Opérateurs d'agrégation pour la mesure de similarité. Application à l'ambiguïté en reconnaissance de formes.

Hoel Le Capitaine

▶ **To cite this version:**

Hoel Le Capitaine. Opérateurs d'agrégation pour la mesure de similarité. Application à l'ambiguïté en reconnaissance de formes.. Traitement du signal et de l'image [eess.SP]. Université de La Rochelle, 2009. Français. NNT : . tel-00438516

HAL Id: tel-00438516

https://theses.hal.science/tel-00438516v1

Submitted on 3 Dec 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université de La Rochelle — École Doctorale S2I

Opérateurs d'agrégation pour la mesure de similarité
Application à l'ambiguïté en reconnaissance de formes

Directeur de thèse : Carl FRÉLICOT

Aggregation operators for similarity measures
Application to ambiguity in pattern recognition

THÈSE

présentée et soutenue publiquement le 17 Novembre 2009,
En vue de l'obtention du grade et du titre de
Docteur de l'Université de La Rochelle
(spécialité AUTOMATIQUE, IMAGE ET SIGNAL)

par

Hoel LE CAPITAINE

devant le jury composé de

Président du jury :
Michel GRABISCH, Professeur, Université de Paris I, Panthéon-Sorbonne

Rapporteurs :
Didier DUBOIS, Directeur de Recherche CNRS, Université Paul Sabatier, Toulouse

Bernard DUBUISSON, Professeur, Université de Technologie de Compiègne

Examinateurs :
Michel BERTHIER, Professeur, Université de La Rochelle

Carl FRELICOT, Professeur, Université de La Rochelle

Laboratoire de Mathématiques, Image et Applications – EA 3165

©Hoel Le Capitaine, La Rochelle, 2009

Résumé

Dans cette thèse, nous nous intéressons à deux problèmes de reconnaissance de formes : l'option de rejet en classification supervisée, et la détermination du nombre de classes en classification non supervisée. Le premier problème consiste à déterminer les zones de l'espace des attributs où les observations n'appartiennent pas clairement à une seule classe. Le second problème repose sur l'analyse d'un nuage d'observations pour lesquelles on ne connait pas les classes d'appartenance. L'objectif est de dégager des structures permettant de distinguer les différentes classes, et en particulier de trouver leur nombre. Pour résoudre ces problèmes, nous fondons nos propositions sur des opérateurs d'agrégation, en particulier des normes triangulaires. Nous définissons de nouvelles mesures de similarité permettant la caractérisation de situations variées. En particulier, nous proposons de nouveaux types de mesures de similarité : la similarité d'ordre, la similarité par blocs, et enfin la similarité par une approche logique. Ces différentes mesures de similarité sont ensuite appliquées aux problèmes évoqués précédemment. Le caractère générique des mesures proposées permet de retrouver de nombreuses propositions de la littérature, ainsi qu'une grande souplesse d'utilisation en pratique. Des résultats expérimentaux sur des jeux de données standard des domaines considérés viennent valider notre approche.

Mots-clés : Opérateurs d'agrégation, similarité de valeurs numériques, implications floues, comparaison d'ensembles flous, classification, coalescence, options de rejet, validité de partition.

Abstract

In this thesis, we are interested in two problems of pattern recognition : the reject option in supervised classification, and determining the number of classes in unsupervised classification. The first problem consists in finding the areas in the feature space where samples do not clearly belong to one class. The second problem is based on the analysis of a cloud of observations which are unlabeled. The objective is to define structures that distinguish different classes, and in particular to find the number of classes. In order to solve these problems, our approach is based on the use of similarity measures allowing to discriminate various situations. In particular, we propose new kinds of similarity measures: the order similarity, the blockwise similarity, and eventually a logic-based similarity measure. These different measures are then applied to the aforementioned problems. The genericity of the proposed measure enables to retrieve usual criterions from the literature, as well as a great versatility in practice. Experimental results on benchmarks datasets for both problems validate our approach.

Key-words: Aggregation operators, similarity of numerical values, fuzzy implications, comparison of fuzzy sets, classification, clustering, reject option, cluster validity.

Remerciements

Après avoir passé quelque temps à rédiger le manuscrit de thèse, s'attaquer aux remerciements est généralement un soulagement, et en même temps un souci, puisque synonyme d'une page qui se tourne.

Je tiens tout d'abord à remercier l'ensemble du jury d'avoir accepté d'évaluer mon travail de thèse : Didier Dubois et Bernard Dubuisson pour avoir rapporté cette thèse, ainsi que Michel Grabisch pour avoir examiné mon travail, et présidé mon jury.

Je tiens également à adresser un vif remerciement à Carl Frélicot, pour avoir proposé puis encadré cette thèse. Il aura été d'une aide précieuse tout au long de ces trois années, aussi bien au niveau de conseils scientifiques, de temps passé à la rédaction d'articles, que sur le plan humain.

Je remercie aussi Michel Berthier, pour avoir accepté d'être examinateur de cette thèse, mais aussi pour les nombreuses discussions, scientifiques ou non, que nous avons eues.

Je voudrais remercier Christophe et Renaud pour leur accueil chaleureux lors de mon arrivée au laboratoire, et pour les nombreux repas, discussions que nous avons partagé.

Je ne pourrais pas terminer ces remerciements sans penser aux autres doctorant(e)s. Dans le désordre, grand To, ptit to, Guillaume, Sullivan, Agathe, Matéo, Benjamin, Sloven. Merci pour les pauses devant Pascal! Une pensée particulière pour Thomas, avec qui j'ai eu une collaboration scientifique plus étroite.

Un petit mot également pour Christine, secrétaire du laboratoire. J'ai usé de sa disponibilité et de sa gentillesse pour la gestion de dossiers urgentissimes ... merci!

Enfin, je voudrais remercier Jean-Yves et Annie, mes parents, sans qui tout cela n'aurait pas été possible. Last but not least, je remercie Maud pour m'avoir encouragé durant des moments parfois difficiles, et Zoé pour ces moments de détente entre deux séances de travail : elles m'auront permis de vivre cette expérience plus sereinement.

Merci à tous!

Hoel Le Capitaine, La Rochelle, Octobre 2009.

Table des matières

Notations, Symboles et Abréviations

Notations	Description	Définition
	Opérateurs d'Agrégation	
I	Interval unité: $[0, 1]$	page 10
$\mathcal{A}$	Opérateur d'agrégation	page 10
$\top$	Norme triangulaire (t-norme)	page 17
$\bot$	Conorme triangulaire (t-conorme)	page 17
$\mathcal{U}$	Uninorme	page 173
$\mathcal{V}$	Nullnorme	page 174
$\wedge$	Conjonction	page 10
$\vee$	Disjonction	page 10
N	Ensemble des sources $\{1, \cdots, n\}$	page 10
μ	Mesure floue	page 22
C_n^i	Combinaison de i parmi n	page 22
$\mathcal{C}_\mu$	Intégrale de Choquet de mesure floue μ	page 23
$\mathcal{S}_\mu$	Intégrale de Sugeno de mesure floue μ	page 24
$\mathcal{D}$	Mesure de distance	page 62
$\mathcal{I}$	Mesure d'inclusion	page 59
$\mathcal{S}$	Mesure de similarité	page 61
	Reconnaissance De Formes	
$\mathbf{x}$	vecteur forme, objet	page 74
p	nombre de variables descriptives de la forme	page 74
n	nombre d'exemples dans les données d'apprentissage	page 74
χ	ensemble de données pour l'apprentissage, $card(\chi) = n$	page 74
c	nombre de classes	page 74
Ω	ensemble des classes, $card(\Omega) = c$	page 74
ω_i	i-ème classe, $i = 1, \cdots, c$	page 74
$\mathbf{v}_i$	prototype de la i-ème classe	page 82
Σ_i	matrice de covariance de la i-ème classe, ω_i	page 82
u_{ik}	degré d'appartenance du k-ième point de χ à la classe ω_i	page 89
A^T	transposé de la matrice A	page 82

Introduction générale

Introduction

Au fond, Dieu veut que l'homme désobéisse. Désobéir, c'est chercher.

— Tas de pierres
Victor Hugo (1901)

Contexte et problématique

Dans cette thèse, nous allons aborder deux problèmes de reconnaissance de formes. Le premier concerne la confiance dans l'assignation d'une forme à une classe : si elle est faible, il vaut mieux refuser le classement. On doit faire face à ce genre de situations lorsque les descriptions des formes pour des classes différentes sont presque identiques, ou au contraire, qu'elles ne ressemblent à aucune de celles connues. Le second problème consiste à déterminer le nombre de classes à prendre en considération à partir des seules descriptions des formes. Il s'agit ici de distinguer les séparations entre les classes afin de pouvoir les compter, et ce même si elles se chevauchent, ou que le bruit dans les données est important. Le développement de méthodes pour la résolution de ces problèmes repose en partie sur l'agrégation d'étiquettes liant formes et classes, dans un cadre mathématique prédéfini parfois étroit (par exemple des probabilités).

L'agrégation d'information est un concept important dans le mécanisme de raisonnement humain. Un effort important a été porté au développement de principes et outils pour l'agrégation tant du point de vue théorique que pratique depuis bientôt trois décennies. Les cadres mathématiques sur lesquels ils sont fondés aujourd'hui sont moins étroits et ils sont utilisés pour caractériser les éléments d'information analysés de divers types, pour des applications variées, par exemple une distribution non uniforme des données (reconnaissance de formes), l'importance d'un ensemble d'alternatives par rapport à d'autres (décision multi-critère), ou encore un désaccord complet entre plusieurs "évaluateurs". Les cadres mathématiques actuels permettent, par leur souplesse, de définir des opérateurs d'agrégation, par essence plus génériques, qui s'avèrent mieux adaptés au traitement d'informations plus complexes ou correspondant à des situations plus délicates. C'est pourquoi les méthodes modernes de reconnaissance de formes s'appuient sur ces derniers.

Ce lien étroit entre ces deux domaines de recherche que sont l'agrégation et la reconnaissance de formes constitue le fondement de cette thèse. Nous allons principalement nous intéresser à l'utilisation et la combinaison d'opérateurs d'agrégation particuliers, les normes triangulaires. Celles-ci étant des extensions floues des notions classiques d'union et d'intersection ensemblistes, leur utilisation paraît toute indiquée pour des problèmes de reconnaissance de formes où les classes se chevauchent, ou que l'on constate la présence de points isolés. Au cœur de nos préoccupations, il existe un autre lien fondamental entre les deux champs disciplinaires considérés : le concept d'incertitude, ou de doute, qui est incontournable pour la prise de décision. L'utilisation d'outils plus souples et tout aussi performants permettant de gérer l'incertitude dans le cadre de la classification d'objets conforte donc pleinement notre démarche.

Nous allons donc aborder ces problèmes d'un point vue tout d'abord théorique sur les opérateurs d'agrégation, puis sur les opérateurs de similarité fondés sur ces opérateurs d'agrégation usuels. Ce travail a pour but d'étudier la notion de similarité définie à l'aide de normes triangulaires, ainsi qu'une analyse de son comportement en fonction de la famille de normes triangulaires choisie. Le caractère général des normes triangulaires permet de formaliser un cadre générique d'opérations de comparaison et de caractérisation des quantités analysées. Ces opérateurs de similarité, très généraux, seront ensuite appliqués aux situations spécifiques décrites ci-avant de reconnaissance de formes.

Plan du mémoire

Ce mémoire est organisé en cinq chapitres principaux, regroupés en deux parties. Nous avons choisi de distinguer le travail plus fondamental sur les opérateurs d'agrégation dans la Partie I, et de consacrer la Partie II aux problèmes de reconnaissance de formes où les opérateurs proposés tiendront un rôle important. Dans chacune des parties, un bref état de l'art est proposé, suivi de nos contributions au domaine concerné.

La première partie est composée de deux chapitres. Le premier (page 9) est consacré aux opérateurs d'agrégation. Nous donnons dans ce chapitre d'un côté les propriétés mathématiques que ces opérateurs peuvent respecter si l'on adopte une approche axiomatique, et de l'autre, nous donnons également les définitions de comportement plus généraux des opérateurs d'agrégation, si l'utilisateur désire construire son opérateur par une approche comportementale. Une fois ces propriétés présentées, nous les illustrons en donnant un catalogue, le plus récent possible, des différents types d'opérateurs que l'on peut trouver dans la littérature. Nous donnons enfin quelques résultats et propositions sur l'opérateur de similarité d'ordre fondé sur la combinaison de normes triangulaires. Le second chapitre (page 35) est dédié dans un premier temps à la description des mesures de caractérisation de vecteurs susceptibles d'être agrégés. Nous présentons ensuite les travaux théoriques réalisés concernant la définition de nouveaux opérateurs de similarité par blocs et de similarité fondée sur des implications floues. Enfin, nous présentons les mesures de comparaison d'ensembles que nous proposons, fondées elles-aussi sur des implications floues.

La deuxième partie est composée de trois chapitres. Dans le premier (page 73), nous commençons par présenter plusieurs approches de classification supervisée et non supervisée en reconnaissance de formes statistique. Nous terminons ce chapitre en décrivant quelques problèmes que l'on peut rencontrer lors de la mise en place de ces méthodes de classification : présence de points isolés ou de bruit, chevauchement de classes pour la classification supervisée, détermination du nombre de groupes en classification non supervisée. Nous appliquons les outils proposés dans la première partie à ces problèmes dans les deux chapitres suivants (pages 97 et 125). L'introduction de ces mesures permet de définir des modèles génériques pour les deux problèmes abordés, et aboutit à des performances supérieures aux méthodes existant dans l'état de l'art.

Enfin, nous donnons dans la conclusion (page 149) un résumé global de la thèse et des apports que nous avons réalisé, ainsi que des perspectives concernant nos futurs travaux.

Première partie

Théorie

Chapitre 1

Agrégation et agrégation d'ordre

Agréger [agreʒe] Action de réunir des éléments distincts pour former un tout homogène.

Larousse (2006)

Résumé : *Ce chapitre a pour but de présenter, de manière non exhaustive, un panorama des différents opérateurs d'agrégation qui auront été proposés au fil du temps, et de donner un bref aperçu de leur possible utilisation dans le domaine qui nous intéresse : la reconnaissance de formes. Nous proposons également un nouvel opérateur de similarité d'ordre en fin de chapitre.*

1.1 Introduction et définitions

Alors que le volume d'information disponible à tout un chacun est de plus en plus grand, le besoin de résumer cette information, de la caractériser, afin d'en tirer avantage au maximum, augmente lui aussi. A l'instar de la reconnaissance des formes, où nous voulons, parmi une quantité d'observations, dégager de grandes tendances et ainsi les classer, les opérations d'agrégation ont pour but de résumer ces tendances afin d'un éventuel regroupement postérieur. C'est ainsi que l'on pourra retrouver dans les méthodes de reconnaissance de formes de nombreux outils qui auront été proposés dans le cadre théorique des opérateurs d'agrégation (approches bayésiennes [Clemen and Winkler, 1999], possibilistes [Dubois and Prade, 1988], intégrales floues [Grabisch, 2000b]). Une opération d'agrégation, ou de fusion, consiste à combiner plusieurs valeurs (cardinales, ordinales) de manière à en obtenir une seule. Cette valeur de sortie devra évidemment représenter au mieux l'ensemble des valeurs d'entrée, et donc tenir compte de chacune d'elles. À partir de cette conception très générale, on voit tout de suite que ce processus d'agrégation peut être un outil utilisé dans de nombreux domaines. À titre d'exemple, on pourra citer les sciences économiques, sociales, l'ingénierie, et pour des domaines plus particuliers, la prise de décision, la reconnaissance de formes, l'apprentissage machine ou encore le traitement d'image. Les problèmes d'agré-

gation sont donc, comme on peut le constater, assez larges et hétérogènes. Ainsi, on trouve dans la littérature associée des articles portant sur l'agrégation d'un nombre infini de valeurs réelles [Grabisch et al., 2000], de valeurs ordinales [Grabisch, 2000a; Yager, 2007], de distribution de probabilités [Schweizer and Sklar, 1983; Nelsen, 2006], ou encore d'ensembles flous [Dubois and Prade, 1985] et de différentes sources incertaines [Dubois and Prade, 2004]. Dans ce mémoire, nous nous restreindrons à l'étude de fonctions réelles de n variables appartenant à l'intervalle unité, $I = [0,1]$ ayant également valeur dans I. Par la suite, nous noterons $\vee$ et $\wedge$ les opérateurs *maximum* et *minimum*, respectivement. Nous prendrons également comme convention $N = \{1, \cdots, n\}$ l'ensemble des sources, et $X = (x_1, \cdots, x_n)$ la suite d'arguments, critères, entrées à agréger. Dans un cadre général, le nombre de valeurs (arguments) d'entrée à agréger est un nombre quelconque n, et un opérateur d'agrégation $\mathcal{A}$ est alors défini comme une fonction n-aire

$$\mathcal{A} : \bigcup_{n \in \mathbb{N}} [0,1]^n \to [0,1]. \tag{1.1}$$

Le premier opérateur satisfaisant ces contraintes et auquel on pense immédiatement, car utilisé tous les jours, est la moyenne, voir exemple 1.1.

Exemple 1.1. La moyenne arithmétique définie par

$$\mathcal{M}(x_1, \cdots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

est un opérateur d'agrégation où n est le nombre de valeurs à agréger, et les différents x_i sont les valeurs numériques que l'on considère pour l'agrégation.

De nombreuses propriétés à satisfaire ont été proposées pour définir les opérations d'agrégation. Suivant la définition de [Calvo et al., 2002, chap. 1, p. 8], un opérateur d'agrégation doit satisfaire les conditions aux bornes et être monotone. En d'autres termes, l'agrégation de valeurs minimales doit donner une sortie minimale, l'agrégation de valeurs maximales doit résulter en une valeur maximale, et si l'une des valeurs d'entrée augmente, alors la valeur de sortie doit elle aussi augmenter. Au delà de ces propriétés basiques, on peut souhaiter spécifier certains comportements de l'opérateur en fonction de la nature des valeurs que l'on agrège, et adopter pour cela une approche axiomatique. Dans la suite de ce chapitre, nous donnerons et illustrerons dans un premier temps des propriétés que l'on pourrait désirer pour ces opérateurs par une approche soit axiomatique soit comportementale. Ensuite, les opérateurs basés sur l'opération de moyenne seront décrits en section 1.2. Une partie importante sera consacrée en section 1.3 à la description d'une famille particulière d'opérateurs : les normes triangulaires. Suivront enfin les intégrales floues en section 1.4. On pourra trouver d'autres détails sur les opérateurs d'agrégation dans les travaux de thèse de Marichal [Marichal, 1998], plus récemment de Detyniecki [Detyniecki, 2000], des livres consacrés aux opérateurs d'agrégation, [Calvo et al., 2002; Torra and Narukawa, 2007; Beliakov et al., 2008], ou encore [Grabisch et al., 2009] à paraître dans les mois qui viennent. Cette prolifération de publications au sujet des opérateurs d'agrégation prouve à quel point le sujet est vaste et propice à investigations.

Nous donnons ci-après les définitions principales, et des définitions complémentaires sont disponibles en Annexe A.

1.1.1 Conditions aux bornes

L'une des premières propriétés que l'on requiert généralement d'un opérateur d'agrégation n-aire est la condition aux deux bornes de I. Ainsi, indépendamment de n, on veut que

$$\mathcal{A}(0,\cdots,0) = 0 \tag{1.2}$$

et

$$\mathcal{A}(1,\cdots,1) = 1 \tag{1.3}$$

La condition (1.2) revient à dire que si les arguments, ou critères, ne sont pas satisfaits, alors la sortie ne le sera pas. Inversement, la condition (1.3) correspond au fait que si les critères sont complètement satisfaits, alors la sortie le sera. Cette propriété parait fondamentale, et la quasi-totalité des opérateurs que l'on trouve la respecte.

1.1.2 Monotonie

Due aux conditions au bornes (1.2) et (1.3), la monotonie de l'équation (1.1) revient finalement à sa non décroissance par rapport à chaque argument. Si un argument en entrée augmente, on désire alors que le résultat final augmente également, ou tout au moins ne diminue pas.

Définition 1.1. Un opérateur d'agrégation $\mathcal{A}$ est dit monotone si

$$\forall n \in \mathbb{N} : x_1 \leq y_1,\cdots,x_n \leq y_n \Rightarrow \mathcal{A}(x_1,\cdots,x_n) \leq \mathcal{A}(y_1,\cdots,y_n). \tag{1.4}$$

Devançant une notion dont nous aurons besoin par la suite, nous introduisons la définition de la stricte monotonie.

Définition 1.2. Un opérateur d'agrégation $\mathcal{A}$ est dit strictement monotone si

$$\begin{gathered}\forall n \in \mathbb{N}, i \in \{1,\cdots,n\} : (x_i \leq y_i \wedge (x_1,\cdots,x_n) \neq (y_1,\cdots,y_n)) \\ \Downarrow \\ \mathcal{A}(x_1,\cdots,x_n) < \mathcal{A}(y_1,\cdots,y_n).\end{gathered} \tag{1.5}$$

Par exemple, l'opérateur *maximum* n'est pas strictement monotone, alors que la *moyenne arithmétique* l'est.

Comme il existe de nombreuses propriétés, nous ne les présentons pas toutes ici, mais nous donnons un panorama plus complet en Annexe A.

1.1.3 Comportement

Après avoir décrit des propriétés mathématiques concernant les opérateurs d'agrégation, intéressons-nous au comportement général et global que l'utilisateur, ou la personne prenant la décision, attend de celui-ci. En effet les opérateurs d'agrégation peuvent être classés en plusieurs catégories, voir [Grabisch et al., 1998] : conjonctifs, disjonctifs, de compromis, de compensation, de renforcement, pondérés, ceci dépendant de la manière dont sont agrégés les arguments.

1.1.3.1 Opérateurs conjonctifs

Les opérateurs conjonctifs modélisent une agrégation connectant chacun des arguments par une opération correspondant au *et* logique. En d'autres termes, le score final ne sera élevé que si l'ensemble des arguments est élevé, ce qui correspondra finalement à un comportement *intolérant*, puisque tous les critères doivent être satisfaits pour engendrer un score positif.

Définition 1.3. Un opérateur d'agrégation $\mathcal{A}$ sera dit conjonctif si

$$\mathcal{A}(x_1, \cdots, x_n) \leq \min(x_1, \cdots, x_n) \tag{1.6}$$

Si à ce comportement on ajoute des propriétés mathématiques de non décroissance (monotonie), commutativité (symétrie) et associativité que l'on a énoncé en amont, on obtient alors une famille bien connue d'opérateurs d'agrégation : les normes triangulaires. Cette famille sera décrite plus en détail dans la section 1.3, puisque nos travaux reposent sur la combinaison de ces opérateurs.

1.1.3.2 Opérateurs disjonctifs

Les opérateurs disjonctifs modélisent une agrégation connectant chacun des arguments par une opération correspondant au *ou* logique. En d'autres termes, le score final sera haut si au moins un des arguments est haut, ce qui correspondra finalement à un comportement *tolérant*, puisque un seul critère satisfait suffit pour engendrer un score positif.

Définition 1.4. Un opérateur d'agrégation $\mathcal{A}$ sera dit disjonctif si

$$\mathcal{A}(x_1, \cdots, x_n) \geq \max(x_1, \cdots, x_n) \tag{1.7}$$

Si à ce comportement on ajoute des propriétés mathématiques de non décroissance (monotonie), commutativité (symétrie) et associativité que l'on a énoncé en amont, on obtient alors une famille bien connue d'opérateurs d'agrégation : les conormes triangulaires (voir section 1.3).

1.1.3.3 Opérateurs de compromis

Les opérateurs de compromis mènent à un score qui sera compris entre le minimum et le maximum. Le but ici est d'obtenir une sorte de compromis entre tous les arguments en entrée, l'influence d'un faible score pouvant être atténué par un score plus élevé et inversement.

Définition 1.5. Un opérateur d'agrégation $\mathcal{A}$ sera dit de compromis si

$$\min(x_1, \cdots, x_n) \leq \mathcal{A}(x_1, \cdots, x_n) \leq \max(x_1, \cdots, x_n) \tag{1.8}$$

Clairement, les opérateurs de compromis ne sont ni conjonctifs ni disjonctifs. L'exemple le plus connu et sûrement le plus utilisé d'un tel opérateur est certainement la moyenne arithmétique (Exemple 1.1). Parmi ces opérateurs, on trouvera aussi les opérateurs de type moyenne et les statistiques d'ordre (voir section 1.2). D'autres opérateurs rentrant dans cette catégorie seront également décrits par la suite (opérateurs *OWA*, intégrales floues). On pourra par ailleurs trouver une étude détaillée du comportement optimiste ou pessimiste de ces opérateurs dans le cadre de la prise de décision dans [Luo and Jennings, 2007].

1.1.3.4 Opérateurs de compensation

À l'instar des opérateurs de compromis, les opérateurs de compensation servent à exprimer une certaine interaction entre les valeurs; un argument fort pourra compenser un argument faible et vice versa. Ils sont souvent confondus avec les opérateurs de compromis de manière erronée dans la mesure où le résultat final n'est pas nécessairement compris entre le minimum et le maximum. Parmi ces opérateurs, on retrouvera l'opérateur γ introduit par Zimmermann and Zysno dans [Zimmermann and Zysno, 1980], ou de manière plus générale les opérateurs de combinaisons exponentielles ou linéaires [Turksen, 1992], mais aussi les sommes symétriques développées par Silvert [Silvert, 1979].

1.1.3.5 Opérateurs de renforcement

Une propriété intéressante souvent utilisée par les humains est ce que Yager et Rybalov [Yager and Rybalov, 1998] appellent le renforcement *complet*. Ce renforcement complet se divise en fait en deux parties, se décomposant en renforcement par le haut et par le bas. Par renforcement par le haut, les auteurs entendent que si l'ensemble des critères sont satisfaits, alors la sortie doit l'être d'autant plus. Inversement si aucun des critères n'est satisfait, la sortie doit l'être d'autant moins. Ces deux concepts sont en fait équivalents du comportement conjonctif et disjonctif, mais réunis dans un seul et même opérateur. On pourra également relier ces opérateurs aux uninormes, que l'on décrira en section A.2.

Exemple 1.2. L'exemple le plus connu d'opérateur de renforcement complet est le triple $\prod$:

$$\mathcal{A}(x_1, \cdots, x_n) = \frac{\prod_{i=1}^{n} x_i}{\prod_{i=1}^{n} x_i + \prod_{i=1}^{n} \bar{x_i}} \tag{1.9}$$

où $\bar{x}$ est la négation stricte, c'est à dire $\bar{x} = 1 - \bar{x}$.

Un autre type de renforcement, introduit dans [Yager, 2003], et que l'auteur dénomme renforcement *noble*, consiste à considérer le renforcement par le haut, mais sans avoir l'effet collatéral où de nombreuses entrées faibles aboutiraient, par effet de la non décroissance, à une sortie haute.

1.2 Type moyenne

Initialement vue par Cauchy [Cauchy, 1821] en 1821 comme une fonction $\mathcal{M}$ de $(x_1, \cdots, x_n)$ ayant une valeur interne à $(x_1, \cdots, x_n)$, c'est à dire, du point de vue des définitions connues à ce jour, un opérateur de compromis, la notion de moyenne a connu plusieurs évolutions au cours du temps. Plus d'un siècle plus tard, en 1930, Kolmogorov [Kolmogorov, 1930] définit une valeur moyenne comme une séquence infinie continue, symétrique et strictement croissante de fonctions réelles idempotentes (équation (A.1)) et associatives (équation (A.3)). Depuis, d'autres définitions ont été proposées: Dubois et Prade requièrent la continuité en excluant le minimum et le maximum [Dubois and Prade, 1985]. Fodor et Marichal proposent la notion de moyennes croissantes (et non strictement croissantes) [Fodor and Marichal, 1997], puis Marichal proposera une moyenne quasi-arithmétique non symétrique [Marichal, 2000]. Dans la suite de cette section, nous présenterons dans un premier temps les moyennes quasi-arithmétiques, puis les moyennes ordonnées, et enfin un bref aperçu des statistiques d'ordre.

Fonction f	Opérateur $\mathcal{A}$ obtenu	Nom usuel
x	$\frac{1}{n}\sum x_i$	arithmétique
$\log x$	$\sqrt[n]{\prod x_i}$	géométrique
x^{-1}	$\frac{1}{\frac{1}{n}\sum \frac{1}{x_i}}$	harmonique
x^2	$\sqrt{\frac{1}{n}\sum x_i^2}$	quadratique
x^α	$\left(\frac{1}{n}\sum x_i^\alpha\right)^{\frac{1}{\alpha}}$	puissance (Hölder)
$\exp(\alpha x)$	$\frac{1}{\alpha}\log\left(\frac{1}{n}\sum \exp(\alpha x_i)\right)$	exponentielle

Tab. 1.1: Moyennes quasi-arithmétiques incluant de nombreuses moyennes différentes selon la définition de f.

1.2.1 Moyennes (quasi)-arithmétiques

Soit E un intervalle réel, fini ou infini et un opérateur d'agrégation $\mathcal{A}$ symétrique, continu, strictement croissant et idempotent. $\mathcal{A}$ est une moyenne *quasi-arithmétique* si et seulement si il existe une fonction $f : E \to \mathbb{R}$ continue strictement monotone telle que :

$$\mathcal{A}(x_1, \cdots, x_n) = f^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\Big) \tag{1.10}$$

Les moyennes quasi-arithmétiques sont évidemment des opérateurs de compromis, et, selon la fonction f, couvrent un large spectre de moyennes, incluant les moyennes de Tab. 1.1.

On pourra noter que la moyenne de Hölder (puissance) généralise les notions de moyenne harmoniques, géométriques, arithmétiques, quadratiques, et des opérateurs minimum et maximum pour des valeurs de $\alpha = -1, \to 0, 1, 2, +\infty$ et $-\infty$, respectivement. L'introduction de poids associés à chacun des arguments permet de donner une importance à chacun des critères considérés. Ces poids sont généralement utilisés soit en dupliquant les entrées auxquelles on veut donner de l'importance, soit après normalisation de ces importances, d'associer à chaque x_i un poids $w_i \geq 0$ respectant la contrainte $\sum_{i=1}^{n} w_i = 1$. Lorsque l'on associe un poids à chaque critère, on perd évidemment la propriété de symétrie. Utilisant cette notion de pondération, on peut alors introduire la moyenne quasi-arithmétique pondérée définie par

$$\mathcal{A}_{\mathbf{w}}(x_1, \cdots, x_n) = f^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} w_i\, f(x_i)\Big) \tag{1.11}$$

1.2.2 Moyennes ordonnées

Comme précisé dans la section précédente, associer des poids à chacun des critères revient à supprimer la propriété de symétrie de l'opérateur. Si l'on ordonne les valeurs à agréger, il faut nécessairement associer un poids à chacune des valeurs sous peine d'obtenir la moyenne quasi-arithmétique usuelle. C'est ainsi que sont apparues les moyennes ordonnées pondérées introduites par Yager dans [Yager, 1988] (communément appelées *OWA* pour *Ordered Weighted Average* operators) et objets d'un ouvrage [Yager and Kacprzyk,

1997]. L'opérateur d'agrégation *OWA* est défini par

$$OWA_{\mathbf{w}}(x_1,\cdots,x_n) = \sum_{i=1}^{n} w_i\, x_{\sigma(i)} \tag{1.12}$$

où $\sigma(.)$ est une permutation telle que l'on a $x_{\sigma(1)} \geq \cdots \geq x_{\sigma(n)}$, et le vecteur de poids $\mathbf{w}$ respecte les mêmes conditions énoncées lors de la définition de la moyenne quasi-arithmétique pondérée (section 1.2.1). On voit immédiatement que cet opérateur est de compromis, monotone, symétrique et idempotent. Cet opérateur possède en outre la particularité de généraliser un certain nombre d'opérateurs que nous avons déjà mentionné jusqu'ici: minimum, maximum, statistiques d'ordre (voir section 1.2.3), moyenne arithmétique. Une manière de décrire, ou caractériser un opérateur *OWA*, consiste à évaluer $\mathbf{w}$. Dans [Yager, 1988], Yager propose un degré de *"maxitude"* défini par

$$maxitude(w_1,\cdots,w_n) = \frac{1}{n-1}\sum_{i=1}^{n} w_i\,(n-i) \tag{1.13}$$

Pour l'opérateur maximum, nous avons ainsi $maxitude(1,0,\cdots,0) = 1$, et inversement, pour l'opérateur minimum, nous avons $maxitude(0,\cdots,0,1) = 0$. Une seconde manière de caractériser w est de quantifier la dispersion des poids, ou tout simplement d'évaluer le caractère uniforme de leur distribution. Pour cela, une mesure de type entropique

$$dispersion(w_1,\cdots,w_n) = -\sum_{i=1}^{n} w_i \log w_i \tag{1.14}$$

est proposée dans [Yager, 1988]. On reliera immédiatement ce type de mesures aux indices de validité de partition qui seront présentés et développés dans le Chapitre 5. Le processus de détermination des poids de cet opérateur est très important, puisqu'il le caractérise, et de nombreuses propositions ont été faites dans cette optique. La première [Yager, 1988] consiste à utiliser des quantifieurs linguistiques représentés par des ensembles flous Q tels que $Q(0) = 0$ et $Q(1) = 1$, et définir le poids w_i de la manière suivante:

$$w_i = Q(\frac{i}{n}) - Q(\frac{i-1}{n}) \tag{1.15}$$

Une deuxième manière consiste à utiliser les mesures de dispersion précédemment introduites, et, avec une maxitude donnée, de maximiser la dispersion des poids, toujours sous la contrainte de somme unitaire et qu'ils soient compris dans l'intervalle unité. On pourra se référer à [Beliakov, 2005; Xu, 2005; Fuller, 2007] pour une revue récente des différentes méthodes d'obtention des poids pour l'opérateur *OWA*. Suite à son introduction en 1988, cet opérateur a donné lieu à un nombre impressionnant de nouvelles propositions, parmi lesquelles on citera, sans les détailler par mesure de concision, OWA pondéré (Weighted OWA) [Torra, 1997], OWA induit (Induced OWA) [Yager and Filev, 1999], OWA incertain (Uncertain OWA) [Xu and Da, 2002], OWA lourd (Heavy OWA) [Yager, 2002], OWA généralisé (Generalized OWA) [Yager, 2004], OWA dépendant (Dependant OWA) [Xu, 2006], OWA généralisé induit (Generalized Induced OWA) [Merigó and Gil-Lafuente, 2009].

1.2.3 Statistiques d'ordre

Très souvent, on ne dispose que d'informations à caractère ordinal, et des opérateurs d'agrégation spécifiques sont alors nécessaires. L'opérateur *médian* est souvent utilisé dans

cette situation, mais une définition plus générale peut être donnée, la statistique d'ordre k. Celle-ci est définie par

$$SO_k(x_1,\cdots,x_n) = x_{(k)} \tag{1.16}$$

où $x_{(k)}$ désigne la k-ième plus petite valeur du n-uplet $(x_1,\cdots,x_n)$. À partir de cet opérateur, on pourra distinguer plusieurs cas particuliers. Lorsque $k=1$ et $k=n$, on obtient les opérateurs minimum et maximum, respectivement. Un opérateur qui peut être qualifié de compromis est l'opérateur *médian*, et correspond à définir la valeur milieu sur les valeurs triées ($k=(n+1)/2$ si n est impair, et on prend la moyenne de $k=n/2$ et $k=(n+1)/2$ si n est pair). Dans [Calvo and Mesiar, 2001], les auteurs introduisent une généralisation de cette notion basée sur l'utilisation de distance. Enfin, plus récemment, Mascarilla et al. proposent un opérateur d'ordre 2, puis k, par combinaison de normes triangulaires (voir section suivante). Cet opérateur devient une nouvelle version de la statistique d'ordre k, puisque si les normes triangulaires minimum et maximum sont utilisées, l'ordre k correspond à la k-ième plus grande valeur du n-uplet $(x_1,\cdots,x_n)$, [Mascarilla et al., 2008].

1.2.4 Min-Max ordonnés pondérés

Avec le même but que les opérateurs OWA permettant de relâcher les contraintes des opérateurs conjonctifs et disjonctifs, les minimum et maximum ordonnés pondérés sont introduits dans [Dubois et al., 1997]. Ils sont respectivement définis par

$$OWmin(x_1,\cdots,x_n) = \max_{i=1,n}\Big(\min(x_{\alpha(i)},w_i)\Big) \tag{1.17}$$

et

$$OWmax(x_1,\cdots,x_n) = \min_{i=1,n}\Big(\max(x_{\alpha(i)},w_i)\Big) \tag{1.18}$$

où $\alpha(.)$ est une permutation telle que l'on a $x_{\alpha(1)} \leq \cdots \leq x_{\alpha(n)}$ et les poids w_i respectent les contraintes suivantes : $w_i \in [0,1]$ et

- $\max_i(w_i) = 1$ pour $OWmin$,
- $\min_i(w_i) = 0$ pour $OWmax$.

Si $w_k = 1$ et $w_i = 0$ pour $i > k$, alors $OWmin$ est la statistique d'ordre k.

1.3 Normes triangulaires

Karl Menger [Menger, 1942] introduit la notion de normes triangulaires, ou t-normes, en 1942. Dans leur premier usage, ces opérateurs sont produits afin de généraliser l'inégalité triangulaire des espaces métriques classiques à des espaces métriques statistiques ou probabilistes. Les fonctions telles que définies par Menger forment une classe large et hétérogène d'opérateurs binaires, symétriques et non décroissants, satisfaisant $\mathcal{A}(1,x) > 0$ si $x > 0$. Pourtant, aujourd'hui, la définition communément admise des normes triangulaires est celle proposée par Schweizer et Sklar dans [Schweizer and Sklar, 1983], où les propriétés d'associativité et d'un élément neutre fixé à 1, impliquant $\mathcal{A}(1,x) = x$, sont ajoutées. On peut remarquer que l'associativité permet d'étendre l'opérateur binaire à son équivalent n-aire, obtenant ainsi une généralisation de l'inégalité polygonale. Depuis, ces opérateurs ont été largement étudiés [Klement et al., 2000; Klement and Mesiar, 2005; Alsina et al., 2006].

1.3.1 Définitions et propriétés

Le terme norme triangulaire vient de la définition de Menger, et les axiomes caractérisant cet opérateur proviennent des travaux de Schweizer et Sklar.

Définition 1.6. Une norme triangulaire, ou t-norme, est une opération binaire $\top$ sur l'intervalle unité, c'est à dire une fonction $\top : [0,1]^2 \to [0,1]$, telle que pour tout x, y et z dans [0,1], les quatre axiomes suivants sont satisfaits:

$$\begin{aligned}
&\textit{(P1)} && \top(x,y) = \top(y,x) && \text{commutativité} && (1.19)\\
&\textit{(P2)} && \top(x,\top(y,z)) = \top(\top(x,y),z) && \text{associativité} && (1.20)\\
&\textit{(P3)} && \top(x,y) \leq \top(x,z) \textit{ si } y \leq z && \text{monotonie} && (1.21)\\
&\textit{(P4)} && \top(x,1) = x && \text{élément neutre} && (1.22)
\end{aligned}$$

Il existe énormément de normes triangulaires (en réalité une infinité), mais nous présentons dans cette section les quatre t-normes dites de base à partir desquelles on peut construire toutes les autres [Klement et al., 2000].

Exemple 1.3. Les quatre normes triangulaires de base $\top_M$, $\top_P$, $\top_L$ et $\top_D$ sont respectivement définies par :

$$\begin{aligned}
&\top_M(x,y) = \min(x,y) && \text{Minimum}\,[1] && (1.23)\\
&\top_P(x,y) = x \cdot y && \text{Produit}\,[2] && (1.24)\\
&\top_L(x,y) = \max(x+y-1,0) && \text{Łukasiewicz} && (1.25)\\
&\top_D(x,y) = \begin{cases} 0 & \text{si } (x,y) \in [0,1[^2 \\ \min(x,y) & \text{sinon} \end{cases} && \text{Drastique} && (1.26)
\end{aligned}$$

On peut noter que la t-norme *minimum* est aussi connue sous le nom de *standard*, ou *Zadeh*, et la t-norme produit sous le nom de *algébrique*. Une visualisation des valeurs de sortie de ces quatres normes triangulaires est donnée en Fig. 1.1.

Définition 1.7. Une conorme triangulaire, ou t-conorme, est une opération binaire $\perp$ sur l'intervalle unité, c'est à dire une fonction $\perp : [0,1]^2 \to [0,1]$, telle que pour tout x, y et z dans [0,1], les axiomes *(P1), (P2), (P3)* de la Définition 1.6, et

$$\textit{(P4')} \qquad \perp(x,0) = x \qquad \text{élément neutre} \qquad (1.27)$$

sont satisfaits.

Exemple 1.4. Les quatre conormes triangulaires de base $\perp_M$, $\perp_P$, $\perp_L$ et $\perp_D$ sont respectivement définies par :

$$\begin{aligned}
&\perp_M(x,y) = \max(x,y) && \text{maximum}\,[3] && (1.28)\\
&\perp_P(x,y) = x+y-x\cdot y && \text{somme probabiliste}\,[4] && (1.29)\\
&\perp_L(x,y) = \min(x+y,1) && \text{Łukasiewicz} && (1.30)\\
&\perp_D(x,y) = \begin{cases} 1 & \text{si } (x,y) \in]0,1]^2 \\ \max(x,y) & \text{sinon} \end{cases} && \text{Drastique} && (1.31)
\end{aligned}$$

1. Par la suite, la notation $\top_M$ ou $\top_S$ sera utilisée pour cet opérateur
2. Par la suite, la notation $\top_P$ ou $\top_A$ sera utilisée pour cet opérateur

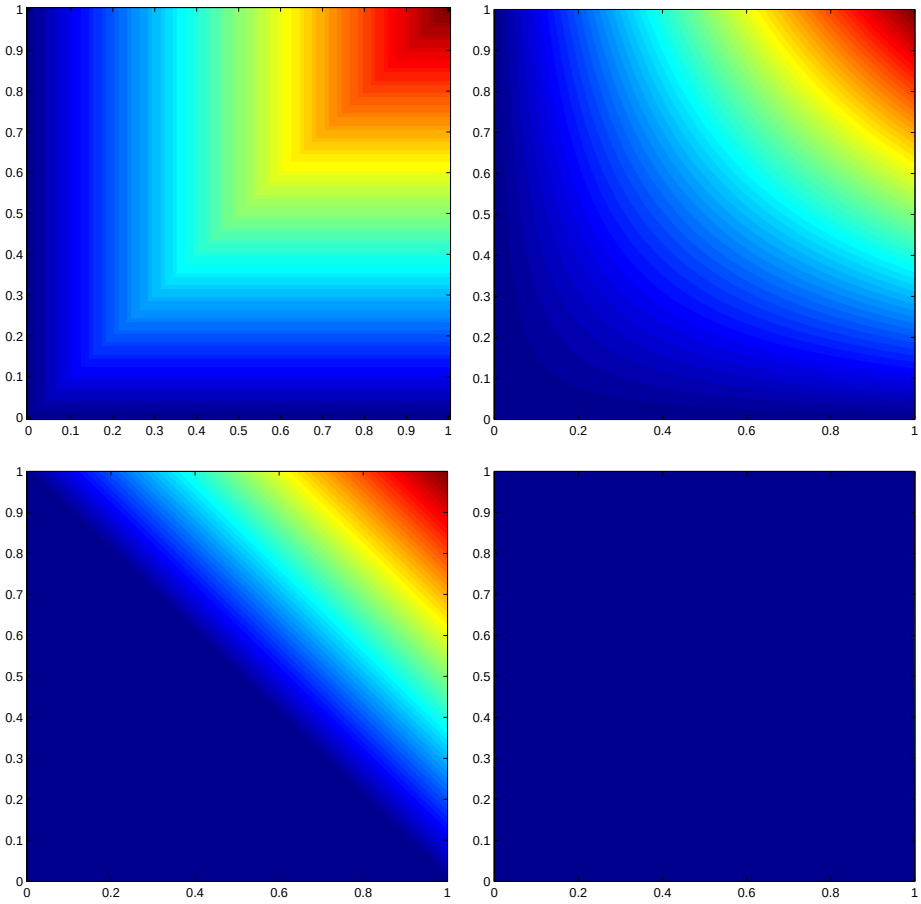

FIG. 1.1: Première ligne : Iso-surfaces des normes triangulaires min et produit ($\top_M$ et $\top_P$). Deuxième ligne : Iso-surfaces des normes triangulaires de Łukasiewicz et Drastique ($\top_L$ et $\top_D$).

De la même manière que pour les t-normes, certaines de ces t-conormes sont connues sous d'autres noms, en particulier $\perp_L$ est aussi appelée *somme bornée*. Une visualisation des valeurs de sortie de ces quatre t-conormes triangulaires est donnée en FIG. 1.2. On peut aussi définir une t-conorme à partir de la définition d'une négation stricte telle que $\bar{x} = 1 - x$. Ainsi, on aura la relation duale entre $\perp$ et $\top$:

$$\top(x, y) = 1 - \perp(1 - x, 1 - y) \tag{1.32}$$

ou, de manière équivalente,

$$\perp(x, y) = 1 - \top(1 - x, 1 - y) \tag{1.33}$$

c'est à dire qu'elles satisfont les lois de De Morgan généralisées. Des différentes propriétés de ces opérateurs, on peut rapidement en déduire de nouvelles. Clairement, on a

$$\top(x, y) \leq x \quad et \quad \top(x, y) \leq y \tag{1.34}$$

$$\perp(x, y) \geq x \quad et \quad \perp(x, y) \geq y \tag{1.35}$$

3. Par la suite, la notation $\perp_M$ ou $\perp_S$ sera utilisée cet opérateur
4. Par la suite, la notation $\perp_P$ ou $\perp_A$ sera utilisée cet opérateur

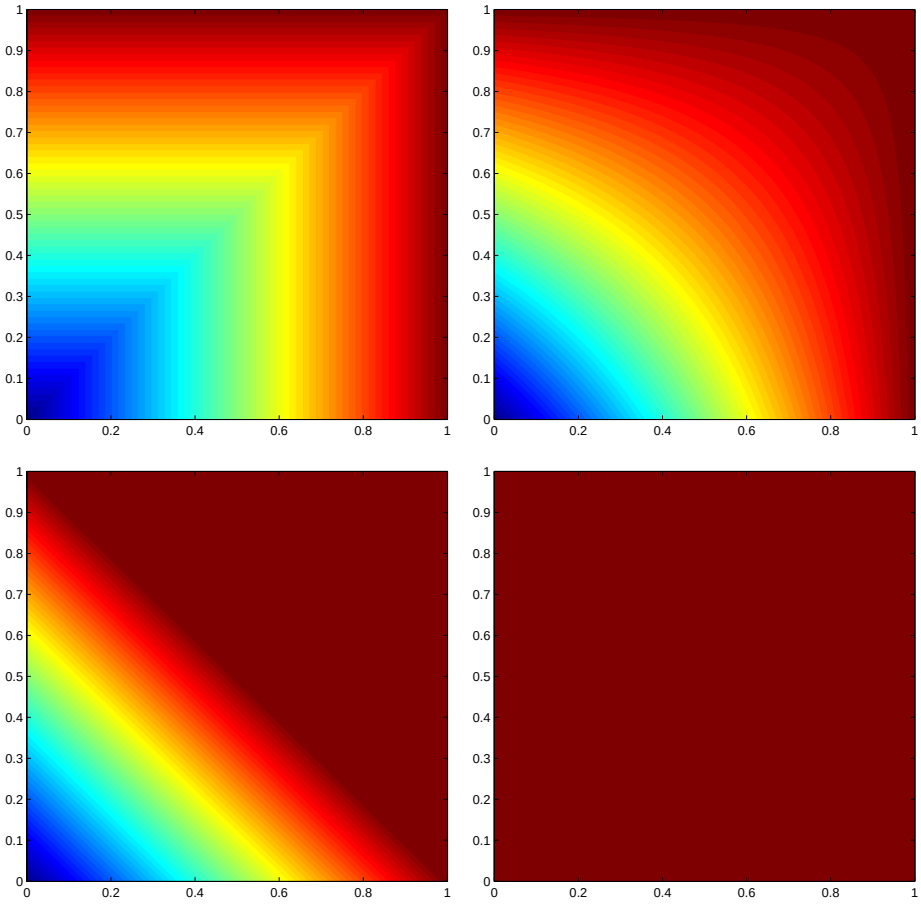

FIG. 1.2: Première ligne : Iso-surface des conormes triangulaires max et somme probabiliste ($\perp_M$ et $\perp_P$). Deuxième ligne : Iso-surface des conormes triangulaires de Łukasiewicz et drastique ($\perp_L$ et $\perp_D$).

ayant pour conséquence que 0 et 1 sont des éléments absorbants des t-normes et t-conormes, respectivement. On aura aussi l'inégalité suivante

$$\top(x, y) \leq \min(x, y) \leq \max(x, y) \leq \perp(x, y) \tag{1.36}$$

L'équation (1.36) permet ainsi de borner les couples de normes triangulaires, et d'aboutir en particulier à la conclusion suivante : le minimum est la plus grande des t-normes, tandis que le maximum est la plus petite des t-conormes. À partir de ces observations, on peut donc classer les normes triangulaires dans les opérateurs conjonctifs, et les conormes triangulaires dans les opérateurs disjonctifs. La propriété d'associativité permet également d'étendre ces opérateurs initialement binaires à leurs pendants n-aire. Une propriété intéressante que l'on peut associer aux normes triangulaires est leur propriété archimédienne. On dit qu'une t-norme continue est archimédienne si

$$\top(x, x) < x,$$

et inversement une t-conorme continue est archimedienne si

$$\perp(x, x) > x.$$

Les premiers résultats de Abel sur l'équation fonctionnelle de l'associativité, puis les travaux de Aczél [Aczél, 1949], suivis par Schweizer et Sklar, permettent de donner une représenta-

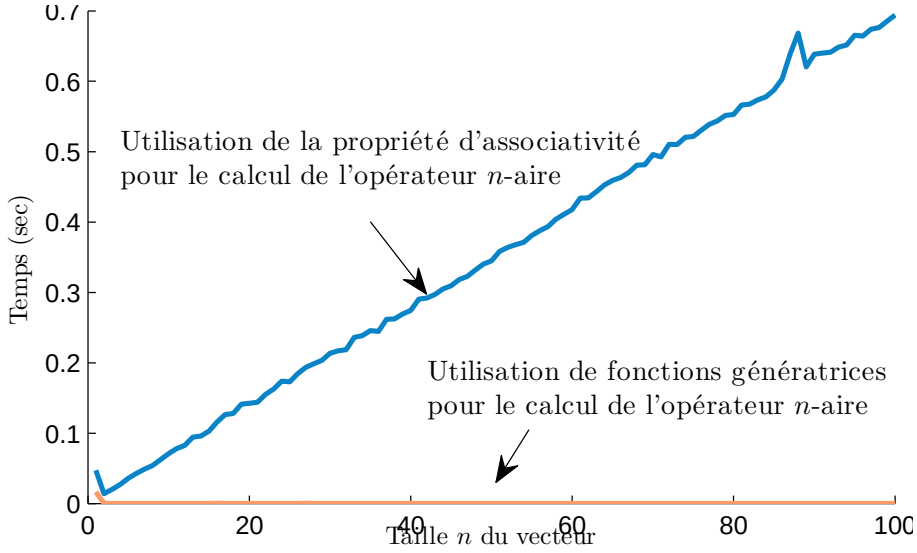

FIG. 1.3: Influence sur le temps de calcul (Matlab™) de l'utilisation de fonctions génératrices et de la propriété d'associativité des normes triangulaires.

tion fondamentale des t-normes, puisqu'il est prouvé que pour toute t-norme continue archimedienne, il existe une fonction $f : I \to [0,\infty]$ continue décroissante respectant $f(0) = +\infty$ et $f(1) = 0$, permettant de représenter cette t-norme de la manière suivante :

$$\top(x, y) = f^{(-1)}(f(x) + f(y)) \tag{1.37}$$

ou plus généralement,

$$\top_{i=1}^{n} x_i = f^{(-1)}(\sum_{i=1}^{n} f(x_i)) \tag{1.38}$$

avec $f^{(-1)}$ le pseudo-inverse défini par :

$$f^{(-1)}(u) = \begin{cases} f^{-1}(u) & \text{si } 0 \leq u < f(0) \\ 0 & \text{si} f(0) \leq u \leq \infty \end{cases} \tag{1.39}$$

On dit que f est un *générateur additif* de $\top$. Notons qu'il existe aussi des générateurs multiplicatifs $g : [0, 1] \to [0, 1]$, et on a alors

$$\top_{i=1}^{n} x_i = g^{(-1)}(\prod_{i=1}^{n} g(x_i)) \tag{1.40}$$

où g est une fonction strictement croissante qui satisfait $g(1) = 1$. Cette génération de normes triangulaires par des fonctions à une seule variable possède un intérêt tout particulier d'un point de vue calculatoire : le temps de calcul de l'agrégation d'un vecteur de taille n est grandement diminué par cette technique, comme le montre la FIG. 1.3. Par la suite, de nombreux autres couples ont été proposés, la plupart possédant un paramètre permettant d'obtenir un comportement différent selon sa valeur, et en particulier de retrouver certains des couples dits basiques dans ce mémoire. Nous proposons en ANNEXE C la définition de la majorité des couples de normes triangulaires paramétriques que l'on peut trouver dans la littérature, ainsi qu'une visualisation des valeurs de sortie sur le carré unité. Un tableau récapitulatif des couples paramétriques est tout de même donné dans cette section, voir TAB. 1.2.

Nom	Connecteur	Expression
Dombi, [Dombi, 1982a]	$\top$	$\left(1+\left(\left(\frac{1-x}{x}\right)^{\gamma}+\left(\frac{1-y}{y}\right)^{\gamma}\right)^{1/\gamma}\right)^{-1}$
$\gamma \in [0,\infty]$	$\bot$	$1-\left(1+\left(\left(\frac{x}{1-x}\right)^{\gamma}+\left(\frac{y}{1-y}\right)^{\gamma}\right)^{1/\gamma}\right)^{-1}$
Frank, [Frank, 1979]	$\top$	$\log_{\gamma}\left(1+\frac{(\gamma^{x}-1)(\gamma^{y}-1)}{\gamma-1}\right)$
$\gamma \in [0,\infty]$	$\bot$	$1-\log_{\gamma}\left(1+\frac{(\gamma^{1-x}-1)(\gamma^{1-y}-1)}{\gamma-1}\right)$
Hamacher, [Hamacher, 1978]	$\top$	$\frac{x\,y}{\gamma+(1-\gamma)\,(x+y-x\,y)}$
$\gamma \in [0,\infty]$	$\bot$	$\frac{x+y+(\gamma-2)\,x\,y}{1+(\gamma-1)\,x\,y}$
Yager, [Yager, 1980]	$\top$	$\max\left(1-\left((1-x)^{\gamma}+(1-y)^{\gamma}\right)^{1/\gamma},0\right)$
$\gamma \in [0,\infty]$	$\bot$	$\min\left(\left(x^{\gamma}+y^{\gamma}\right)^{1/\gamma},1\right)$
Dubois-Prade, [Dubois and Prade, 1980]	$\top$	$\frac{xy}{\max(x,y,\gamma)}$
$\gamma \in [0,1]$	$\bot$	$1-\frac{(1-x)(1-y)}{\max((1-x),(1-y),\lambda)}$

Tab. 1.2: Couples de normes triangulaires paramétriques parmi les plus utilisés.

1.3.2 Combinaisons de normes triangulaires

Il existe également de nombreux opérateurs fondés sur les normes triangulaires, et où celles-ci sont combinées en fonction d'un élément neutre ou absorbant défini : uninormes et nullnormes, respectivement. Nous citerons aussi la classe des sommes symétriques, ainsi que les combinaisons convexes linéaires ou exponentielles. Ces opérateurs sont généralement construits dans le but d'obtenir un comportement dit de compensation (voir section 1.1.3) : les caractères conjonctifs et disjonctifs des normes triangulaires sont associés. Par mesure de concision, ces opérateurs sont détaillés en Annexe A.

1.4 Intégrales et mesures floues

1.4.1 Mesures floues

Le concept de mesure est très important en mathématiques, en particulier pour la théorie des intégrales. Ces mesures classiques sont supposées respecter l'additivité. Bien que cette propriété soit intéressante dans certains cas, il apparaît que dans des applications de la théorie de la décision, la théorie des jeux et l'intelligence artificielle, il devient indispensable de définir des mesures non-additives. Un exemple criant de ce genre de situation est le travail d'un groupe de personnes. Si l'on représente l'efficacité d'une personne par une mesure, il est évident que l'efficacité globale du groupe ne sera pas l'addition des différentes mesures mais dépendra bien des interactions entre les personnes.

Sugeno [Sugeno, 1974] propose donc de remplacer cette notion d'additivité par de la monotonie, et introduit ainsi les mesures floues, qui sont en fait des mesures non additives

monotones. On pourra tout de même observer que cette notion de mesure floue existait avant cette introduction sous des noms différents. En particulier Choquet [Choquet, 1954] les avait décrites sous le nom de *capacité*.

Définition 1.8. Une mesure floue (discrète) sur X est une fonction d'ensemble $\mu : \mathcal{P}(X) \to [0,1]$, où $\mathcal{P}(X)$ dénote l'ensemble des parties de X, ayant les propriétés suivantes :

$$\textit{(P1)} \quad \mu(A) \leq \mu(B) \text{ si } A \subseteq B \qquad \text{monotonie} \tag{1.41}$$

$$\textit{(P2)} \quad \mu(\emptyset) = 0 \qquad \text{condition aux bornes} \tag{1.42}$$

$$\textit{(P3)} \quad \mu(X) = 1 \qquad \text{condition aux bornes} \tag{1.43}$$

On notera que la propriété *(P3)* n'est pas indispensable pour définir une mesure floue, et une mesure respectant cette propriété sera dite *normale*. L'interprétation de cette mesure peut être assez variée, mais elle sert généralement à exprimer l'importance de la combinaison des critères $A \subseteq N$. Ceci permet ainsi de modéliser les diverses interactions entre critères et de leur donner un poids. Ainsi définies, les mesures floues requièrent la spécification de $2^n - 2$ coefficients. Afin de faciliter leur utilisation, plusieurs approches ont été considérées dans la littérature. En particulier, nous détaillerons ici les mesures restreintes, c'est à dire des mesures satisfaisant des contraintes supplémentaires, qui permettent de s'affranchir de spécifier l'ensemble de tous les $2^n - 2$ coefficients. Sugeno [Sugeno, 1974] introduit ainsi les λ-mesures floues. Celles-ci sont en fait des mesures normalisées et λ-additives. Si l'on pose $\lambda \in [-1,\infty]$, une manière d'obtenir, une fois les poids pour les singletons fixés, les mesures correspondant à l'union de deux sous-ensembles disjoints de critères A et B est

$$\mu(\{A,B\}) = \mu(\{A\}) + \mu(\{B\}) + \lambda\mu(\{A\})\mu(\{B\}) \tag{1.44}$$

Puisque cette mesure est normalisée, on a alors :

$$\lambda + 1 = \prod_{i=1}^{n} \Big(1 + \lambda\mu(\{x_i\})\Big) \tag{1.45}$$

Cette équation permet de déterminer la valeur de lambda correspondante. Dubois et Prade [Dubois and Prade, 1982b] proposent les mesures floues $\perp$-décomposables. Celles-ci respectent $\mu(\emptyset) = 0$ et $\mu(X) = 1$ et l'on calcule $\mu(\{A,B\})$ par l'intermédiaire d'une t-conorme; si $A \cap B = \emptyset$:

$$\mu(\{A,B\}) = \perp\Big(\mu(\{A\}), \mu(\{B\})\Big) \tag{1.46}$$

On peut ainsi obtenir des mesures de probabilités ($\perp_L$), possibilités ($\perp_M$) ou λ-mesures floues selon la t-conorme utilisée. Dans [Trillas and Alsina, 1993], Trillas et Alsina proposent une définition générale fondée sur des relations d'ordre $\prec$. Ils introduisent ainsi les $\prec$-mesures floues respectant *(P2)*, *(P3)* de la Définition 1.8 et :

$$\textit{(P1')} \qquad \text{si } x \prec y \text{ , alors } \mu(x) \leq \mu(y) \tag{1.47}$$

Plus tard, Grabisch [Grabisch, 1997] introduit les mesures floues k-additives. La représentation est cette fois-ci fondée sur une décomposition polynomiale de fonctions pseudo-booléennes. Dans ce cas, la définition complète d'une mesure k-additive requiert $\sum_{i=1}^{k} C_n^i$ coefficients. Une proposition visant encore à réduire la complexité de calcul des $2^n - 2$ coefficients consiste à supposer que tous les indices n'ont pas la même importance vis-à-vis de

la décision à prendre. Les auteurs [Miranda et al., 2002] introduisent ainsi les mesures floues p-symétriques. À partir de la notion d'ensembles d'indifférence, ils proposent une nouvelle équation de mise à jour où le nombre de coefficients nécessaires à la définition de la mesure floue dépend du nombre d'ensembles d'indifférence (c'est à dire p) et de leur cardinalité respective (dont la somme vaut n). Plus récemment, à partir du travail de Edwards [Edwards, 1953] sur les probabilités déformées[5], Narukawa et Torra proposent une nouvelle famille de mesures, voir [Narukawa and Torra, 2005]. Dans ce cadre, une mesure floue μ est une probabilité déformée si elle est représentée par une distribution de probabilité P sur X et une fonction f strictement croissante avec P.

À noter aussi que la théorie des fonctions de croyance et de plausibilité peut aussi rentrer dans ce cadre, puisque une fonction de masse est également une mesure floue avec une règle de mise à jour propre [Shafer, 1976]. Plus récemment encore, les *bi-capacités* ont été introduites afin de gérer les situations où l'on doit faire face à des échelles bi-polaires [Grabisch and Labreuche, 2005a;b], mais ceci sort du cadre de ce tour d'horizon. Nous terminerons ce paragraphe sur les mesures floues en précisant que des outils de caractérisation et d'interprétation existent, en particulier la transformée de Möbius, l'indice de Shapley [Shapley, 1953], et le concept d'interaction entre critères [Murofushi and Soneda, 1993; Grabisch, 1997].

1.4.2 Intégrale de Choquet

Une intégrale par rapport aux mesures floues (dans ce contexte, on parlait de *capacité*) qui viennent d'être évoquées est introduite par Choquet [Choquet, 1954]. Initialement vue dans le domaine continu où l'intégrale de Choquet de f par rapport à μ s'écrit $\int_0^1 \mu(\{x|f(x) \geq v\})dv$, elle sera ensuite proposée dans sa version discrète.

Définition 1.9. L'intégrale de Choquet discrète de $\mathbf{x} \in \mathbb{R}^n$ par rapport à une mesure floue μ est définie par

$$\mathcal{C}_\mu(\mathbf{x}) = \sum_{i=1}^{n} x_{(i)} \Big(\mu(A_{(i)}) - \mu(A_{(i+1)})\Big) \tag{1.48}$$

où $(\cdot)$ est une permutation sur N telle que $x_{(1)} \leq \cdots \leq x_{(n)}$. Par ailleurs, $A_{(i)}$ dénote l'ensemble $\{(i), \cdots, (n)\}$, et par convention $A_{(n+1)} = \emptyset$. On pourra écrire (1.48) de manière équivalente :

$$\mathcal{C}_\mu(\mathbf{x}) = \sum_{i=1}^{n} \mu(A_{(i)}) \Big(x_{(i)} - x_{(i-1)}\Big) \tag{1.49}$$

Un aspect particulièrement intéressant de l'intégrale de Choquet est qu'elle permet de généraliser un certain nombre d'opérateurs d'agrégation déjà évoqués, selon les mesures floues que l'on se donne. Ainsi, on pourra obtenir les opérateurs minimum, maximum, des statistiques d'ordre, la moyenne arithmétique et moyenne pondérée, et enfin la moyenne ordonnée pondérée OWA. Cette intégrale est monotone, continue et idempotente, et présente évidemment un comportement de compromis. Elle est de plus stable par transformation linéaire positive, elle est donc adaptée à l'agrégation de valeurs cardinales, à la différence de l'intégrale de Sugeno décrite ci-après.

5. traduction libre de "distorted probabilities"

1.4.3 Intégrale de Sugeno

En 1974, Sugeno introduit les concepts de mesure floue et d'intégrales floues, intégrales classiques au sens de Lebesgue mais par rapport à une mesure floue [Sugeno, 1974] :

Définition 1.10. L'intégrale de Sugeno discrète de $\mathbf{x} \in \mathbb{R}^n$ par rapport à une mesure floue μ est définie par

$$\mathcal{S}_\mu(\mathbf{x}) = \bigvee_{i=1}^{n} \Big(x_{(i)} \wedge \mu(A_{(i)}) \Big) \tag{1.50}$$

où $(\cdot)$ est toujours une permutation sur N telle que $x_{(1)} \leq \cdots \leq x_{(n)}$, $A_{(i)}$ dénote l'ensemble $\{(i), \cdots, (n)\}$, et par convention $A_{(n+1)} = \emptyset$.

Cette définition utilise les opérateurs minimum et maximum, mais plusieurs auteurs ont proposé l'utilisation de t-normes, si bien que la définition suivante parait importante.

Définition 1.11. La quasi-intégrale de Sugeno discrète de $\mathbf{x} \in \mathbb{R}^n$ par rapport à une mesure floue μ est définie par

$$\mathcal{S}_\mu(\mathbf{x}) = \bigvee_{i=1}^{n} \Big(x_{(i)} \top \mu(A_{(i)}) \Big) \tag{1.51}$$

où $(\cdot)$ est une permutation sur N telle que $x_{(1)} \leq \cdots \leq x_{(n)}$. Par ailleurs, $A_{(i)}$ dénote l'ensemble $\{(i), \cdots, (n)\}$, et par convention $A_{(n+1)} = \emptyset$.

De la même manière que l'intégrale de Choquet généralisait des opérateurs d'agrégation assez variés, l'intégrale de Sugeno permet elle aussi d'obtenir d'autres opérateurs d'agrégation. Parmi eux, on recensera les opérateurs minimum, maximum, des statistiques d'ordre, et enfin les minimum et maximum pondérés. Cette intégrale est monotone, continue et idempotente, et présente évidemment un comportement de compromis. La souplesse de la représentation de l'information grâce aux intégrales floues a conduit à son utilisation dans de nombreux domaines, et en particulier en décision multi-critères [Grabisch and Labreuche, 2008], décision dans l'incertain [Sabbadin, 1998], et en reconnaissance de formes [Bezdek et al., 1999b; Grabisch, 2000b]. Cette intégrale de Sugeno est mieux adaptée à l'agrégation de valeurs ordinales. Ceci implique donc que pour combiner les mesures floues et les valeurs de la fonction, il faut qu'elles appartiennent au même sous-domaine, et doivent dans un sens dénoter le même concept [Torra and Narukawa, 2006]. Puisque les mesures floues modélisent une notion d'importance, de confiance dans les sources, ceci devrait être le cas pour f. En somme, l'intégrale de Sugeno parait appropriée lorsque l'utilisateur veut obtenir une mesure de confiance globale du système. Enfin, nous noterons que cette intégrale procède par saturation : elle recherche l'importance (confiance) dépassant un certain degré, puis réalise une opération de compromis entre les valeurs sélectionnées. Un autre point de vue consiste à voir cette combinaison comme la disjonction de conjonctions, en d'autres termes obtenir le couple fonction-mesure floue maximum. À chaque fois qu'il faut gérer des situations incertaines, où des réponses discordantes peuvent apparaître, et qu'il faut prendre une décision en tenant compte de ces prises de position, on retrouve le concept d'intégrale, qui permet de modéliser la situation via l'introduction de mesures floues adaptées. C'est ainsi qu'elles sont utilisées en classification, traitement d'image [Tahani and Keller, 1990; Shi et al., 1998], fusion de classifieurs [Kuncheva, 2004; Temko et al., 2008], systèmes d'inférence floue [Torra and Narukawa, 2006], etc ...

1.4.4 $\perp$-intégrales

Murofushi et Sugeno proposent une généralisation des intégrales de Choquet et Sugeno, et construisent ainsi une intégrale fondée sur un couple de t-conormes et une mesure floue [Murofushi and Sugeno, 1991]:

Définition 1.12. Soient deux t-conormes ayant pour générateurs les fonctions h et g, avec $g(1) = 1$, et μ une mesure floue, la $\perp$-intégrale de $\mathbf{x}$ par rapport à μ est définie par

$$\mathcal{F}_\mu(\mathbf{x}) = h^{-1}\Big(\mathcal{C}_{g\circ\mu}(h \circ x)\Big) \tag{1.52}$$

Depuis, d'autres propositions ont vu le jour. Nous citerons à titre d'exemple les travaux de Benvenuti et Vivona [Benvenuti and Vivona, 1996]. Leur intégrale se fonde sur deux opérations binaires $\oplus$ et $\odot$. Pour une constante $a \in \mathbb{R}^{*+}$, l'opération $\oplus : [0, a]^2 \to [0, a]$ est supposée être une t-conorme. Pour une autre constante $b \in \mathbb{R}^{*+}$, l'opération $\odot : [0, a] \times [0, b] \to [0, a]$ est distributive à droite par rapport à $\oplus$, c'est à dire que pour tout x,$y \in [0, a]$ et $z \in [0, b]$, on a

$$(x \oplus y) \odot z = (x \odot z) \oplus (y \odot z) \tag{1.53}$$

De plus, il est défini une troisième opération binaire $\ominus$ associée à $\oplus$ de la manière suivante

$$x \ominus y = inf\{t \in [0, a] | y \oplus t \geq x\} \tag{1.54}$$

Définition 1.13. L'intégrale de Benvenuti discrète de $\mathbf{x} \in \mathbb{R}^n$ par rapport à une mesure floue μ est définie par

$$\mathcal{B}_\mu = \bigoplus_{i=1}^{n} \mu(A_{(i)}) \odot \Big(x_{(i)} \ominus x_{(i-1)}\Big) \tag{1.55}$$

Selon les opérations binaires utilisées, on retrouvera successivement les intégrales de Choquet, de Sugeno ou de Shilkret [Shilkret, 1971]. Plus récemment, Narukawa et Torra [Narukawa and Torra, 2009] proposent à leur tour une intégrale floue, généralisation de la proposition de Benvenuti et Vivona, car étendue au cas multidimensionnel. On pourra se référer à [Mesiar and Mesiarova, 2008] pour une revue récente sur les intégrales floues.

1.5 Similarité d'ordre

Initialement introduit dans [Frélicot et al., 2004], le concept de OU flou d'ordre est présenté, mais restreint à l'ordre 2. Celui-ci est défini par

$$\overset{2}{\perp}(\mathbf{u}) = \underset{i=1,c}{\top}\left(\underset{j=1,c;j\neq i}{\perp} u_j\right) \tag{1.56}$$

où $\mathbf{u}$ est un vecteur de valeurs dans $[0, 1]$ de taille c. Cet opérateur respecte un certain nombre de propriétés, telles que la symétrie, conditions aux bornes, idempotence, continuité de l'opérateur avec la continuité du couple $(\top, \perp)$. Dans le cas particulier ou les normes triangulaires min max sont utilisées, $\overset{2}{\perp}(\mathbf{u})$ est exactement la deuxième plus grande valeur de $\mathbf{u}$. Cet opérateur présente également un certain comportement de compensation (faible), dans la mesure où

$$\top(\mathbf{u}) \leq \overset{2}{\perp}(\mathbf{u}) \leq \perp(\mathbf{u})$$

Les auteurs, en introduisant cet opérateur, désirent savoir si au moins les deux plus grandes valeurs sont comparables. Dans de nombreuses situations, on est amené à considérer non seulement la deuxième plus grande valeur, mais aussi les valeurs inférieures, qui peuvent apporter de précieuses informations. Une extension naturelle de cet opérateur consiste donc à considérer l'opérateur $\overset{k}{\perp}(\mathbf{u})$, $k = 1,c$, [Mascarilla et al., 2008] dont nous allons détailler les propriétés car certaines de nos propositions en découlent.

Définition 1.14 ([Mascarilla et al., 2008],[Le Capitaine and Frélicot, 2009c])**.** Soit $\mathcal{P}$ l'ensemble de tous les sous-ensembles de $C = \{1, \cdots, c\}$, et $\mathcal{P}_k = \{A \in \mathcal{P} : |A| = k\}$, où $|A|$ dénote le cardinal du sous-ensemble A. Le OU flou d'ordre k est défini pour tout couple $(\top, \perp)$ par

$$\underset{i=1,\,c}{\overset{k}{\perp}} u_i = \underset{A\in\mathcal{P}_{k-1}}{\top} \left(\underset{j\in C\setminus A}{\perp} u_j \right) \tag{1.57}$$

Proposition 1.1 ([Mascarilla et al., 2008])**.** *L'opérateur défini par (1.57) est un opérateur d'agrégation satisfaisant les conditions aux bornes (1.2-1.3), la monotonie (1.4) et la symétrie (A.2).*

Démonstration.
Venant des propriétés d'éléments neutres de $\top$ et $\perp$, les égalités suivantes sont immédiates

$$\underset{i=1,c}{\overset{k}{\perp}}(\mathbf{0}) = 0 \qquad \text{et} \qquad \underset{i=1,c}{\overset{k}{\perp}}(\mathbf{1}) = 1$$

Par monotonie de $\top$ et $\perp$, on a facilement

$$\underset{i=1,c}{\overset{k}{\perp}} u_i \leq \underset{i=1,c}{\overset{k}{\perp}} w_i$$

si $u_i \leq w_i$ pour tout $i = 1, \cdots, c$.
Enfin, encore par propriétés de $\top$ et $\perp$ (ici la symétrie), on obtient

$$\underset{i=1,c}{\overset{k}{\perp}} u_i = \underset{i=1,c}{\overset{k}{\perp}} u_{\sigma(i)}$$

voir (A.2), pour la définition. □

On soulignera la rôle de la valeur 0 si elle est présente parmi les éléments à agréger.

Proposition 1.2 ([Mascarilla et al., 2008])**.** *Calculer (1.57) sur un vecteur* $\mathbf{u} = (u_1, \cdots, u_{c-1}, 0)$ *revient à évaluer le OU flou de même ordre k sur ce vecteur privé de* 0, $\mathbf{u}' = (u_1, \cdots, u_{c-1})$ *et sa conjonction avec le OU flou d'ordre* $k-1$ *de* $\mathbf{u}'$*.*

Démonstration.
On peut écrire

$$\overset{k}{\perp}(u_1, \cdots, u_{c-1}, 0) = \left(\underset{A\in\mathcal{P}_{k-1}, c\in A}{\top} \left(\underset{j\in C\setminus A}{\perp} u_j \right) \right) \top \left(\underset{A\in\mathcal{P}_{k-1}, c\notin A}{\top} \left(\underset{j\in C\setminus A}{\perp} u_j \right) \right)$$

Soit $C' = \{1, \cdots, c-1\}$, et $\mathcal{P}'_k$ l'ensemble des sous-ensembles de cardinal k associés.

$$\begin{aligned}
\overset{k}{\perp}(u_1, \cdots, u_{c-1}, 0) &= \left(\underset{B\in\mathcal{P}'_{k-2}}{\top}\left(\underset{j\in C\setminus(B\cup\{c\})}{\perp} u_j\right)\right) \top \left(\underset{A\in\mathcal{P}'_{k-1}}{\top}\left(\underset{j\in C'\setminus A}{\perp} u_j\right)\right) \\
&= \left(\underset{B\in\mathcal{P}'_{k-2}}{\top}\left(\underset{j\in C'\setminus B}{\perp} u_j\right)\right) \top \left(\overset{k}{\perp}(u_1, \cdots, u_{c-1})\right) \\
&= \left(\overset{k-1}{\perp}(u_1, \cdots, u_{c-1})\right) \top \left(\overset{k}{\perp}(u_1, \cdots, u_{c-1})\right)
\end{aligned}$$

□

Exemple 1.5. Considérons les cas particuliers $\overset{1}{\perp}$ et $\overset{c}{\perp}$. A partir de (1.57), on peut réécrire

$$\begin{aligned}
\overset{1}{\perp}(\mathbf{u}) &= \underset{A\in\mathcal{P}_0}{\top}\left(\underset{j\in C\setminus A}{\perp} u_j\right) \\
&= \underset{\emptyset}{\top}\left(\underset{j\in C}{\perp} u_j\right) \\
&= \underset{j\in C}{\perp} u_j \\
&= \perp(\mathbf{u})
\end{aligned}$$

Le OU flou d'ordre 1 est donc le OU flou usuel défini par une t-conorme sur un c-uplet. De la même manière, on peut réécrire

$$\begin{aligned}
\overset{c}{\perp}(\mathbf{u}) &= \underset{A\in\mathcal{P}_{c-1}}{\top}\left(\underset{j\in C\setminus A}{\perp} u_j\right) \\
&= \underset{C\setminus\{i\}, i=1,c}{\top} u_i \\
&= \underset{j\in C}{\top} u_j \\
&= \top(\mathbf{u})
\end{aligned}$$

Donc le OU flou d'ordre c est le ET flou usuel défini par une t-norme sur un c-uplet.

A partir de cet exemple, on pourra noter que, selon la valeur donnée à k, $\overset{k}{\perp}$ est un opérateur conjonctif, de compromis (et/ou compensation), ou disjonctif. Le proposition suivante relie les statistiques d'ordre à l'opérateur présenté.

Proposition 1.3 ([Mascarilla et al., 2008])**.** *Si l'on prend le couple* $(\min, \max)$*, alors* $\overset{k}{\perp}_M(\mathbf{u}) = u_{(k)}$*, où* $u_{(k)}$ *dénote la k-ième plus grande valeur de* $\mathbf{u}$ *.*

Démonstration.
On peut écrire

$$\overset{k}{\perp}(\mathbf{u}) = u_{(k)} = \left(\underset{j\in C\setminus\kappa}{\perp} u_j\right) \top \left(\underset{A\in\mathcal{P}_{k-1\setminus\kappa}}{\top}\left(\underset{j\in C\setminus A}{\perp} u_j\right)\right)$$

où $\kappa = \{1, \cdots, k-1\}$. Si $A \in \mathcal{P}_{k-1\setminus\kappa}$, alors il existe i_0 dans $\kappa \cap (C\setminus A)$, et donc

$$\underset{j\in C\setminus A}{\perp} u_j \geq u_{i_0} \geq u_{(k)}$$

Comme on sait que $\underset{j\in C\setminus\kappa}{\perp} u_j = u_{(k)}$ et que $\underset{A\in\mathcal{P}_{k-1}\setminus\kappa}{\top}\left(\underset{j\in C\setminus A}{\perp} u_j\right) \geq u_{(k)}$, ceci conclut la preuve. □

Bien que l'on ne puisse déterminer une borne pour chacun des opérateurs, où k varie de 1 à c, on peut néanmoins donner une borne sur k pour que la monotonie soit respectée.

Proposition 1.4 ([Mascarilla et al., 2008])**.** *Soit $b = c/2+1$ si c est pair, et $b = (c+3)/2$ si c est impair. On a alors, pour tout couple $(\top, \perp)$*

$$\perp(\mathbf{u}) = \overset{1}{\underset{i=1,c}{\perp}} u_i \geq \overset{2}{\underset{i=1,c}{\perp}} u_i \geq \cdots \geq \overset{b}{\underset{i=1,c}{\perp}} u_i \tag{1.58}$$

Démonstration.
Soit $2 \leq k \leq b$, on montre que $\overset{k-1}{\perp}(\mathbf{u}) \geq \overset{k}{\perp}(\mathbf{u})$. Par définition (1.57),

$$\begin{aligned}\overset{k-1}{\perp}(\mathbf{u}) &= \underset{A\in\mathcal{P}_{k-2}}{\top}\left(\underset{j\in C\setminus A}{\perp} u_j\right)\\ &= \underset{\{i_1,\cdots,i_{c-k+2}\}}{\top}(u_{i_1}\perp\cdots\perp u_{i_{c-k+2}})\end{aligned}$$

D'autre part,

$$\overset{k}{\perp}(\mathbf{u}) = \underset{\{j_1,\cdots,j_{c-k+1}\}}{\top}(u_{j_1}\perp\cdots\perp u_{j_{c-k+1}})$$

Comme on sait que $C_c^{k-2} \leq C_c^{k-1}$, en utilisant une partition de $\mathcal{P}$ en chaînes symétriques [Bruijn et al., 1951], et en construisant une injection τ de $\mathcal{P}_{c-k+2}$ dans $\mathcal{P}_{c-k+1}$ telle que $\tau(A) \subset A$ pour tout A dans $\mathcal{P}_{c-k+2}$, on écrit $\tau(\{i_1,\cdots,i_{c-k+2}\}) = \{l_1,\ldots,l_{c-k+1}\}$. Par monotonie de $\perp$, on obtient

$$\overset{k-1}{\perp}(\mathbf{u}) = \underset{\{l_1,\cdots,l_{c-k+1}\}}{\top}(u_{l_1}\perp\cdots\perp u_{l_{c-k+1}})$$

et

$$\overset{k}{\perp}(\mathbf{u}) \leq \underset{\{l_1,\cdots,l_{c-k+1}\}}{\top}(u_{l_1}\perp\cdots\perp u_{l_{c-k+1}})$$

ce qui conclut la preuve. □

L'opérateur ainsi défini est une mesure qui permet de déterminer si les k plus grandes valeurs sont fortes (et donc dans un certain sens similaires). Afin d'observer le comportement de celui-ci, prenons 4 vecteurs $\mathbf{u}$ dénotant des situations différentes (aussi bien en terme de décision multi-critères que de reconnaissance de formes) :

- $\mathbf{u}_1 = (0.70, 0.10, 0.85, 0.80)$
- $\mathbf{u}_2 = (0.20, 0.10, 0.85, 0.80)$
- $\mathbf{u}_3 = (0.20, 0.10, 0.85, 0.15)$
- $\mathbf{u}_4 = (0.20, 0.10, 0.05, 0.15)$

Dans la mesure où le choix des couples de normes triangulaires est quasiment infini, nous avons choisi de reporter les résultats des couples que nous estimons les plus représentatifs, voir la discussion à ce sujet dans la suite de cette section. Les valeurs successives de l'opérateur pour $k = 1,2,3,4$ et pour les différents $\mathbf{u}_i$ sont données dans les Tab. 1.3, 1.4, 1.5 et 1.6, respectivement.

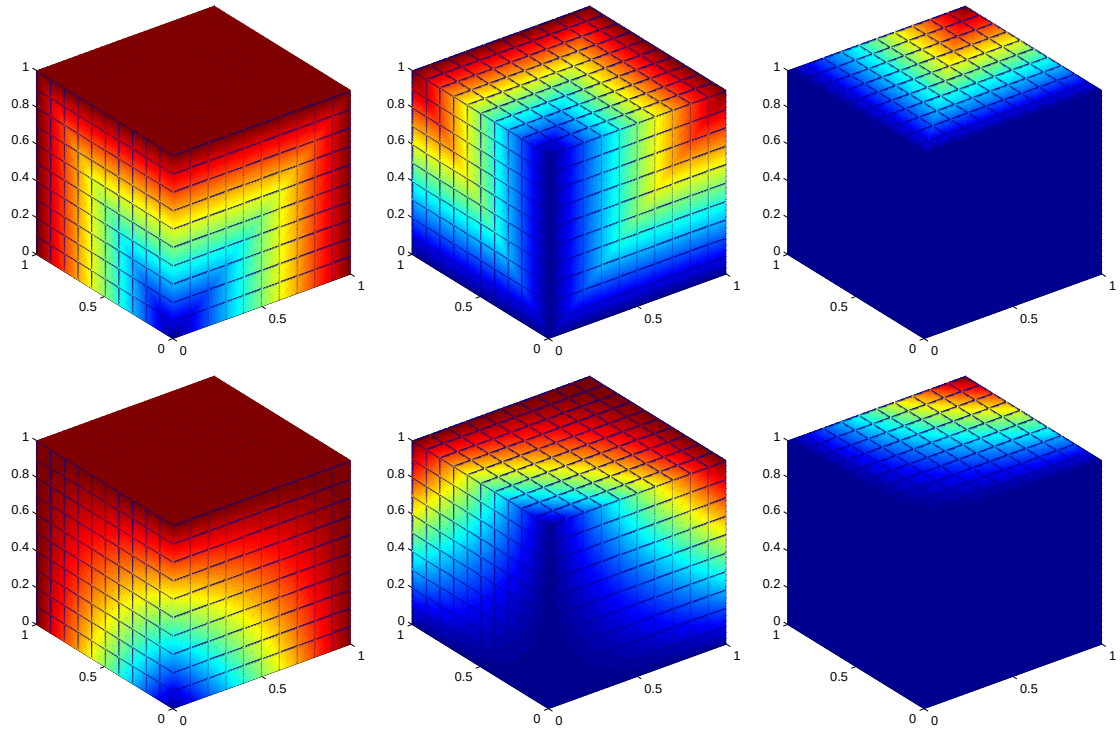

FIG. 1.4: *Haut* : Valeurs de sortie du OU flou d'ordre k, pour k = 1,2,3 (respectivement *gauche*, *centre*, *droite*), avec le couple (min , max) pour l'ensemble des vecteurs $\mathbf{u} \in [0,1]^3$. La couleur rouge dénote la valeur la plus haute, 1, et la couleur bleue la plus faible, 0. *Bas* : Valeurs de sortie du OU flou d'ordre k, pour k = 1,2,3 (respectivement *gauche*, *centre*, *droite*), avec le couple produit .

$\mathbf{u}_1 = (0.70, 0.10, 0.85, 0.80)$	$k =$	1	2	3	4
$\overset{k}{\perp}$	$(\top,\perp)_M$	0.85	0.80	0.70	0.10
	$(\top,\perp)_P$	0.99	0.87	0.45	0.04
	$(\top,\perp)_L$	1	1	0.65	0
	$(\top,\perp)_D$	1	0	0	0
	$(\top,\perp)_{H_0}$	0.92	0.68	0.45	0.09
	$(\top,\perp)_{D_2}$	0.87	0.74	0.62	0.10
	$(\top,\perp)_{Y_3}$	1	0.94	0.66	0.08
	$(\top,\perp)_{DP_{0.5}}$	0.96	0.88	0.70	0.10
	$(\top,\perp)_{F_{0.5}}$	0.98	0.84	0.45	0.05

TAB. 1.3: Valeur du OU flou d'ordre k pour k = 1,2,3,4, sur $\mathbf{u}_1$. Les couples utilisés sont successivement Standard, Algébrique, Lukasiewicz, Drastique, Hamacher ($\gamma = 0$), Dombi ($\gamma = 2$), Yager ($\gamma = 2$), Dubois-Prade ($\gamma = 0.5$) et Frank ($\gamma = 0.5$).

$\mathbf{u}_2 = (0.20, 0.10, 0.85, 0.80)$	$k =$	1	2	3	4
$\overset{k}{\perp}$	$(\top, \perp)_M$	0.85	0.80	0.20	0.10
	$(\top, \perp)_P$	0.97	0.72	0.14	0.01
	$(\top, \perp)_L$	1	1	0.15	0
	$(\top, \perp)_D$	1	0	0	0
	$(\top, \perp)_{H_0}$	0.90	0.62	0.21	0.06
	$(\top, \perp)_{D_2}$	0.87	0.73	0.21	0.09
	$(\top, \perp)_{Y_3}$	1	0.78	0.19	0
	$(\top, \perp)_{DP_{0.5}}$	0.94	0.80	0.20	0.04
	$(\top, \perp)_{F_{0.5}}$	0.97	0.70	0.15	0.02

TAB. 1.4: Valeur du OU flou d'ordre k pour k = 1,2,3,4, sur $\mathbf{u}_2$. Les couples utilisés sont successivement Standard, Algébrique, Lukasiewicz, Drastique, Hamacher ($\gamma = 0$), Dombi ($\gamma = 2$), Yager ($\gamma = 2$), Dubois-Prade ($\gamma = 0.5$) et Frank ($\gamma = 0.5$).

$\mathbf{u}_3 = (0.20, 0.10, 0.85, 0.15)$	$k =$	1	2	3	4
$\overset{k}{\perp}$	$(\top, \perp)_M$	0.85	0.20	0.15	0.10
	$(\top, \perp)_P$	0.90	0.27	0.01	0.00
	$(\top, \perp)_L$	1	0.45	0	0
	$(\top, \perp)_D$	1	0	0	0
	$(\top, \perp)_{H_0}$	0.86	0.29	0.10	0.05
	$(\top, \perp)_{D_2}$	0.85	0.24	0.12	0.08
	$(\top, \perp)_{Y_3}$	0.85	0.22	0	0
	$(\top, \perp)_{DP_{0.5}}$	0.85	0.20	0.02	0.01
	$(\top, \perp)_{F_{0.5}}$	0.89	0.27	0.02	0.00

TAB. 1.5: Valeur du OU flou d'ordre k pour k = 1,2,3,4, sur $\mathbf{u}_3$. Les couples utilisés sont successivement Standard, Algébrique, Lukasiewicz, Drastique, Hamacher ($\gamma = 0$), Dombi ($\gamma = 2$), Yager ($\gamma = 2$), Dubois-Prade ($\gamma = 0.5$) et Frank ($\gamma = 0.5$).

$\mathbf{u}_4 = (0.20, 0.10, 0.05, 0.15)$	$k =$	1	2	3	4
$\overset{k}{\perp}$	$(\top, \perp)_M$	0.20	0.15	0.10	0.05
	$(\top, \perp)_P$	0.42	0.01	0.00	0.00
	$(\top, \perp)_L$	0.5	0	0	0
	$(\top, \perp)_D$	1	0	0	0
	$(\top, \perp)_{H_0}$	0.37	0.09	0.04	0.03
	$(\top, \perp)_{D_2}$	0.24	0.12	0.07	0.04
	$(\top, \perp)_{Y_3}$	0.23	0	0	0
	$(\top, \perp)_{DP_{0.5}}$	0.20	0.00	0.00	0.00
	$(\top, \perp)_{F_{0.5}}$	0.40	0.01	0.00	0.00

TAB. 1.6: Valeur du OU flou d'ordre k pour k = 1,2,3,4, sur $\mathbf{u}_4$. Les couples utilisés sont successivement Standard, Algébrique, Lukasiewicz, Drastique, Hamacher ($\gamma = 0$), Dombi ($\gamma = 2$), Yager ($\gamma = 2$), Dubois-Prade ($\gamma = 0.5$) et Frank ($\gamma = 0.5$).

À la lecture de ces tableaux, plusieurs choses apparaissent. D'abord, on vérifie que l'utilisation du couple standard $(\top, \bot)_M$ mène à la statistique d'ordre k. On retrouve également le résultat de l'exemple 1.5, ainsi que les propriétés d'ordre sur les différents couples : on a vu que $\overset{1}{\bot} = \bot$ et que $\overset{c}{\bot} = \top$. Or, on sait, par exemple, que $\top_M \geq \top_P \geq \top_L \geq \top_D$, ainsi que $\bot_M \leq \bot_P \leq \bot_L \leq \bot_D$. On retrouve cet ordre dans les valeurs du tableau, en observant les colonnes correspondant à $k = 1$ et $k = 4$. Comme l'ordre change entre $k = 1$ et $k = c$ pour différents couples, par exemple ($\underset{M}{\overset{2}{\bot}}\mathbf{u}_2 > \underset{P}{\overset{2}{\bot}}\mathbf{u}_2$, alors que $\underset{M}{\overset{2}{\bot}}\mathbf{u}_3 < \underset{P}{\overset{2}{\bot}}\mathbf{u}_3$), on ne peut pas retrouver une notion d'ordre parmi les k intermédiaires pour les différents couples comme on pouvait le faire dans le cas précédent ($k = 1$ et $k = c$).

D'autre part, comme attendu, les valeurs de sortie sont élevées lorsque exactement k valeurs sont élevées (et donc similaires), et ce, quel que soit le choix du couple $(\top, \bot)$. Le cas où $k = 1$ est un peu particulier, dans le sens où l'opérateur cherche à déterminer si une seule valeur est forte. Ceci se fait de manière indépendante des autres valeurs dans le cas de (min , max), mais conjointement pour tout autre choix de couple strictement monotone. On voit également que porter son choix sur des couples comme $(\top, \bot)_D$ ou $(\top, \bot)_L$ ne sera que très peu utilisable en pratique, comme on pouvait raisonnablement l'attendre.

Enfin, on observe que lorsque k valeurs sont similaires, alors c'est uniquement le OU flou d'ordre k qui est modifié de manière significative, c'est à dire que pour $l = 1, \cdots k-1$, ainsi que pour $l = k+1, \cdots c$, les valeurs associées ($\underset{l}{\bot}(\mathbf{u})$) restent pratiquement inchangées. Ceci montre ainsi la faculté de l'opérateur à localiser correctement une similarité d'ordre donné.

L'opérateur OU flou d'ordre k évalue les k plus grandes valeurs de $\mathbf{u}$. Dans d'autres circonstances, on peut être intéressé par les k plus petites valeurs de $\mathbf{u}$, et nous proposons le ET flou d'ordre k comme l'opérateur dual du OU flou d'ordre k.

Définition 1.15. Soit $\overset{k}{\bot}(\mathbf{u})$ le OU flou d'ordre k, alors le ET flou d'ordre k est défini par dualité

$$\overset{k}{\top}(\mathbf{u}) = 1 - \overset{k}{\bot}(1 - \mathbf{u}) \tag{1.59}$$

Proposition 1.5. *On peut écrire* $\overset{k}{\top}(\mathbf{u})$ *sous la forme*

$$\overset{k}{\top}(\mathbf{u}) = \underset{A \in \mathcal{P}_{k-1}}{\bot} \left(\underset{j \in C \setminus A}{\top} u_j \right) \tag{1.60}$$

Démonstration.
On a

$$\begin{aligned} \overset{k}{\top}(\mathbf{u}) &= 1 - \overset{k}{\bot} 1 - (\mathbf{u}) \\ &= 1 - \underset{A \in \mathcal{P}_{k-1}}{\top} \left(\underset{j \in C \setminus A}{\bot} 1 - u_j \right) \end{aligned}$$

Par propriété de dualité (1.32) puis (1.33), il s'en suit

$$\begin{aligned}\overset{k}{\top}(\mathbf{u}) &= 1 - \underset{A\in\mathcal{P}_{k-1}}{\top}\left(1 - \underset{j\in C\setminus A}{\top} u_j\right)\\ &= 1 - \left(1 - \underset{A\in\mathcal{P}_{k-1}}{\bot}\left(\underset{j\in C\setminus A}{\top} u_j\right)\right)\\ &= \underset{A\in\mathcal{P}_{k-1}}{\bot}\left(\underset{j\in C\setminus A}{\top} u_j\right)\end{aligned}$$

□

L'opérateur $\overset{k}{\top}(\mathbf{u})$ partage les mêmes propriétés duales que $\overset{k}{\bot}(\mathbf{u})$ de conditions aux bornes, monotonie et symétrie. Les propriétés duales des propositions 1.2–1.3–1.4 peuvent également être démontrées; elles ne sont pas données car les preuves sont pratiquement identiques. Un seul exemple cependant : dans le cas de l'utilisation de $(\top,\bot)_M$, $\underset{i=1,c}{\overset{k}{\top}}(\mathbf{u})$ est exactement la k-ième plus petite valeur de $\mathbf{u}$. Si l'on prend $k = 1$ et $k = c$, on obtient respectivement les opérateurs de t-norme et t-conorme.

1.6 Conclusion

Dans ce chapitre, nous avons donc présenté un aperçu général des opérateurs d'agrégation. Cet aperçu ne peut évidemment être complet, tant les outils disponibles sont nombreux. A ce titre, nous suggérons au lecteur de se référer à [Grabisch et al., 2009]. D'autres opérateurs, tels que les *copules*, n'ont pas été mentionnés dans ce chapitre, mais mériteraient qu'on leur porte attention dans un avenir proche, dans la mesure où ils permettent de modéliser des unions et intersections de distributions de probabilités, et sont proches des normes triangulaires. Une introduction claire et concise à ce sujet pourra être trouvée dans le livre de Nelsen [Nelsen, 2006].

Nous avons proposé un nouveau couple d'opérateur de similarité d'ordre, le OU flou d'ordre, et son opération duale, le ET flou d'ordre. C'est une généralisation des statistiques d'ordre k à l'aide d'une combinaison de normes triangulaires. Les applications potentielles de cet opérateurs sont multiples. En effet, déterminer si k valeurs parmi n sont fortes et similaires revient dans de nombreux problèmes. Nous citerons, sans être exhaustif, le filtrage médian en traitement d'image, la sélection de valeurs propres (composantes principales), le rejet en classification, ou encore l'appariement d'images.

Comme on a pu le voir tout au long de ce chapitre, les opérateurs d'agrégation ont pour but, avec une seule valeur, de représenter au mieux l'ensemble des valeurs d'entrée. Ceci ne se fait évidemment pas sans une perte d'information. L'objectif est alors de minimiser cette perte d'information, de manière à rendre cet opérateur pertinent. Une façon de minimiser cette perte est de déterminer les éléments caractéristiques, et donc discriminants, d'une série de valeurs. L'apparition des mesures floues, puis des intégrales floues qui leur sont associées, aura permis un pouvoir de modélisation très important, puisque les interactions entre valeurs sont déterminées de cette manière. Cette puissance a évidemment quelques défauts : elle requiert la spécification de nombreuses valeurs pour la définition de la mesure.

Afin de remédier à ce problème, de nombreuses solutions ont été proposées pour diminuer le nombre de valeurs nécessaires.

Dans le chapitre suivant, nous aborderons les notions de similarité entre valeurs d'entrée et/ou de sortie d'opérateurs d'agrégation, ou plus simplement la similarité de vecteurs que l'on pourrait soumettre à un opérateur d'agrégation. On verra que les opérateurs d'agrégation tiennent deux rôles distincts lorsque l'on désire évaluer la compatibilité d'observations. Le premier type de mesures de compatibilité utilisant des opérateurs d'agrégation vise à déterminer les caractères communs de deux objets par le biais d'opérations ensemblistes telles que l'intersection ou l'union. Dans l'autre cas, l'opérateur d'agrégation est utilisé sur les mesures de compatibilité : un vecteur de compatibilité est évalué par un opérateur d'agrégation afin de déterminer certaines particularités, ou tout simplement pour obtenir une valeur scalaire de compatibilité. Nous nous intéresserons principalement à la deuxième situation dans le chapitre qui suit.

Chapitre 2

Similarité de valeurs numériques dans [0,1]

... the similarity of two simple qualities may consist in the slightness difference that exists between them.

— Oswald Külpe
Outlines of psychology (1895)

Résumé : *Après avoir introduit le concept de similarité ainsi que les notions voisines et les différentes définitions qui s'y rapportent, le cœur de ce chapitre regroupe nos différentes propositions pour l'évaluation de différents types de similarités : similarité par blocs, similarité par une approche logique. Chaque proposition sera agrémentée d'exemples numériques sur des vecteurs ayant valeur dans l'intervalle unité, ainsi qu'une visualisation dans l'hypercube unité. Différents exemples d'application de ces opérateurs sont enfin proposés pour conclure ce chapitre.*

2.1 Introduction

Alors que nous avons utilisé jusqu'à maintenant la notation $\mathbf{x} = (x_1, \cdots, x_n)$ pour la description des opérateurs d'agrégation, nous utiliserons indifféremment $\mathbf{u} = (u_1, \cdots, u_c)$. Ce changement de notation est dû au fait que les principales applications auxquelles nous nous intéressons utilisent la notation $\mathbf{u}$ pour qualifier les degrés d'appartenance aux classes, et que ce nombre de classes est souvent dénoté c.

La similarité entre objets est sans doute la plus utilisée, mais aussi la plus difficile à quantifier des mesures de compatibilité [Cross and Sudkamp, 2002]. Cette difficulté est liée au fait que la similarité ne possède pas une définition précise autre qu'intuitive. Les notions de *similarité* et de *différence* sont très liées, certains auteurs allant jusqu'à dire qu'une similarité n'est rien d'autre qu'une faible différence entre deux objets, [Külpe, 1895]. Notons toutefois que définir la *dissimilarité* comme le complément strict de la similarité fait légitimement débat [Dubois and Prade, 1982a]. La suite du chapitre est organisée de la manière suivante. Dans un premier temps, nous donnons et rappelons quelques notions, définitions qui sont liées au concept de similarité, et créons le lien avec les mesures d'ambiguïté, d'incertitude,

et d'imprécision. Nous proposons ensuite de nouveaux opérateurs et de nouvelles mesures fondées sur la combinaison de normes triangulaires. Ces propositions seront illustrées par des exemples numériques ainsi que par des visualisations. Nous donnons enfin quelques applications possibles de ces propositions, et concluons en donnant des perspectives de travail.

Lorsque l'on compare deux objets, on désire mettre en avant les relations, et les différences qui existent entre eux. Ceci peut se formuler grâce à des degrés de similarité caractérisant leurs différences et points communs. Les mesures permettant cette comparaison seront citées sous le terme général **mesures de compatibilité**, parmi lesquelles on trouvera les mesures d'inclusion, de distance, de similarité [Cross and Sudkamp, 2002; Della Riccia et al., 2008]. Nous sommes donc bien en présence de **deux** objets dont on souhaite estimer la compatibilité. Inversement, nous aborderons le thème des **mesures de caractérisation**, comprenant les mesures d'incertitude, spécificité, ambiguïté [1], imprécision, entropie, etc ... Dans ce cas, nous souhaitons caractériser, pour **un** objet les différentes notions précédemment énoncées. Bien évidemment, nous verrons qu'il existe des liens entre ces deux types de mesure.

2.2 Mesures de caractérisation

2.2.1 Mesures entropiques et d'incertitude

Les ensembles flous permettent une représentation aisée de l'imprécision, mais l'évaluation de l'incertitude d'un ensemble flou est quelque chose qui dépend fortement de ce que l'on souhaite mesurer. Selon certains auteurs, une distinction doit être faite entre ambiguïté et imprécision [Cross and Sudkamp, 2002]. On peut interpréter l'entropie comme une notion d'ordre, ou de structuration, c'est à dire que si l'entropie augmente, alors le désordre augmente aussi. Cette définition est clairement liée à la divergence de Kullback-Leibler, [Kullback and Leibler, 1951]. Dorénavant, on identifiera la valeur $f_A(x_i)$ à la valeur u_i, et l'ensemble A à $\mathbf{u}$. La définition de Shannon [Shannon, 1948]

$$H_S(\mathbf{u}) = -\sum_{i=1}^{n} u_i \log_2 u_i \tag{2.1}$$

est réputée être la seule mesure sensible à l'incertitude dans le cadre de la théorie probabiliste. L'une des premières mesures entropiques pour des ensembles flous est proposée par De Luca et Termini [De Luca and Termini, 1972], étendant le principe de la mesure d'entropie de Shannon

$$H_{DT}(\mathbf{u}) = -K \sum_{i=1}^{n} u_i \log_2 u_i + (1-u_i)\log_2(1-u_i), \tag{2.2}$$

où K est constante de normalisation, et sera plus tard généralisée par Knopfmacher [Knopfmacher, 1975]

$$H(\mathbf{u}) = h\left(\sum_{i=1}^{n} g_i\left(u_i\right)\right) \tag{2.3}$$

où h est une fonction monotone croissante de $\mathbb{R}^+$ dans $\mathbb{R}^+$, et $g_i : [0,1] \to \mathbb{R}^+$ des fonctions associées à chaque u_i. Elles satisfont $g_i(0) = g_i(1) = 0$, $g_i(1/2)$ est un maximum unique, et g_i est monotone croissante sur $[0, 1/2]$, monotone décroissante sur $[1/2, 1]$. On notera que

1. En particulier lorsqu'il y a ambiguïté sur la décision à prendre

Dombi introduira une mesure fondée sur l'utilisation d'opérateurs conjonctifs et de négation [Dombi and Porkolab, 1991]

$$H_D(\mathbf{u}) = \frac{K}{n} \sum_{i=1}^{n} u_i \top \overline{u_i} \tag{2.4}$$

On pourra se référer à [Karmeshu, 2003; Klir, 2006] pour un panorama complet des mesures entropiques et des mesures d'incertitude s'appliquant à des probabilités, des possibilités, des ensembles flous voir même des capacités (ou mesures floues).

2.2.2 Mesures de spécificité

L'U-incertitude est une mesure de non-spécificité provenant d'une généralisation de la mesure d'information de Hartley [Hartley, 1928], dont dérivera par la suite la célèbre entropie de Shannon :

$$U(\mathbf{u}) = \sum_{i=2}^{n} \left(u_{(i)} - u_{(i+1)} \right) \log_2 i \tag{2.5}$$

où les $u_{(i)}$ sont triés de manière décroissante, et par convention $u_{(n+1)} = 0$. Cette mesure est l'unique mesure de non-spécificité d'une représentation possibiliste de l'information [Klir and Mariano, 1987]. En 1982, Yager introduit le concept de mesure de *spécificité* [Yager, 1982]. Cette mesure fournit la quantité d'information contenue dans un ensemble flou. En particulier, elle évalue à quel point un sous-ensemble flou possède une valeur proche de 1, et seulement une. On notera que cette notion est fortement liée à l'inverse de la cardinalité floue, ainsi qu'au concept de granularité [Zadeh, 1997]. Toutefois, on fera la distinction entre spécificité et incertitude, voir la discussion dans [Yager, 2008] à ce sujet, ainsi que pour un panorama des applications potentielles de cette mesure. On pourra également consulter [Dubois and Prade, 1986] dans le cadre de la théorie de l'évidence où cette fois la spécificité sera minimisée pour la sélection d'un ensemble parmi des éléments d'évidence.

Définition 2.1. Une mesure de spécificité $\mathcal{SP}$ est une fonction $\mathcal{SP} : [0,1]^c \to [0,1]$ telle que

(P1) $\mathcal{SP}(\mathbf{u}) = 1$ si et seulement si $\mathbf{u}$ est un singleton.

(P2) $\mathcal{SP}(\emptyset) = 0$.

(P3) Si $\mathbf{u}$ et $\mathbf{u}'$ sont des ensembles flous normaux[2], et $\mathbf{u} \subset \mathbf{u}'$, alors $\mathcal{SP}(\mathbf{u}) \geq \mathcal{SP}(\mathbf{u}')$.

La première proposition tient son origine au lien existant entre la spécificité et la cardinalité d'un ensemble flou, [Yager, 1982]

$$\mathcal{SP}(\mathbf{u}) = \int_0^{\alpha_{\max}} \frac{1}{|\mathbf{u}_\alpha|} \mathrm{d}\alpha \tag{2.6}$$

où $\alpha_{\max}$ est le plus haut degré d'appartenance de $\mathbf{u}$, et $\mathbf{u}_\alpha$ l'α-coupe de $\mathbf{u}$ (voir Annexe B). Plus tard, le même auteur introduit les mesures de spécificité linéaire [Yager, 1990]

$$\mathcal{SP}(\mathbf{u}) = u_{(1)} - \sum_{j=2}^{n} w_j u_{(j)} \tag{2.7}$$

où les $u_{(j)}$ sont triés de manière décroissante, et les poids w_j respectent les contraintes suivantes :

– $w_j \in [0,1]$,

2. $\mathbf{u}$ est dit normal si il existe un i tel que $u_i = 1$ (voir Annexe B)

– $\sum_{j=2}^{n} w_j = 1$,
– $w_j \geq w_i$ pour $j < i$.

Une combinaison de deux t-normes $\top$, $\top'$, une t-conorme $\perp$, et une négation est utilisée dans [Garmendia et al., 2003], et aboutit à la mesure de spécificité suivante

$$\mathcal{SP}(\mathbf{u}) = \top\left(u_{(1)}, \overline{\underset{j=2,n}{\perp}\left(\top'(u_{(j)}, w_j)\right)}\right) \tag{2.8}$$

où $\top$, $\top'$ et $\perp$ ne sont pas forcément liées. On verra cet opérateur comme une seule grande valeur et aucune autre. Enfin, plus récemment, une version continue se fondant sur des mesures floues est proposée [Yager, 2008]

$$\mathcal{SP}(\mathbf{u}) = \int_0^{\alpha_{\max}} F\left(\mu\left(\mathbf{u}_\alpha\right)\right) \mathrm{d}\alpha \tag{2.9}$$

et $F : [0,1] \to [0,1]$ satisfait $F(0) = 1$, $F(1) = 0$, et $0 \leq F(x) \leq F(y)$ pour $x > y$. De manière indépendante, une mesure de spécificité ayant le même but que (2.8) est donnée dans [Mascarilla and Frélicot, 2001; Frélicot and Le Capitaine, 2009]. Il s'agit d'une version floue de l'opérateur OU exclusif, le XOR, défini par

$$\underline{\perp}(\mathbf{u}) = \overset{1}{\perp}(\mathbf{u}) \top \overline{(\overset{2}{\perp}(\mathbf{u}) / \overset{1}{\perp}(\mathbf{u}))} \tag{2.10}$$

où $\overset{k}{\perp}$ est le OU flou d'ordre, cas particulier de l'opérateur présenté en section 1.5. Il doit également être compris dans le sens où l'on veut une seule valeur forte (la plus grande, $\overset{1}{\perp}$), et que toutes les autres soient faibles : le deuxième plus grande, $\overset{2}{\perp}$, doit être faible par rapport à la plus grande.

2.2.3 Un nouveau type de mesure : la similarité par blocs

Dans le chapitre précédent, nous avons présenté un opérateur évaluant la similarité d'ordre sur des valeurs appartenant à l'intervalle unité. Ici, nous adoptons une démarche différente, puisque nous voulons quantifier les similarités par blocs de valeurs, non nécessairement les plus grandes ou les plus petites valeurs comme c'était le cas pour les opérateurs $\overset{k}{\perp}\mathbf{u}$ et $\overset{k}{\top}\mathbf{u}$, voir Fig. 2.1. Cette approche permet d'évaluer un vecteur de valeurs, non plus

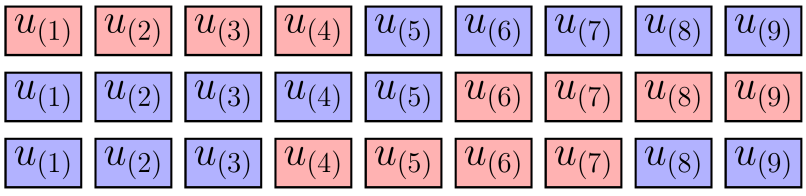

Fig. 2.1: Un vecteur de $c = 9$ valeurs triées. Première ligne : les blocs en rouge sont considérés par le OU flou d'ordre 4. Deuxième ligne : les blocs en rouge sont considérés par le ET flou d'ordre 4. Troisième ligne : les blocs en rouge sont considérés par un opérateur de similarité d'un bloc (parmi d'autres) de taille 4.

avec la restriction de prendre en compte les valeurs extrêmes, mais de spécifier la zone (ou bloc) d'intérêt. Ceci conduit à plusieurs propriétés que l'opérateur doit satisfaire :

1. La similarité des valeurs extrêmes 1 et 0 est nulle.

2. Si les degrés d'appartenance $\mathbf{u}$ aux objets j et k sont égaux, alors on demande d'avoir une similarité totale.
3. Il apparaît naturel que la similarité augmente lorsque $\mathbf{x}$ évolue de manière à ce que l'un des deux degrés d'appartenance à j ou k se rapproche de l'autre.
4. Inversement, la similarité doit diminuer lorsqu'ils s'éloignent.

Définition 2.2. Un opérateur de similarité par blocs est une famille de fonctions $\Phi_{j,k} : [0,1]^c \to [0,1]$, $(j,k) \in C \times C$, $j < k$, vérifiant les propriétés suivantes, sous la contrainte que le c-uplet $\mathbf{u}$ soumis à l'agrégation est ordonné de manière décroissante (et donc noté $\mathbf{u} = (u_{(1)}, \cdots, u_{(c)})$) :

$$(P1) \qquad \Phi_{j,k}(\mathbf{u}) = 0 \text{ si } u_{(j)} = 1 \text{ et } u_{(k)} = 0 \tag{2.11}$$

$$(P2) \qquad \Phi_{j,k}(\mathbf{u}) = 1 \Leftrightarrow u_{(j)} = u_{(k)} \tag{2.12}$$

$$(P3) \qquad \forall\, 0 \leq \varepsilon \leq u_{(j-1)} - u_{(j)},$$
$$\Phi_{j,k}(u_{(1)}, \cdots, u_{(j)} + \varepsilon, \cdots, u_{(c)}) \leq \Phi_{j,k}(u_{(1)}, \cdots, u_{(j)}, \cdots, u_{(c)}) \tag{2.13}$$

$$(P4) \qquad \forall\, 0 \leq \varepsilon \leq u_{(k-1)} - u_{(k)},$$
$$\Phi_{j,k}(u_{(1)}, \cdots, u_{(k)} + \varepsilon, \cdots, u_{(c)}) \geq \Phi_{j,k}(u_{(1)}, \cdots, u_{(k)}, \cdots, u_{(c)}) \tag{2.14}$$

Ainsi, les fonctions $\Phi_{j,k}$ ont pour but de mesurer la similarité entre le j-ème et le k-ème élément. Plus précisément, $\Phi_{j,k}$ mesure la similarité de tous les $u_{(i)}$ appartenant au bloc borné par les indices j et k. La valeur $\Phi_{1,c}$ reflète donc la similarité du plus grand bloc possible, c'est à dire la similarité globale de l'ensemble des degrés d'appartenance. Une première approche pour la définition d'une telle fonction serait le simple quotient $\overset{k}{\perp}/\overset{j}{\perp}$, mais celui-ci ne respecte pas les propriétés requises, par exemple *(P2)* de la Définition 2.2 : prenons par exemple $\mathbf{u} = (0.5, 0.2, 0.5)$ et $(j,k) = (1,2)$. Avec les normes (min, max), on a bien $\overset{2}{\perp}(\mathbf{u})/\overset{1}{\perp}(\mathbf{u}) = u_{(2)}/u_{(1)} = 1$. Pourtant, si on utilise le couple $(\top, \perp)_P$, on trouve $\overset{2}{\perp}(\mathbf{u})/\overset{1}{\perp}(\mathbf{u}) = 0.27/0.8 \neq 1$.

Nous avons présenté dans le chapitre précédent l'intégrale de Sugeno (section 1.4.3). Dans le cadre présent, $\mathcal{S}_\mu(\mathbf{u})$ est l'intégrale floue de $\mathbf{u}$ par rapport à une mesure floue μ, et nous l'utilisons pour mesurer la similarité associée à $\mathbf{u}$. Soit $A_i = \{j, u_{(j)} \geq u_{(i)}\}$, et la mesure floue μ_k définie par

$$\mu_k(A_i) = \begin{cases} 0 & \text{si } Card(A_i) < k \\ 1 & \text{sinon} \end{cases} \tag{2.15}$$

et nous fixons

$$\overset{k}{\mathcal{S}}_\mu(\mathbf{u}) = \underset{i=1,\cdots,c}{\perp} \Big(u_{(i)} \top \mu_k(A_i) \Big) \tag{2.16}$$

Proposition 2.1. *Si l'on se fixe une mesure floue respectant (2.15), alors* $\overset{k}{\mathcal{S}}_\mu$ *peut s'écrire*

$$\overset{k}{\mathcal{S}}_\mu(\mathbf{u}) = \begin{cases} \underset{i=k,\cdots,c}{\perp} u_{(i)} & \textit{si } u_{(k-1)} > u_{(k)} \\ \underset{i=j,\cdots,c}{\perp} u_{(i)} & \textit{où } j \textit{ est défini par } u_{(j-1)} > u_{(j)} = \cdots = u_{(k)} \end{cases} \tag{2.17}$$

Démonstration. On peut écrire

$$\overset{k}{\mathcal{S}}_\mu(\mathbf{u}) = \Big(\underset{i=1,\cdots,k-1}{\perp} u_{(i)} \top \mu_k(A_i)\Big) \perp \Big(\underset{i=k,\cdots,c}{\perp} u_{(i)} \top \mu_k(A_i)\Big)$$

D'après (2.15), on a $\mu_k(A_i) = 0$ pour tout $i \in \{1, \cdots, k-1\}$, et $\mu_k(A_i) = 1$ pour tout $i \in \{k, \cdots, c\}$. Les valeurs 0 et 1 étant respectivement les éléments absorbant et neutre de la norme triangulaire $\top$, on obtient facilement (2.17). □

Proposition 2.2. *L'ensemble des fonctions*

$$\Phi_{j,k}(\mathbf{u}) = \frac{\overset{k}{\mathcal{S}}_\mu(\mathbf{u})}{\overset{j}{\mathcal{S}}_\mu(\mathbf{u})} \tag{2.18}$$

pour $(j,k) \in C \times C, j < k$ fournit un opérateur qui respecte les propriétés (P1), (P2), (P3) et (P4) de la Définition *2.2, c'est à dire un opérateur de similarité par blocs, si l'on utilise les normes triangulaires (*min, max*) et $(\top, \perp)_P$.*

Démonstration.

(P1) Soit $\mathbf{u} = (1, \cdots, 1, u_{(j+1)}, \cdots, u_{(k-1)}, 0, \cdots, 0)$. Alors on a $\overset{k}{\mathcal{S}}_\mu(\mathbf{u}) = 0$ et $\overset{j}{\mathcal{S}}_\mu(\mathbf{u}) = 1$ puisque les valeurs 0 et 1 sont respectivement les éléments neutre et absorbant de la conorme triangulaire $\perp$. On a donc $\Phi_{j,k} = 0$.

(P2) Si $u_{(j)} = u_{(k)}$, alors $\overset{k}{\mathcal{S}}_\mu(\mathbf{u}) = \overset{j}{\mathcal{S}}_\mu(\mathbf{u})$ (voir (2.17)). Et on a donc $\Phi_{j,k} = 1$. Réciproquement, supposons que $\Phi_{j,k} = 1$. Alors il existe i dans $\{1, \cdots, k\}$ tel que

$$\overset{k}{\mathcal{S}}_\mu(\mathbf{u}) = u_{(i)} \perp \cdots \perp u_{(k)} \perp \cdots \perp u_{(c)}$$

et il existe un $l \leq i$ tel que

$$\begin{aligned} \overset{j}{\mathcal{S}}_\mu(\mathbf{u}) &= u_{(l)} \perp \cdots \perp u_{(i)} \perp \cdots \perp u_{(c)} \\ &= u_{(l)} \perp \cdots \perp u_{(i-1)} \perp \overset{k}{\mathcal{S}}_\mu(\mathbf{u}) \end{aligned}$$

Finalement, on a

$$\Phi_{j,k}(\mathbf{u}) = \frac{\overset{k}{\mathcal{S}}_\mu(\mathbf{u})}{u_{(l)} \perp \cdots \perp u_{(i-1)} \perp \overset{k}{\mathcal{S}}_\mu(\mathbf{u})} = 1$$

Si la t-conorme est strictement monotone, on a nécessairement $l = i$, et par conséquent $u_{(i)} = u_{(k)}$. Dans le cas contraire, on ne peut pas conclure. Parmi les t-conormes non strictement monotones, on pourra trouver par exemple $\perp_M$ et $\perp_L$. On peut prouver facilement que l'utilisation de $\perp_M$ vérifie *(P2)*, tandis que $\perp_L$ ne le permet pas.

(P3) Soit $u_{(j-1)} - u_{(j)} \geq \varepsilon > 0$. On a alors

$$\begin{aligned} \overset{k}{\mathcal{S}}_\mu(u_{(1)}, \cdots, u_{(j-1)}, u_{(j)} + \varepsilon, \cdots, u_{(k)}, \cdots u_{(c)}) &\leq \overset{k}{\mathcal{S}}_\mu(u_{(1)}, \cdots, u_{(j-1)}, u_{(j)}, \cdots, u_{(k)}, \cdots u_{(c)}) \\ &\textit{et} \\ \overset{j}{\mathcal{S}}_\mu(u_{(1)}, \cdots, u_{(j-1)}, u_{(j)} + \varepsilon, \cdots, u_{(k)}, \cdots u_{(c)}) &\geq \overset{j}{\mathcal{S}}_\mu(u_{(1)}, \cdots, u_{(j-1)}, u_{(j)}, \cdots, u_{(k)}, \cdots u_{(c)}) \end{aligned}$$

Ainsi,

$$\Phi_{j,k}(u_{(1)},\cdots,u_{(j-1)},u_{(j)}+\varepsilon,\cdots,u_{(k)},\cdots u_{(c)}) \leq \Phi_{j,k}(u_{(1)},\cdots,u_{(j-1)},u_{(j)},\cdots,u_{(k)},\cdots u_{(c)})$$

et ce quelles que soient les normes triangulaires utilisées.

(P4) Soit $u_{(k-1)} - u_{(k)} \geq \varepsilon > 0$. Il s'agit ici de comparer les quantités

$$\frac{u_{(l)} \perp \cdots \perp u_{(k)} + \varepsilon \perp \cdots \perp u_{(c)}}{u_{(i)} \perp \cdots \perp u_{(j)} \perp \cdots \perp u_{(l)} \perp \cdots \perp u_{(k)} + \varepsilon \perp \cdots \perp u_{(c)}}$$

et

$$\frac{u_{(k)} \perp \cdots \perp u_{(c)}}{u_{(i)} \perp \cdots \perp u_{(j)} \perp \cdots \perp u_{(l)} \perp \cdots \perp u_{(k)} \perp \cdots \perp u_{(c)}},$$

avec $l > j$. On suppose que $u_{(j)} > u_{(k)} + \varepsilon$, puisque si $u_{(j)} = u_{(k)} + \varepsilon$, alors

$$\Phi_{j,k}(u_{(1)},\cdots,u_{(k)}+\varepsilon,\cdots u_{(c)}) = 1,$$

et donc *(P4)* est vérifiée. Le résultat étant dépendant des normes triangulaires utilisées, nous ne présentons les résultats que pour les normes triangulaires (min, max). Pour la t-conorme max, on voit immédiatement que la relation est vérifiée, puisque $\Phi_{j,k}(u_{(1)},\cdots,u_{(k)}+\varepsilon,\cdots,u_{(c)}) = \Phi_{j,k}(u_{(1)},\cdots,u_{(k)},\cdots,u_{(c)})$.

□

Si $u_{(j)} = 0$, alors $\overset{j}{\mathcal{S}}_{\mu}(\mathbf{u}) = 0$, et l'on pose donc, pour éviter une division par zéro, $\overset{k}{\mathcal{S}}_{\mu}/\overset{j}{\mathcal{S}}_{\mu} = 1$. Cette convention est tout à fait logique, puisque si $u_{(j)} = 0$, alors $u_{(k)} = 0$, la similarité est donc totale.

Cet opérateur est donc une première proposition pour la mesure de similarité, mais il comporte plusieurs défauts. D'une part, il tient compte, lors du calcul, des valeurs $u_{(i)}$ appartenant au bloc $\{k+1,\cdots,c\}$ et ce même si $u_{(k)} > u_{(k+1)}$, alors qu'il ne donne aucun poids aux $u_{(i)}$ du bloc $\{1,\cdots,j-1\}$, même si $u_{(j-1)} > u_{(j)}$. Or il parait logique de souhaiter que seules les valeurs appartenant au bloc $\{j,\cdots,k\}$ interviennent. D'autre part, la même importance est donnée à chacune des valeurs par l'intermédiaire de la mesure floue cardinale quel que soit leur rang. Nous proposons donc de donner de moins en moins d'importance à la valeur $u_{(i)}$, $j \leq i \leq k$, au fur et à mesure que $u_{(i)}$ de $\mathbf{u}$ s'éloigne de $u_{(j)}$ et $u_{(k)}$. Comme cela ne peut se faire de manière arbitraire, nous proposons d'utiliser des fonctions noyaux permettant d'obtenir une certaine symétrie autour du degré considéré. Il existe évidemment de nombreuses fonctions noyaux, et nous détaillerons ce choix à la fin de cette section. Afin d'introduire les importances de manière progressive, on se place dans un premier temps au centre du bloc considéré, $(k+j)/2$ si $k-j$ est pair, et l'importance associée, c'est à dire la valeur du noyau, est minimum puisque nous sommes éloignés de l'indice considéré. La valeur du noyau augmente ensuite lorsque l'on se rapproche de j et de k au dénominateur

et au numérateur, respectivement. L'importance maximum est enfin atteinte aux indices j et k, puisque le noyau est normalisé : $\mathcal{K}(i,i) = 1$.

Proposition 2.3 ([Le Capitaine et al., 2007a])**.** *Soit une fonction noyau $\mathcal{K}_\lambda(i,l)$ centrée en l, l'opérateur défini par*

$$\Phi_{j,k}^{\mathcal{K}_\lambda}(\mathbf{u}) = \begin{cases} \dfrac{\perp_{i=\frac{k+j}{2}}^{k} u_{(i)} \top \mathcal{K}_\lambda(i,k)}{\perp_{i=\frac{k+j}{2}}^{j} u_{(i)} \top \mathcal{K}_\lambda(i,j)} & \text{si } k-j \text{ est pair} \\ \dfrac{\perp_{i=\frac{k+j+1}{2}}^{k} u_{(i)} \top \mathcal{K}_\lambda(i,k)}{\perp_{i=\frac{k+j-1}{2}}^{j} u_{(i)} \top \mathcal{K}_\lambda(i,j)} & \text{si } k-j \text{ est impair} \end{cases} \tag{2.19}$$

avec la convention $\Phi_{j,k}^{\mathcal{K}_\lambda}(\mathbf{u}) = 1$ si $u_{(j)} = 0$, est un opérateur de similarité par blocs si l'on utilise des normes triangulaires strictement monotones.

Démonstration.
On montre directement que *(P1)*, *(P2 ($\Leftarrow$))*, *(P3)* et *(P4)* de la Définition 2.2 sont vérifiées pour l'ensemble des couples de normes triangulaires $(\top, \perp)$ par propriété de celles-ci et la symétrie des noyaux. Nous montrons donc *(P2 ($\Rightarrow$))* pour les couples de t-normes strictement monotones.
Supposons dans un premier temps que $k - j$ est pair, ce qui ne change pas la nature du résultat. Posons

$$x = \Big(u_{([k+j]/2)} \top \mathcal{K}((k+j)/2,k)\Big) \perp \cdots \perp \Big(u_{(k-1)} \top \mathcal{K}((k-1,k)\Big)$$

et

$$y = \Big(u_{(j+1)} \top \mathcal{K}(j+1,j)\Big) \perp \cdots \perp \Big(u_{([k+j]/2)} \top \mathcal{K}((k+j)/2,j)\Big)$$

Par convention, on a $u_{(j)} > u_{(k)}$, on obtient alors

$$\frac{x \perp u_{(k)}}{u_{(j)} \perp y} = 1 \Leftrightarrow x \perp u_{(k)} = u_{(j)} \perp y$$

Dans ce cas, il faut nécessairement que $x = y = 1$, par stricte monotonie, et puisque 1 est élément absorbant de $\perp$. On aura donc $x = 1 \Leftrightarrow \mathcal{K}_\lambda = 1$ et $u_{([k+j]/2)} = \cdots = u_{(k-1)} = 1$ (toujours par stricte monotonie). De manière identique, $y = 1 \Leftrightarrow \mathcal{K}_\lambda = 1$ et $u_{(j+1)} = \cdots = u_{([k+j]/2)} = 1$.
En somme, si le couple est strictement monotone ou si $\mathcal{K}_\lambda \neq 1$, ou si $(u_{(j+1)}, \cdots, u_{(k-1)}) \neq (1, \cdots, 1)$, alors $\Phi_{j,k}^{\mathcal{K}_\lambda} \neq 1$. Ce qui nous amène à conclure que *(P2 ($\Rightarrow$))* est démontrée, par contraposée. □

Remarque 2.1. Pour les couples de t-normes non strictement monotones, par exemple $(\top,\perp)_M$ et $(\top,\perp)_L$, l'implication ($\Rightarrow$) de *(P2)* fait défaut. Pourtant, en pratique, la probabilité de considérer un vecteur comportant $(u_{(j+1)}, \cdots, u_{(k-1)}) = (1, \cdots, 1)$ est très faible, en particulier dans le domaine qui nous intéresse, la reconnaissance de formes.

Nom	Expression
Uniforme	$\mathcal{K}(y) = \frac{1}{2} 1_{(y \leq 1)}$
Gaussien	$\mathcal{K}(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y^2\right)$
Exponentiel	$\mathcal{K}(y) = \frac{1}{2} \exp(-y^2)$
Epanechnikov	$\mathcal{K}(y) = \frac{3}{4}(1-y^2) 1_{(y \leq 1)}$
Triangulaire	$\mathcal{K}(y) = (1-y) 1_{(y \leq 1)}$
Cauchy	$\mathcal{K}(y) = \frac{1}{\pi(1+y^2)}$

Tab. 2.1: Exemples de noyaux $\mathcal{K}(i,l)$ où y représente $|i-l|$. La notation $1_{(x)}$ est adoptée pour dénoter 1 lorsque la proposition x est vraie, 0 sinon.

On notera que la convention $\Phi_{j,k}^{\mathcal{K}_\lambda}(\mathbf{u}) = 1$ si $u_{(j)} = 0$ permet, comme dans la proposition précédente, d'éviter la division par zéro. Ici encore, $u_{(j)} = 0 \Rightarrow \forall i \geq j,\ u_{(i)} = 0$. Ce qui implique $u_{(k)} = 0$, la similarité totale étant alors naturelle dans ce cas.

Comme nous l'avons précisé plus haut, on peut choisir de nombreuses fonctions noyaux, voir Tab. 2.1 pour des exemples. Par mesure de simplicité, nous choisissons un noyau gaussien (normal) défini par

$$\mathcal{N}_\lambda(i,l) = \exp\left(\frac{(i-l)^2}{\lambda}\right) \tag{2.20}$$

où λ est un paramètre de résolution supérieur à 0. Lorsque λ tend vers 0, ce noyau devient un dirac δ_k centré en k, la convergence n'étant pas uniforme par continuité de $\mathcal{N}_\lambda$ et discontinuité de δ_k.

Proposition 2.4. *Si les normes triangulaires utilisées sont continues, alors :*

$$\lim_{\lambda \to 0} \Phi_{j,k}^{\mathcal{K}_\lambda}(\mathbf{u}) = \begin{cases} 1 & \text{si } u_j = 0 \\ \frac{u_{(k)}}{u_{(j)}} & \text{sinon} \end{cases} \tag{2.21}$$

Démonstration. Si une fonction f est continue en un point y, alors $\lim_{x \to y} f(x) = f(y)$. □

Lorsque λ tend vers l'infini, $\mathcal{N}_\lambda(i,l)$ tend vers 1, la convergence n'est pas uniforme. Par l'absurde, supposons $\forall \varepsilon > 0$, il existe Λ, $\forall \lambda \geq \Lambda$, $\forall i$, $|\exp\left(\frac{(i-l)^2}{\lambda}\right)| \leq \varepsilon$. Si $\lambda \geq \Lambda$, l'hypothèse n'est plus vérifiée dès que $|i-l| > \sqrt{-\lambda ln(1-\varepsilon)}$.

Proposition 2.5. *Si les normes triangulaires sont continues sur* $[0,1] \times 1$*, alors*

$$\lim_{\lambda \to +\infty} \Phi_{j,k}^{\mathcal{K}_\lambda}(\mathbf{u}) = \begin{cases} \dfrac{\perp_{i=\frac{k+j}{2}}^{k} u_{(i)}}{\perp_{i=\frac{k+j}{2}}^{j} u_{(i)}} & \text{si } k-j \text{ est pair} \\ \dfrac{\perp_{i=\frac{k+j+1}{2}}^{k} u_{(i)}}{\perp_{i=\frac{k+j-1}{2}}^{j} u_{(i)}} & \text{si } k-j \text{ est impair} \\ 1 & \text{si } u_{(j)} = 0 \end{cases} \tag{2.22}$$

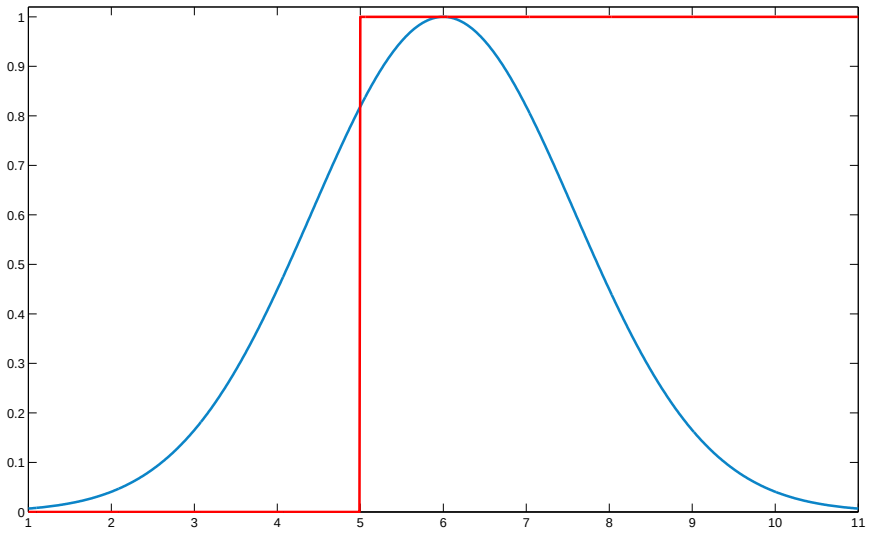

FIG. 2.2: Noyau $\mathcal{N}_5(i,l)$ et mesure cardinale avec $u_{(5)} > u_{(6)}$, pour $l = 6$.

Démonstration. On montre cette proposition on considérant les propriétés de la limite d'une fonction continue, de manière analogue à la proposition précédente. □

Finalement, le terme de lissage (ou variance) λ sert à régler le poids que l'on accorde aux différents $u_{(i)}$, pour $i \in \{j+1, \cdots, k-1\}$ lors de la mesure de similarité du bloc constitué par $(u_{(j)}, \cdots, u_{(k)})$. Une valeur nulle de ce paramètre impliquera un poids nul, et inversement, une grande valeur associera un poids important aux valeurs intermédiaires. Pour autant, on notera que la valeur $\Phi^{\mathcal{K}_\lambda}_{j,j+1} = u_{(j+1)}/u_{(j)}$ ne dépend pas de λ dans $\mathbb{R}^+$. Concrètement, augmenter sa valeur ne rendra pas deux $u_{(i)}$ consécutifs plus similaires, mais augmentera la similarité de blocs de taille supérieure. On notera également que si $\mathbf{u}$ est un vecteur constant sur le bloc considéré, alors $\Phi^{\mathcal{K}_\lambda}_{j,k}(\mathbf{u}) = 1$, quel que soit le couple (j,k). Comme attendu, l'introduction d'un noyau à la place d'une mesure cardinale permet de régler le poids des valeurs, et ainsi de ne plus être contraint à l'approche stricte qu'imposait la première proposition, voir FIG. 2.2.

Si l'on dispose d'un $\mathbf{u}$, le calcul de $\Phi^{\mathcal{K}_\lambda}_{j,k}$ pour tous les couples (j,k) dans $C \times C$ donne lieu à un tableau symétrique. Dans TAB. 2.2, nous donnons les valeurs (tronquées) de l'opérateur $\Phi^{\mathcal{N}_\lambda}_{j,k}(\mathbf{u})$ pour tous les couples (j,k), et pour des valeurs de $\lambda = 0.5$ et $\lambda = 2$, respectivement, utilisant le couple standard $(\perp,\top)_M$ pour $\mathbf{u} = (0.90, 0.80, 0.68, 0.51, 0.50, 0.48, 0.32, 0.1)$.. Nous donnons également, à titre de comparaison sur les valeurs de sortie, les valeurs de $\Phi^{\mathcal{N}_\lambda}_{j,k}$ pour des valeurs de λ identiques, mais en utilisant cette fois-ci le couple $(\perp,\top)_P$ en TAB. 2.3.

Évidemment, pour un j fixé, les valeurs de $\Phi^{\mathcal{N}_\lambda}_{j,k}$ diminuent lorsque k augmente, conformément aux propriétés attendues. La détection de bloc de similarité consiste donc à explorer la partie triangulaire supérieure du tableau, et comparer les valeurs à un seuil s que l'utilisateur aura défini. De manière algorithmique, ceci peut s'exprimer sous la forme de l'ALGORITHME 1. Afin de comprendre le comportement de l'opérateur en fonction de λ,

$\Phi_{j,k}$	u	0.90	0.80	0.68	0.51	0.50	0.48	0.32	0.10
u	$\downarrow j,k \rightarrow$	1	2	3	4	5	6	7	8
0.90	1	1.00	0.88	0.75	0.56	0.55	0.53	0.35	0.15
0.80	2	0.88	1.00	0.85	0.63	0.62	0.60	0.40	0.16
0.68	3	0.75	0.85	1.00	0.75	0.73	0.70	0.47	0.19
0.51	4	0.56	0.63	0.75	***1.00***	***0.98***	***0.94***	0.62	0.26
0.50	5	0.55	0.62	0.73	0.98	1.00	0.96	0.64	0.27
0.48	6	0.53	0.60	0.70	0.94	0.96	1.00	0.66	0.28
0.32	7	0.35	0.40	0.47	0.62	0.64	0.66	1.00	0.31
0.10	8	0.15	0.16	0.19	0.26	0.27	0.28	0.31	1.00

$\Phi_{j,k}$	u	0.90	0.80	0.68	0.51	0.50	0.48	0.32	0.10
u	$\downarrow j,k \rightarrow$	1	2	3	4	5	6	7	8
0.90	1	1.00	0.88	0.75	0.67	0.56	0.55	0.53	0.35
0.80	2	0.88	1.00	0.85	0.75	0.63	0.62	0.60	0.40
0.68	3	0.75	0.85	1.00	0.75	0.75	0.73	0.70	0.47
0.51	4	0.67	0.75	0.75	***1.00***	***0.98***	***0.98***	***0.94***	0.62
0.50	5	0.56	0.63	0.75	0.98	1.00	0.96	0.96	0.64
0.48	6	0.55	0.62	0.73	0.98	0.96	1.00	0.66	0.66
0.32	7	0.53	0.60	0.70	0.94	0.96	0.66	1.00	0.31
0.10	8	0.35	0.40	0.47	0.62	0.64	0.66	0.31	1.00

TAB. 2.2: *Haut* : Valeurs de $\Phi_{j,k}^{\mathcal{K}_\lambda}$ avec $(\bot,\top)_M$ et $\mathcal{K}_\lambda(i,l) = \mathcal{N}_{0.5}(i,l)$ - *Bas* : Valeurs de $\Phi_{j,k}^{\mathcal{K}_\lambda}$ avec $(\bot,\top)_M$ et $\mathcal{K}_\lambda(i,l) = \mathcal{N}_2(i,l)$.

Algorithme 1 : Algorithme de parcours des tableaux symétriques pour la détection de similarité par blocs.

Données : Un vecteur **u** de degrés d'appartenance trié de manière décroissante, un opérateur de similarité par blocs $\Phi_{j,k}^{\mathcal{K}_\lambda}$ et un seuil s de similarité.
Résultat : Blocs dont les valeur de degrés d'appartenance sont similaires.
pour *j de 1 à c (ligne)* **faire**
  pour *k de j+1 à c (colonne)* **faire**
    si $\Phi_{j,k}^{\mathcal{K}_\lambda}(\mathbf{u}) > s$ **alors**
      $(u_{(j)}, \ldots, u_{(k)})$ sont similaires.
    fin
  fin
fin

comparons la quatrième ligne de chaque table, en d'autres termes, $\Phi_{4,k}$ pour $k = 4, \cdots, 8$.

- lorsque $\lambda = 0.5$ (TAB. 2.2), les valeurs de $\Phi_{4,4}^{\mathcal{N}_\lambda}$, $\Phi_{4,5}^{\mathcal{N}_\lambda}$ et $\Phi_{4,6}^{\mathcal{N}_\lambda}$ sont plus grandes que $s = 0.90$, et les $u_{(i)}$ correspondant ($\{0.51,0.50,0.48\}$) sont considérés comme similaires pour ce seuil. Une faible valeur de $\Phi_{4,7}^{\mathcal{N}_\lambda}$ et $\Phi_{4,8}^{\mathcal{N}_\lambda}$ indique enfin que aussi bien $u_{(7)} = 0.32$ que $u_{(8)} = 0.10$ ne sont pas similaires aux autres.
- si maintenant nous observons les valeurs de (TAB. 2.2), où $\lambda = 2$, $\Phi_{4,7}^{\mathcal{N}_\lambda}$ devient supérieur au seuil s, et ainsi $u_{(7)} = 0.32$ est considéré comme similaire aux valeurs du bloc précédent, tandis que $u_{(8)}$ ne l'est toujours pas. Le bloc de valeurs similaires est donc dans ce cas $\{0.51,0.50,0.48,0.32\}$
- si l'on choisit une valeur $s = 0.70$, les blocs détectés sont également différents. Lorsque $\lambda = 0.5$, on distingue les blocs $(u_{(1)}, \cdots, u_{(3)})$ et $(u_{(3)}, \cdots, u_{(6)})$. L'augmentation du paramètre de résolution permet de mettre en lumière des blocs de plus grande taille : $(u_{(1)}, \cdots, u_{(3)})$ et $(u_{(3)}, \cdots, u_{(7)})$.

Très clairement, à partir des résultats de TAB. 2.3, nous voyons que si l'on utilise le couple $(\top, \bot)_P$, il faudra une valeur de paramètre λ du noyau supérieure à celle que l'on

$\Phi_{j,k}$	u	0.90	0.80	0.68	0.51	0.50	0.48	0.32	0.10
u	$\downarrow j,k \rightarrow$	1	2	3	4	5	6	7	8
0.90	1	1.00	0.88	0.78	0.60	0.58	0.56	0.39	0.15
0.80	2	0.88	1.00	0.85	0.67	0.65	0.62	0.44	0.17
0.68	3	0.78	0.85	1.00	0.75	0.76	0.73	0.51	0.19
0.51	4	0.60	0.67	0.75	1.00	0.98	0.94	0.67	0.25
0.50	5	0.58	0.65	0.76	0.98	1.00	0.96	0.68	0.26
0.48	6	0.56	0.62	0.73	0.94	0.96	1.00	0.66	0.27
0.32	7	0.39	0.44	0.51	0.67	0.68	0.66	1.00	0.31
0.10	8	0.15	0.17	0.19	0.25	0.26	0.27	0.31	1.00

$\Phi_{j,k}$	u	0.90	0.80	0.68	0.51	0.50	0.48	0.32	0.10
u	$\downarrow j,k \rightarrow$	1	2	3	4	5	6	7	8
0.90	1	1.00	0.88	0.88	0.75	0.72	0.69	0.58	0.34
0.80	2	0.88	1.00	0.85	0.80	0.74	0.74	0.61	0.36
0.68	3	0.88	0.85	1.00	0.75	0.84	0.81	0.69	0.40
0.51	4	0.75	0.80	0.75	1.00	0.98	0.96	0.78	0.47
0.50	5	0.72	0.74	0.84	0.98	1.00	0.96	0.80	0.42
0.48	6	0.69	0.74	0.81	0.96	0.96	1.00	0.66	0.47
0.32	7	0.58	0.61	0.69	0.78	0.80	0.66	1.00	0.31
0.10	8	0.34	0.36	0.40	0.47	0.42	0.47	0.31	1.00

Tab. 2.3: *Haut* : Valeurs de $\Phi_{j,k}^{\mathcal{K}_\lambda}$ avec $(\perp,\top)_P$ et $\mathcal{K}_\lambda(i,l) = \mathcal{N}_{0.5}(i,l)$ - *Bas* : Valeurs de $\Phi_{j,k}^{\mathcal{K}_\lambda}$ avec $(\perp,\top)_P$ et $\mathcal{K}_\lambda(i,l) = \mathcal{N}_2(i,l)$.

aurait fixé pour le couple $(\top,\perp)_M$ afin d'obtenir les même résultats. Ceci s'explique par le caractère strictement monotone du couple produit, que l'on retrouve pour de nombreux autres couples, de manière plus ou moins intense. Nous pouvons également formuler une autre remarque illustrant les propositions 2.4 et 2.5. Lorsque λ est faible, la différence de résultats entre les couples est faible, ceci s'explique par le résultat limite $\Phi_{j,k} = \frac{u_{(k)}}{u_{(j)}}$ de la proposition 2.4, quel que soit le couple $(\top,\perp)$. Inversement, les différences de valeurs entre le Tab. 2.2 et le Tab. 2.3 sont plus grandes, dans la mesure où le choix du couple revêt plus d'importance.

En Fig. 2.3, nous donnons un exemple de valeurs de sortie de l'opérateur, voir la légende pour la description. On voit aisément la différence induite par le changement d'indice. Alors qu'il suffit que deux valeurs soient similaires pour obtenir une similarité élevée par l'utilisation de $\Phi_{1,2}^{\mathcal{N}_\lambda}$, il faut nécessairement trois valeurs similaires pour obtenir une sortie identique dans le cas de $\Phi_{1,3}^{\mathcal{N}_\lambda}$. Comme nous l'avons précisé, il existe beaucoup de noyaux différents, permettant de choisir le poids de chacune des valeurs assez précisément. En Fig. 2.4, on peut observer le comportement et l'évolution des noyaux précédemment définis, pour un centre fixé en $l = 6$. On notera cependant que les noyaux utilisés dans $\Phi_{j,k}^{\mathcal{K}_\lambda}$ sont normalisés, de sorte que $\mathcal{K}(l,l) = 1$. Reprenons le $\mathbf{u}$ précédent, le Tab. 2.4 présente les valeurs de l'opérateurs $\Phi_{j,k}^{\mathcal{K}_\lambda}$ pour j fixé à 1, et k variant de 2 à 8, et $(\top,\perp)_M$. Évidemment, nous constatons que $\Phi_{1,2} = 0.88$, quel que soit le noyau utilisé, ceci étant dû à la propriété limite de l'opérateur. Nous observons également que les principales différences entre noyaux se situent au niveau des blocs les plus petits (hormis $\Phi_{1,2}^{\mathcal{K}_\lambda}$). Ceci est facilement explicable d'une part en considérant la Fig. 2.4 : lorsque l'on s'éloigne du centre, les valeurs de noyaux tendent à se confondre.

Dans toutes les applications où une mesure de similarité entre valeurs est nécessaire, on peut utiliser cet opérateur. Parmi celles-ci, citons les quelques-unes qui nous intéressent, présentées brièvement pour le moment, mais développées au Chapitre 4 et Chapitre 5.

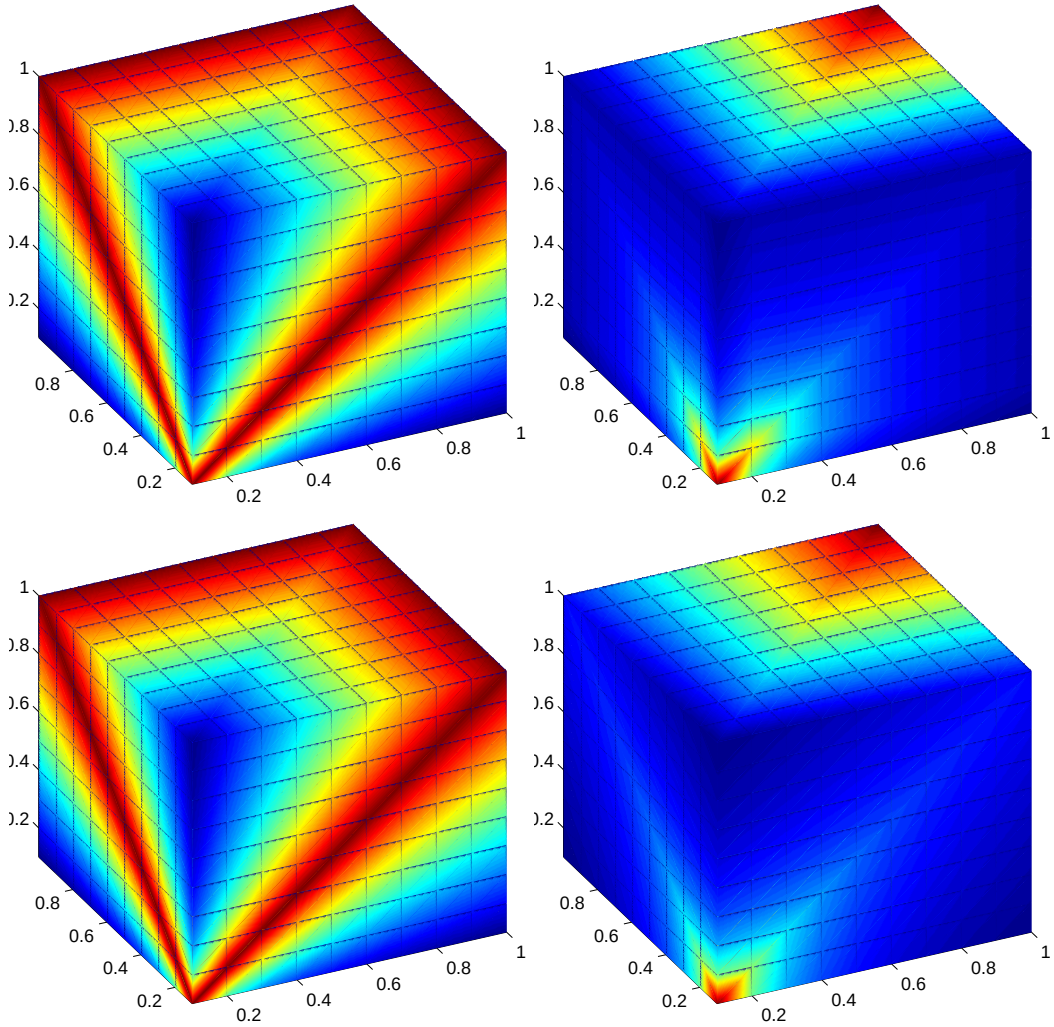

FIG. 2.3: *Haut* : Valeurs de sortie des opérateurs $\Phi_{1,2}^{\mathcal{N}_{0.5}}$ (*gauche*) et $\Phi_{1,3}^{\mathcal{N}_{0.5}}$ (*droite*) avec le couple $(\top, \perp)_M$ pour l'ensemble des vecteurs $\mathbf{u} \in [0,1]^3$. La couleur rouge dénote la valeur la plus haute, 1, et la couleur bleue la plus faible, 0. *Bas* : $\Phi_{1,2}^{\mathcal{N}_{0.5}}$ (*gauche*) et $\Phi_{1,3}^{\mathcal{N}_{0.5}}$ (*droite*) avec le couple $(\top, \perp)_P$

Dans les problèmes de validation de partition, on estime la similarité successive de blocs de taille de plus en plus grande selon le nombre de clusters de la partition, [Le Capitaine et al., 2007a]. En traitement d'image, pour un point (pixel) et un voisinage donné, la similarité des valeurs de ce voisinage est estimée. Si celle-ci est grande, on peut considérer que les pixels du voisinage appartiennent à une région homogène, et inversement qu'il y a présence d'un contour [Le Capitaine et al., 2007b]. On notera que l'on pourra identifier cette approche à une approche de type filtrage robuste, dans la mesure où on peut ne considérer que les valeurs intermédiaires du bloc de pixels (et ainsi éliminer les pixels de bruit). Enfin, nous avons utilisé cet opérateur en sélection de classes multiples en classification supervisée. La similarité (ou degré d'appartenance) d'un point à l'ensemble des classes est évaluée, et permet ainsi d'associer ce point à une ou plusieurs classes du problème, [Le Capitaine and Frélicot, 2008b].

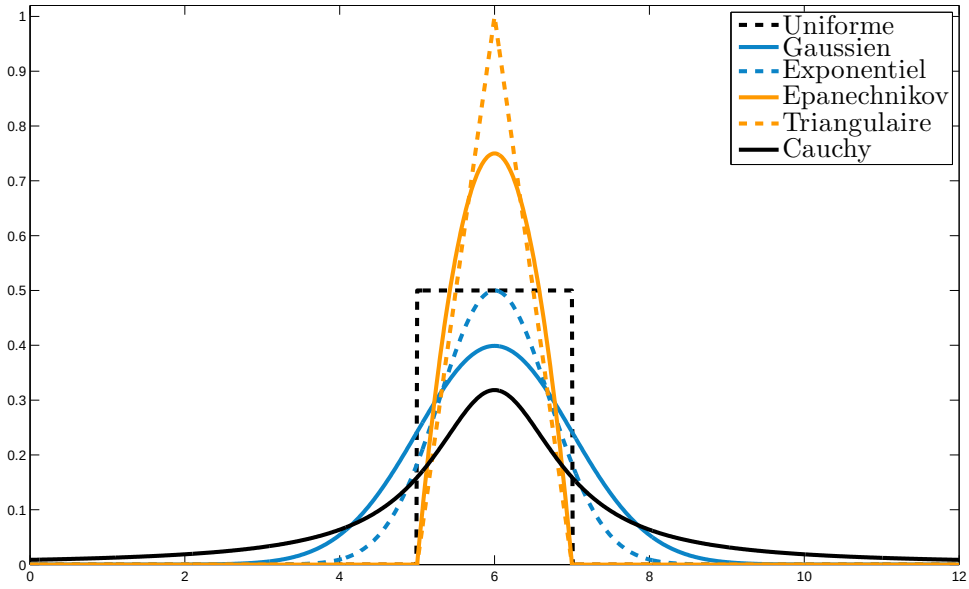

FIG. 2.4: Noyaux centrés en $l = 6$.

$\Phi_{j,k}$	u	0.80	0.68	0.51	0.50	0.48	0.32	0.10
$\mathcal{K}_\lambda$	$k \rightarrow$	2	3	4	5	6	7	8
Uniforme		0.88	0.88	0.75	0.56	0.55	0.53	0.35
Gaussien ($\lambda = 2$)		0.88	0.75	0.67	0.56	0.55	0.53	0.35
Exponentiel ($\lambda = 1$)		0.88	0.75	0.56	0.55	0.53	0.40	0.35
Epanechnikov		0.88	0.83	0.75	0.56	0.55	0.53	0.35
Triangulaire		0.88	0.75	0.56	0.55	0.55	0.53	0.35
Cauchy ($\lambda = 2$)		0.88	0.75	0.74	0.56	0.55	0.53	0.37

TAB. 2.4: Valeurs de $\Phi_{1,k}^{\mathcal{K}_\lambda}$ avec $(\bot,\top)_M$ pour $k = 2, \cdots, c$, et différents noyaux $\mathcal{K}_\lambda(i,l)$.

2.2.4 Vers une approche logique

Un problème général en logique floue est de traiter des propositions conditionnelles de type *si* x, *alors* y, où x et y sont des prédicats flous. Une méthode largement utilisée consiste à les gérer par l'introduction de fonctions $I : [0,1] \times [0,1] \rightarrow [0,1]$ de telle sorte que la valeur de I dépend des propositions initiales x et y.

Définition 2.3. On parle de fonctions d'implications floues (voir [Mas et al., 2007]) si I

(P1) est non croissante avec la première variable,

(P2) est non décroissante avec la seconde variable,

(P3) si $I(0,0) = I(1,1) = 1$,

(P4) et $I(1,0) = 0$.

A partir de ces propriétés, découlent $I(0,x) = 1$ et $I(x,1) = 1$, ce qui fait coïncider I avec l'implication stricte sur $\{0,1\}^2$. On distingue cinq types d'implications :

1. *S-implications*, définies par :

$$I_\bot(x,y) = \bot(\overline{x},y) \tag{2.23}$$

Cette implication est en fait l'immédiate généralisation de l'implication booléenne $x \rightarrow y \equiv \overline{x} \vee y$.

2. *R-(pour résiduelle) implications*, définies par :

$$I_{\top}(x,y) = \sup_t\{t \in [0,1] | \top(x,t) \leq y\}. \tag{2.24}$$

On notera que si $\top$ est continue à gauche, l'opération de *supremum* peut être transformée par le maximum.

3. *QL-(pour Quantum mechanic Logic) implications*, définies par :

$$I_{QL}(x,y) = \perp(\overline{x},\top(x,y)). \tag{2.25}$$

4. *D-implications*, définies par :

$$I_D(x,y) = \perp(\top(\overline{x},\overline{y}),y). \tag{2.26}$$

qui sont la contraposition directe des QL-implications par rapport au complément.

5. *A*-implications [Turksen et al., 1998], définies par :

$$I_A(x,y) = y^x. \tag{2.27}$$

De nombreuses propriétés additionnelles peuvent être requises [Mas et al., 2007], parmi elles citons

(P5) Principe de confinement : $x \leq y$ si et seulement si $I(x,y) = 1$,

(P6) Principe de bord : $I(1,x) = x, \forall x \in [0,1]$,

(P7) Principe d'échange : $I(x,I(y,z)) = I(y,I(x,z))$,

(P8) Principe de contraposition : $I(x,y) = I(\overline{y},\overline{x})$.

Parmi les implications définies à partir de t-normes, sont bien connues : l'implication de Gödel, donnée par

$$I(x,y) = \begin{cases} 1 & \text{si } y \geq x \\ y & \text{si } y < x \end{cases} \tag{2.28}$$

l'implication de Goguen, définie par

$$I(x,y) = \begin{cases} 1 & \text{si } y \geq x \\ \frac{y}{x} & \text{si } y \leq x \end{cases} \tag{2.29}$$

et enfin l'implication de Lukasiewicz, donnée par

$$I(x,y) = \min(1, 1 - x + y) \tag{2.30}$$

obtenues respectivement avec les t-normes minimum, produit et Lukasiewicz. Notons que l'on peut également obtenir des implications paramétriques via des t-normes paramétriques, voir nos propositions à la section 2.2.4.1. On pourra se référer à [Whalen, 2003] pour une description de l'usage d'implications paramétriques dans le cadre de systèmes d'inférence floue. Selon [Hirota and Pedrycz, 1991], pour des éléments x et y de X, le degré d'égalité entre x et y peut être défini à partir d'implications de la manière suivante :

$$(x \equiv y) = \frac{1}{2}((x \to y) \wedge (y \to x) + (\overline{x} \to \overline{y}) \wedge (\overline{y} \to \overline{x})) \tag{2.31}$$

où $\wedge$ est l'opérateur minimum, $\to$ est une implication résiduelle. On peut étendre cette définition de la manière suivante

$$(x \equiv y) = \frac{1}{2}((x \to y)\top(y \to x) + (\overline{x} \to \overline{y})\top(\overline{y} \to \overline{x})) \tag{2.32}$$

Puis, en utilisant l'équation (2.24) avec $x \geq y$ et le principe de confinement *(P5)*, donne

$$(x \equiv y) = \frac{1}{2}((x \to y) + (\overline{y} \to \overline{x})) \tag{2.33}$$

puisque 1 est l'élément neutre des t-normes.
De plus, par la loi de contraposition $I(x,y) = I(\overline{y},\overline{x})$, on obtient finalement

$$(x \equiv y) = (x \to y) \tag{2.34}$$

Une manière convenable de définir une mesure de similarité est de quantifier à quel point deux degrés d'appartenance sont similaires, ce qui relie très clairement ce problème au problème d'appariement de quantités floues, ou similarité d'ensembles flous. Cette similarité à partir d'implications peut être interprétée de la manière suivante :

(x est similaire à y) veut dire que (x implique y *et* y implique x).

Nous proposons ainsi d'utiliser les fonctions d'implications floues comme mesures de similarité. Comme nous supposerons par la suite que $x \geq y$, cette relation se réduira à l'équation (2.34).

2.2.4.1 Mesures paramétriques

Dans cette section, nous proposons des implications résiduelles paramétriques, puisque à notre connaissance, très peu d'entre elles ont été introduites en dépit du grand nombre de normes triangulaires paramétriques.

Proposition 2.6 ([Le Capitaine and Frélicot, 2009a]). *Soit* $(\top_{H_\gamma})$, $\gamma \in [0,+\infty[$, *la famille des t-normes de Hamacher. L'implication résiduelle de Hamacher est alors donnée par*

$$I_{H_\gamma}(x,y) = \begin{cases} 1 & \text{si } y \geq x \\ \dfrac{y\,(\gamma + x - \gamma x)}{y\,(\gamma + x - \gamma x) + x - y} & \text{si } y \leq x \end{cases} \tag{2.35}$$

Démonstration.
Par définition des implications résiduelles (2.24), on peut écrire $I_{H_\gamma}(x,y) = \sup_t\{t \in [0,1] : \top_{H_\gamma}(x,t) \leq y\}$. On a aussi $I_{H_\gamma}(x,y) = \max_t\{t \in [0,1] : \top_{H_\gamma}(x,t) \leq y\}$ puisque $\top_{H_\gamma}$ est une t-norme continue à gauche. Alors, résoudre

$$\frac{x\,t}{\gamma + (1-\gamma)\,(x + t - x\,t)} \leq y \tag{2.36}$$

donne

$$t \leq \frac{y\,(\gamma + x - \gamma x)}{y\,(\gamma + x - \gamma x) + x - y}. \tag{2.37}$$

Puisque $x \geq y$, il est facile de montrer que la partie droite de l'inéquation (2.37) appartient à [0,1], pour enfin nous donner l'équation (2.35). □

Proposition 2.7 ([Le Capitaine and Frélicot, 2009a]). *Soit* $(\top_{D_\gamma})$, $\gamma \in]0,+\infty[$, *la famille des t-normes de Dombi. L'implication résiduelle de Dombi est donnée par*

$$I_{D_\gamma}(x,y) = \begin{cases} 1 & \text{si } y \geq x \\ \left(1 + \left(\left(\dfrac{1-y}{y}\right)^\gamma - \left(\dfrac{1-x}{x}\right)^\gamma\right)^{1/\gamma}\right)^{-1} & \text{si } y < x \end{cases} \tag{2.38}$$

Démonstration.
$I_{D_\gamma}(x,y) = \sup_t\{t \in [0,1] : \top_{H_\gamma}(x,t) \leq y\}$ par (2.24). Comme $\top_{D_\gamma}$ est une t-norme continue à gauche, alors on peut écrire $I_{D_\gamma}(x,y) = \max_t\{t \in [0,1] : \top_{D_\gamma}(x,t) \leq y\}$. Puis, la résolution de

$$\left(1 + \left(\left(\frac{1-x}{x}\right)^\gamma + \left(\frac{1-t}{t}\right)^\gamma\right)^{1/\gamma}\right)^{-1} \leq y \tag{2.39}$$

donne

$$t \leq \left(1 + \left(\left(\frac{1-y}{y}\right)^\gamma - \left(\frac{1-x}{x}\right)^\gamma\right)^{1/\gamma}\right)^{-1}. \tag{2.40}$$

Comme, $x \geq y$, il est facile de montrer que
$\left(\left(\frac{1-y}{y}\right)^\gamma - \left(\frac{1-x}{x}\right)^\gamma\right)^{1/\gamma} \geq 0$, et donc la partie droite de l'inéquation (2.40) est dans [0,1] et (2.38) est obtenue. □

Proposition 2.8. *Soit* $(\top_{Y_\gamma})$, $\gamma \in [0,+\infty[$, *la famille des t-normes de Yager. L'implication résiduelle de Yager est donnée par*

$$I_{Y_\gamma}(x,y) = \begin{cases} 1 & si\ y \geq x \\ 1 - \left((1-y)^\gamma - (1-x)^\gamma\right)^{1/\gamma} & si\ y \leq x \end{cases} \tag{2.41}$$

Démonstration.
$I_{Y_\gamma}(x,y) = \sup_t\{t \in [0,1] : \top_{Y_\gamma}(x,t) \leq y\}$ par (2.24). Comme $\top_{Y_\gamma}$ est une t-norme continue à gauche, alors on peut écrire $I_{Y_\gamma}(x,y) = \max_t\{t \in [0,1] : \top_{Y_\gamma}(x,t) \leq y\}$. Puis, la résolution de

$$1 - \left((1-x)^\gamma + (1-t)^\gamma\right)^{1/\gamma} \leq y \tag{2.42}$$

donne

$$t \leq 1 - \left((1-y)^\gamma - (1-x)^\gamma\right)^{1/\gamma}. \tag{2.43}$$

Comme, $x \geq y$, il est facile de montrer que
$1 - \left((1-y)^\gamma - (1-x)^\gamma\right)^{1/\gamma} \geq 0$, et donc la partie droite de l'inéquation (2.43) est dans [0,1] et (2.41) est obtenue. □

Proposition 2.9 ([Fono et al., 2007])**.** *Soit* $(\top_{F_\gamma})$, $\gamma \in [0,+\infty[$, *la famille des t-normes de Frank. L'implication résiduelle de Frank est donnée par*

$$I_{F_\gamma}(x,y) = \begin{cases} 1 & si\ x \geq y \\ \log_\gamma\left(1 + \dfrac{(\gamma^y - 1)(\gamma - 1)}{\gamma^x - 1}\right) & si\ y \leq x \end{cases} \tag{2.44}$$

Proposition 2.10. *Soit* $(\top_{DP_\gamma})$, $\gamma \in [0,+\infty[$, *la famille des t-normes de Dubois-Prade. L'implication résiduelle de Dubois-Prade est donnée par*

$$I_{DP_\gamma}(x,y) = \begin{cases} 1 & si\ y \geq x \\ \max(\gamma,x)\dfrac{y}{x} & si\ y \leq x \end{cases} \tag{2.45}$$

Démonstration.
On résout

$$\frac{xt}{\max(x,t,\gamma)} \leq y, \tag{2.46}$$

ce qui donne

$$t \leq \max(x,\gamma)\frac{y}{x} \tag{2.47}$$

Comme $x \geq y$ et $\gamma \in [0,1]$, on montre facilement que $\max(x,\gamma)\frac{y}{x}$ appartient à $[0,1]$, et on obtient donc (2.45). □

2.2.4.2 Exemples numériques

Une des difficultés rencontrées lors de l'utilisation de t-norme est le choix de la famille, et éventuellement du paramètre associé. Ce choix peut se révéler d'autant plus épineux que les résultats obtenus peuvent être radicalement différents. Pour cette raison, nous proposons ici une étude sur l'influence du double choix famille-paramètre, permettant ainsi, nous l'espérons, à l'utilisateur de sélectionner son opérateur en connaissance de cause.

Nous reprenons pour cette étude l'un des vecteurs que nous avons utilisé auparavant, $\mathbf{u} = (0.20, 0.10, 0.85, 0.80)$, car il présente plusieurs caractéristiques intéressantes : deux hautes valeurs similaires, et deux faibles valeurs similaires. Afin de faciliter l'écriture, nous noterons $a = 0.85$, $b = 0.80$, $c = 0.20$ et enfin $d = 0.10$.

- *Hamacher*:

Augmenter la valeur du paramètre γ rend deux valeurs floues plus similaires, car $I_{H_\gamma}(x,y) = xy/(x-y+xy)$ si $\gamma = 0$, tandis que $\lim_{\gamma \to +\infty} I_{H_\gamma}(x,y) = 1$ quels que soient $(x,y) \in [0,1]^2$. On peut en fait remarquer que I_{H_γ} est non décroissante avec γ, puisque

$$\frac{\partial I_{H_\gamma}}{\partial \gamma} = \frac{(y-xy)(x-y)}{(y(\gamma+x-\gamma x)+x-y)^2} \geq 0.$$

L'influence de γ est bien plus importante pour de faibles valeurs de x et y que pour de fortes valeurs. En effet, $y(\gamma+x-\gamma x)$ apparaît être de l'ordre de $y\,x$ (respectivement $y\,(\gamma+x)$ pour de fortes valeurs (respectivement faibles) de x et y. Ainsi, si $\gamma_1 >> \gamma_2$, on a $\frac{\gamma_1}{\gamma_1+\varepsilon} >> \frac{\gamma_2}{\gamma_2+\varepsilon}$. De fortes valeurs de γ associées à de faibles valeurs de (x,y) résultera une grande valeur de I_{H_γ}, voir Tab. 2.5 pour des exemples.

γ	0.5	2	5
$I_{H_\gamma}(a,b)$	0.93	0.94	0.96
$I_{H_\gamma}(a,c)$	0.20	0.26	0.33
$I_{H_\gamma}(a,d)$	0.10	0.13	0.17
$I_{H_\gamma}(b,c)$	0.21	0.28	0.37
$I_{H_\gamma}(b,d)$	0.10	0.14	0.20
$I_{H_\gamma}(c,d)$	0.16	0.64	0.80

Tab. 2.5: Exemples d'implications de Hamacher, où $a = 0.85$, $b = 0.80$, $c = 0.20$ et $d = 0.10$.

- *Dombi*:

Cette fois-ci, diminuer la valeur de γ rend deux valeurs floues plus similaires, puisque

$\lim_{\gamma\to+\infty} I_{D_\gamma}(x,y) = y$, tandis que $\lim_{\gamma\to 0} I_{D_\gamma}(x,y) = 1$ quels que soient $(x,y) \in [0,1]^2$. Contrairement à la famille de Hamacher, I_{D_γ} est non croissante avec γ

$$\frac{\partial I_{D_\gamma}}{\partial \gamma} \leq 0.$$

En somme, diminuer la valeur de γ de la famille de Dombi a le même impact que l'augmenter pour la famille de Hamacher, donnant une tendance opposée pour les valeurs du TAB. 2.6. Bien que les deux familles présentent un comportement opposé, ici encore, l'influence de γ sur la valeur d'implication de faibles valeurs est plus importante que pour de fortes valeurs.

γ	0.25	0.75	2
$I_{D_\gamma}(a,b)$	1.00	0.96	0.85
$I_{D_\gamma}(a,c)$	0.74	0.22	0.20
$I_{D_\gamma}(a,d)$	0.42	0.10	0.10
$I_{D_\gamma}(b,c)$	0.80	0.23	0.20
$I_{D_\gamma}(b,d)$	0.47	0.10	0.10
$I_{D_\gamma}(c,d)$	0.98	0.24	0.11

TAB. 2.6: Exemples d'implications de Dombi, où $a = 0.85$, $b = 0.80$, $c = 0.20$ et $d = 0.10$.

- *Yager*:

Ici, à l'instar de la famille de Dombi, diminuer la valeur de γ rend deux valeurs floues plus similaires, puisque $\lim_{\gamma\to+\infty} I_{Y_\gamma}(x,y) = x$, tandis que $\lim_{\gamma\to 0} I_{Y_\gamma}(x,y) = 1$ quels que soient $(x,y) \in [0,1]^2$. On trouve facilement que

$$\frac{\partial I_{Y_\gamma}}{\partial \gamma} \leq 0.$$

impliquant que I_{Y_γ} est non croissante avec γ. Ces résultats sont confirmés par le TAB. 2.7. On retrouve une nouvelle fois que la valeur de l'implication de faibles valeurs est fortement dépendante de γ, ce qui n'est pas le cas pour de fortes valeurs.

γ	0.5	1	2
$I_{Y_\gamma}(a,b)$	0.99	0.95	0.86
$I_{Y_\gamma}(a,c)$	0.74	0.35	0.21
$I_{Y_\gamma}(a,d)$	0.68	0.25	0.11
$I_{Y_\gamma}(b,c)$	0.80	0.40	0.22
$I_{Y_\gamma}(b,d)$	0.74	0.30	0.12
$I_{Y_\gamma}(c,d)$	0.99	0.90	0.58

TAB. 2.7: Exemples d'implications de Yager, où $a = 0.85$, $b = 0.80$, $c = 0.20$ et $d = 0.10$.

- *Frank*:

Cette fois-ci, diminuer la valeur de γ rend deux valeurs floues moins similaires, dans la mesure où $\lim_{\gamma\to+\infty} I_{F_\gamma}(x,y) = 1 - x + y$, tandis que $\lim_{\gamma\to 0} F_{Y_\gamma}(x,y) = x$ quels que soient $(x,y) \in [0,1]^2$. Après quelques calculs, on trouve que

$$\frac{\partial I_{F_\gamma}}{\partial \gamma} \geq 0.$$

La famille d'implications I_{F_γ} est donc non décroissante avec γ. La valeur de l'implication de fortes valeurs est peu modifiée par γ, alors que la valeur de l'implication de faibles valeurs varie plus en fonction de γ, voir TAB. 2.8.

γ	0.1	5	10
$I_{F_\gamma}(a,b)$	0.93	0.94	0.95
$I_{F_\gamma}(a,c)$	0.21	0.26	0.27
$I_{F_\gamma}(a,d)$	0.10	0.13	0.14
$I_{F_\gamma}(b,c)$	0.21	0.28	0.30
$I_{F_\gamma}(b,d)$	0.10	0.14	0.16
$I_{F_\gamma}(c,d)$	0.30	0.64	0.69

TAB. 2.8: Exemples d'implications de Frank, où $a = 0.85$, $b = 0.80$, $c = 0.20$ et $d = 0.10$.

- *Dubois-Prade*:

Ici, diminuer la valeur de γ rend deux valeurs floues moins similaires, puisque $\lim_{\gamma\to 0} I_{DP_\gamma}(x,y) = y$, tandis que $\lim_{\gamma\to 1} F_{Y_\gamma}(x,y) = y/x$ quels que soient $(x,y) \in [0,1]^2$. Le calcul de la dérivée apporte deux résultats intéressants. Si $\gamma \leq x$, alors

$$\frac{\partial I_{DP_\gamma}}{\partial \gamma} = 0.$$

l'implication I_{DP_γ} est donc constante quel que soit γ, résultat que l'on retrouve dans TAB. 2.9 pour les deux premières colonnes, et hormis la dernière ligne, où $\gamma \geq a$. Inversement, si $\gamma \leq x$,

$$\frac{\partial I_{DP_\gamma}}{\partial \gamma} = \frac{y}{x} \geq 0.$$

c'est le cas de la dernière ligne et de la dernière colonne du TAB. 2.9.

γ	0.3	0.5	0.9
$I_{DP_\gamma}(a,b)$	0.80	0.80	0.84
$I_{DP_\gamma}(a,c)$	0.20	0.20	0.21
$I_{DP_\gamma}(a,d)$	0.10	0.10	0.11
$I_{DP_\gamma}(b,c)$	0.20	0.20	0.22
$I_{DP_\gamma}(b,d)$	0.10	0.10	0.11
$I_{DP_\gamma}(c,d)$	0.15	0.25	0.45

TAB. 2.9: Exemples d'implications de Dubois-Prade, où $a = 0.85$, $b = 0.80$, $c = 0.20$ et $d = 0.10$.

En conclusion, deux grandes tendances se dégagent des implications paramétriques en fonction de l'évolution du paramètre γ. On constate que les implications de Hamacher, Frank et Dubois-Prade rendent deux valeurs plus similaires lorsque γ augmente, alors que les implications de Dombi et Yager présentent le comportement inverse : il faut diminuer γ pour rendre deux valeur plus similaires.

2.3 Mesures de compatibilité

Lorsque l'on compare des ensembles flous, on distingue trois grandes familles de mesures de compatibilité, selon la nature de leurs construction :

1. métrique (fondées sur des distances),
2. ensembliste,
3. logique.

On citera également une quatrième approche, moins courante, se fondant sur une approche morphologique [Bloch, 1996]. Par la suite, nous proposons une description de ces trois grandes approches. Ainsi, nous verrons qu'il existe des liens assez étroits entre les inclusions

$\mathcal{I}$, les distances $\mathcal{D}$ et les similarités $\mathcal{S}$, et nous présenterons donc indifféremment des mesures respectant un modèle de construction (métrique, ensembliste, logique) mais n'évaluant pas les mêmes caractéristiques. Nous présentons donc ces liens avant de détailler les différentes approches.
Par la suite, nous noterons

- $X = \{x_1, \cdots, x_n\}$ le (supposé fini) univers de discours.
- $\mathcal{C}(X)$ et $\mathcal{F}(X)$ les ensembles de tous les ensembles stricts et flous, respectivement.
- $f_A(x)$, $\forall x \in X$, la fonction d'appartenance de l'ensemble flou A sur X, qui pourra aussi être notée $\mathbf{u} = (u_1, \cdots, u_c)$, auquel cas $c = n$.
- $[\frac{1}{2}]$ l'ensemble flou constant défini par $[\frac{1}{2}](x) = \frac{1}{2}$ pour tout $x \in X$.

2.3.0.3 Des liens entre $\mathcal{I}$, $\mathcal{S}$ et $\mathcal{D}$ à partir de relations ensemblistes

Dans [Bouchon-Meunier et al., 1996], les auteurs introduisent une approche intéressante permettant de lier, ou connecter de nombreuses mesures de compatibilité (appelée comparaison dans l'article), en se basant sur les travaux de Tversky [Tversky, 1977]. En effet, une mesure de compatibilité est une fonction F de $a = f(A \cap B)$, $b = f(A - B)$ et $c = f(B - A)$. Selon les propriétés de F, on aura

- une mesure de *satisfiabilité*[3] si $F(0,b,c) = 0$, $F(a,0,c) = 1$, et F est croissante avec a, décroissante avec b, et indépendante de c.
- une mesure de *ressemblance* si $F(a,0,0) = 1$, $F(a,b,c) = F(a,c,b)$, et F est croissante avec a, décroissante avec b et c.
- une mesure de *similitude* si F est non décroissante avec a, et non croissante avec b et c.
- une mesure d'inclusion si $F(0,b,c) = 0$, et F est indépendante de c.
- une mesure de dissimilarité si $F = 0$ lorsque $A = B$, et que F est indépendante de a, et croissante avec b et c.

Une distance peut être obtenue grâce à une mesure de dissimilarité, ou par $1 - F$ si F est une mesure de similitude. On notera que les auteurs proposeront plus tard un moyen de sélection de mesures à partir de leurs pouvoir discriminant [Rifqi et al., 2000].

2.3.0.4 Des liens entre $\mathcal{I}$, $\mathcal{S}$ et $\mathcal{D}$ à partir de similarités

Le lien entre mesures de dissimilarité et mesure de similarités a déjà été évoqué en introduction de ce chapitre. En effet, si l'on possède une mesure de similarité à valeurs dans $[0, 1]$, on peut alors obtenir une distance en complémentant cette mesure, et vice versa. Certains auteurs avancent que dissimilarité et non similarité ne sont pas la même chose, et que pour obtenir une dissimilarité à partir d'une similarité, on prend $\mathcal{D}(A, B) = \mathcal{S}(\overline{A}, \overline{B})$. Nous discuterons de l'incidence de cette différence de point de vue en section 2.3.4. A partir d'une mesure de similarité, il est également possible d'obtenir une mesure d'inclusion définie par $\mathcal{I}(A, B) = \mathcal{S}(A, A \cap B)$, ou $\mathcal{I}(A, B) = \mathcal{S}(B, A \cup B)$, voir [Zeng and Li, 2006].

3. satisfiabilité d'une description de référence A par rapport à une nouvelle description B.

2.3.0.5 Des liens entre $\mathcal{I}$, $\mathcal{S}$ et $\mathcal{D}$ à partir d'inclusions

À partir de la définition de l'inclusion d'un ensemble dans un autre, la similarité de deux ensembles A et B peut être évaluée par A *est inclus dans* B et B *est inclut dans* A, comme proposé dans [Bandler and Kohout, 1980; Wang et al., 1995] :

$$\mathcal{S}(A,B) = \min\left(\mathcal{I}(A,B), \mathcal{I}(B,A)\right) \tag{2.48}$$

Plus récemment, [Zeng and Li, 2006] définissent une mesure de similarité par

$$\mathcal{S}(A,B) = \mathcal{I}(A,B)\,\mathcal{I}(B,A) \tag{2.49}$$

Plus généralement,

$$\mathcal{S}(A,B) = \mathcal{I}(A,B) \top \mathcal{I}(B,A) \tag{2.50}$$

est une mesure de similarité.

2.3.1 Mesures fondées sur des distances

Les mesures fondées sur des distances, ou proximité, sont des généralisations de modèles géométrique de distance. Le principe général est fondé sur la définition d'une distance floue entre ensembles flous, on se place ainsi dans un espace pseudo-métrique. De manière générale, une distance (ou métrique) d[4] est définie comme

Définition 2.4. Une métrique sur un ensemble X est une fonction $d : X \times X \to \mathbb{R}$ qui satisfait les conditions suivantes pour trois objets a, b et c :

(P1)	$d(a,b) = d(b,a)$,	symétrie,
(P2)	$d(a,b) = 0$ ssi $a = b$	identité des indiscernables,
(P3)	$d(a,b) \geq 0$,	non-négativité,
(P4)	$d(a,c) \leq d(a,b) + d(b,c)$,	inégalité triangulaire.

Selon les propriétés vérifiées, on pourra obtenir des pseudo-métriques, des quasi-métriques, ou des ultra-métriques, voir [Schroeder, 2006]. Dans le cas de la comparaison d'ensembles flous, des fonctions de distance peuvent être utilisées sur des points ou intervalles. Elles sont fondées en particulier sur le modèle générique des distances de r-Minkowski défini par

$$d_r(A,B) = \left(\int_{-\infty}^{+\infty} |f_A(x) - f_B(x)|^r \mathrm{d}x\right)^{1/r} \tag{2.51}$$

ou dans sa version discrète

$$d_r(A,B) = \left(\sum_{x \in X} |f_A(x) - f_B(x)|^r\right)^{1/r} \tag{2.52}$$

où selon la valeur de r, on retrouve des familles de distances bien connues :

- si $r = 1$, distance de Manhattan (ou Hamming),
- si $r = 2$, distance euclidienne,
- si $r = \infty$, distance de Tchebychev.

4. La notation d, et non $\mathcal{D}$, est ici volontairement adoptée, puisque c'est une métrique au sens général, nous réserverons la notation $\mathcal{D}$ pour exprimer une distance entre ensembles.

Ces distances ayant valeur dans $\mathbb{R}^+$, il convient de les normaliser dans $[0,1]$ par exemple en divisant par $|X|$, et, afin de construire une mesure de compatibilité, de prendre leur complément à 1. Les auteurs de [De Luca and Termini, 1972] définissent une distance entre a et b définie par $d(a,b) = |H_{DT}(a) - H_{DT}(b)|$, où H est définie par (2.2). De manière plus générale, l'objectif est de lier la notion d'entropie individuelle à la distance de deux entités. Toujours avec une approche métrique de la compatibilité, on trouvera les mesures utilisant la différence symétrique $\triangle$ (voir DÉFINITION 2.5), qui définissent une métrique sous certaines conditions [Alsina and Trillas, 2005]. Le lecteur pourra se référer à [Bloch, 1999] pour une revue des distances floues.

2.3.2 Mesures ensemblistes

La seconde approche consiste à considérer les opérations ensemblistes d'union, intersection et négation afin de comparer deux ensembles. L'une des premières propositions vient de Jaccard [Jaccard, 1908], qui utilise les attributs communs pour construire

$$\mathcal{S}(A,B) = \frac{|A \cap B|}{|A \cap B| + |A \cap \overline{B}| + |\overline{A} \cap B|} = \frac{|A \cap B|}{|A \cup B|} \tag{2.53}$$

où A et B sont deux ensembles. Cette définition est ensuite étendue par Tversky dans ses travaux basés sur une approche psychologique [Tversky, 1977]

$$\mathcal{S}(A,B) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)} \tag{2.54}$$

où α et β sont des paramètres permettant d'obtenir différents types de mesure. C'est une version paramétrique de (2.53) où $A - B = A \cap \overline{B}$ et $B - A = \overline{A} \cap B$. Généralement, la fonction f est cardinale, mais elle peut être différente, du moment que $f(X \cup Y) = f(X) + f(Y)$ pour X et Y disjoints. Une nouvelle généralisation est proposée dans [Bouchon-Meunier et al., 1996], où une mesure de comparaison est donnée par

$$F\left(f(A \cap B), f(B - A), f(A - B)\right) \tag{2.55}$$

Plus tard, Santini et Jain [Santini and Jain, 1999] proposeront une version floue [5] du modèle à contraste de Tversky (dénominateur de l'équation (2.54)) pour lequel l'intersection et l'union sont modélisées par les opérations min et max, respectivement. D'autre part, ils proposeront pour f une cardinalité floue usuelle, puis l'intégrale de Choquet afin de prendre en compte les dépendances entre attributs. Enfin, Tolias et al. proposeront d'utiliser la t-norme produit ($\top_P$) pour modéliser l'intersection dans ce cadre, et introduiront ainsi l'indice de Tversky généralisé, [Tolias et al., 2001].

La différence symétrique a souvent été employée pour définir la similarité entre ensembles strictes.

Définition 2.5. La différence symétrique $A \bigtriangleup B$ peut être écrite de deux manières $A \bigtriangleup^- B = (A \cap \overline{B}) \cup (B \cap \overline{A})$ et $A \bigtriangleup^+ B = (A \cup B) \cap \overline{(A \cap B)}$. L'opérateur binaire $\triangle$ sur les degrés d'appartenance de deux éléments respecte les propriétés suivantes

(P1) $\triangle(x,0) = \triangle(0,x) = x$,

(P2) $\triangle(x,x) = 0$,

5. fondée sur l'utilisation de prédicats flous

(P3) $\triangle(x,1) = \triangle(1,x) = \overline{x}$

A partir d'une différence symétrique, des conditions doivent être satisfaites pour définir une mesure de similarité [Dubois and Prade, 1982a],

(P1) $\mathcal{S}(A,B) = 1$ si et seulement si $A \bigtriangleup B = \emptyset$,

(P2) si A et B ont des supports disjoints, alors $\mathcal{S}(A,B) = 0$,

(P3) $\mathcal{S}(A,B) = \mathcal{S}(B,A)$,

(P4) $\mathcal{S}(A,B)$ dépend de $f(\overline{A \bigtriangleup B})$, ou d'une fonction symétrique de $f(A \cup \overline{B})$ et $f(\overline{A} \cup B)$.

Par exemple, on pourra définir la mesure de similarité suivante

$$\mathcal{S}(A,B) = \frac{f(\overline{A \bigtriangleup B} - f(\overline{A \cup B})}{1 - f(\overline{A \cup B})} \tag{2.56}$$

On peut trouver des approches alternatives, en particulier dans [De Baets et al., 2002; Bosteels and Kerre, 2007]

2.3.3 Mesures logiques

L'approche logique est fondée sur le degré d'implication des éléments de A sur les éléments de B, et inversement [Bandler and Kohout, 1980]. Pour cela, un opérateur d'implication $\rightarrow$ est défini, voir [Mas et al., 2007]. Il peut s'agir d'implications résiduelles, de S-implications, ou de QL et D-implications, voir section 2.2.4 pour les définitions. De manière générale, le degré d'égalité (ou d'appariement) [Hirota and Pedrycz, 1991] entre $f_A(x)$ et $f_B(x)$ est défini comme

$$\begin{gathered}(f_A(x) \equiv f_B(x)) \\ = \\ \frac{(f_A(x) \rightarrow f_B(x)) \wedge (f_B(x) \rightarrow f_A(x)) + (\overline{f_A(x)} \rightarrow \overline{f_B(x)}) \wedge (\overline{f_B(x)} \rightarrow \overline{f_A(x)})}{2}\end{gathered} \tag{2.57}$$

Si $\rightarrow$ satisfait le principe de contraposition ($x \rightarrow y = \overline{y} \rightarrow \overline{x}$, voir section 2.2.4), alors (2.57) peut s'écrire

$$(f_A(x) \equiv f_B(x)) = (f_A(x) \rightarrow f_B(x)) \wedge (f_B(x) \rightarrow f_A(x)) \tag{2.58}$$

Ces relations fournissent un degré d'égalité, ou de similarité, entre degrés d'appartenance. Une mesure de similarité entre A et B peut alors être construite [Le Capitaine and Frélicot, 2009b]

$$\mathcal{S}(A,B) = \mathcal{A}_{x \in X}\left(f_A(x) \equiv f_B(x)\right) \tag{2.59}$$

De nombreuses solutions ont été proposées dans la littérature, une étude plus détaillée sur l'utilisation d'implications floues pour les mesures de comparaison est donc donnée en section 2.3.4, en parallèle avec notre propre proposition.

2.3.4 De nouvelles mesures de comparaison d'ensembles flous

Alors que l'on distingue trois grandes approches pour la génération de mesures de comparaison (voir section 2.3), nous nous focaliserons sur la troisième, l'approche logique. Avant toute chose, nous aurons besoin par la suite du théorème suivant.

Théorème 2.1. *Si $\mathcal{A}$ est un opérateur d'agrégation strictement monotone (au sens de (1.5)), et qu'il présente un comportement de compromis (1.8), alors*

$$\mathcal{A}(u_1,\cdots,u_c)=1 \Leftrightarrow u_1=\cdots=u_c=1 \tag{2.60}$$

et

$$\mathcal{A}(u_1,\cdots,u_c)=0 \Leftrightarrow u_1=\cdots=u_c=0 \tag{2.61}$$

Démonstration.

(2.60) ($\Leftarrow$) évident avec l'équation (1.3).
($\Rightarrow$) par contraposition, nous montrons qu'il existe i tel que $u_i<1$ implique $\mathcal{A}(u_1,\cdots,u_c)<1$: si $u_i<1$ pour un i, alors, puisque $u_j\leq 1$ pour tout j, par stricte monotonie de $\mathcal{A}$, on a $\mathcal{A}(u_1,\cdots,u_c)<\mathcal{A}(1,\cdots,1)=1$, ce qui termine la contraposition.

(2.61) ($\Leftarrow$) évident avec l'équation (1.2).
($\Rightarrow$) par contraposition, nous montrons qu'il existe i tel que $u_i>0$ implique $\mathcal{A}(u_1,\cdots,u_c)>0$: si $u_i>0$ pour un i, alors, puisque $u_j\geq 0$ pour tout j, par stricte monotonie de $\mathcal{A}$, on a $\mathcal{A}(u_1,\cdots,u_c)>\mathcal{A}(0,\cdots,0)=0$, ce qui termine la contraposition..

□

Exemple 2.1. La famille des opérateurs de type moyenne sont des opérateurs d'agrégation strictement monotones présentant un comportement de compromis.

Une mesure d'inclusion est une relation entre deux ensembles flous indiquant dans quelle mesure un ensemble flou est inclus dans un autre. Dans sa définition originelle, $A\subseteq B$ si et seulement si $f_A(x)\leq f_B(x)$, pour tout x dans X, ce qui est une conclusion stricte. Bandler et Kohout ont élargi ce point de vue en donnant un degré d'inclusion [Bandler and Kohout, 1980], plus cohérent avec l'esprit de la théorie de la logique floue. Les mesures d'inclusion sont généralement définies par un ensemble d'axiomes [Young, 1996], et par l'utilisation d'opérateurs d'implications floues[6] [Bandler and Kohout, 1980; Zhang and Zhang, 2009]. Dans ce travail, nous avons choisi les axiomes proposés par Young, bien que d'autres aient été par ailleurs proposés, voir [De Baets et al., 2002].

Définition 2.6 ([Young, 1996])**.** Une fonction $\mathcal{I}:\mathcal{F}(X)\times\mathcal{F}(X)\to[0,1]$ est appelée une mesure d'inclusion si elle satisfait

(P1) $\mathcal{I}(A,B)=1$ ssi $A\subseteq B,\quad \forall A,B\in\mathcal{F}(X)$,

(P2) si $[\frac{1}{2}]\subseteq A$, alors $\mathcal{I}(A,A^c)=0$ ssi $A=X$,

(P3) $\forall A,B,C\in\mathcal{F}(X)$, si $A\subseteq B\subseteq C$, alors $\mathcal{I}(C,A)\leq\mathcal{I}(B,A)$, et si $A\subseteq B$, $\mathcal{I}(C,A)\leq\mathcal{I}(C,B)$.

La propriété *(P1)* est naturelle et découle de la proposition originelle introduite par Zadeh. La propriété *(P2)* est sujette à plus de controverses : elle est par exemple remplacée par $\mathcal{I}(A,B)=0$ si $A\neq 0$ et $A\cap B=0$ dans [Fan et al., 1999]. Le fait que $[\frac{1}{2}]\subseteq A$ assure que $A^c\subset A$. Si $A\neq X$, alors A sera, en partie, inclut dans A^c : $A\cap A^c$ est non nulle, il faut

6. si, évidemment, on se restreint à l'approche logique.

donc $A = X$ pour que l'inclusion soit nulle. Notons que plusieurs auteurs ont également remplacé *(P2)* par : $\mathcal{I}(X,\emptyset) = 0$. Enfin, la propriété *(P3)* est quant à elle assez naturelle. Si une mesure $\mathcal{I}$ satisfait seulement les deux dernières propriétés, elle est alors appelée une mesure d'inclusion faible.

Théorème 2.2 ([Le Capitaine and Frélicot, 2009b]). *Soit $I_\top$ une fonction d'implication résiduelle. Soient X l'univers du discours, et des ensembles $A,B \in \mathcal{F}(X)$, posons*

$$\mathcal{I}(A,B) = \underset{i=1}{\overset{n}{\mathcal{A}}} I_\top\left(f_A(x_i),f_B(x_i)\right) \tag{2.62}$$

pour tout x_i dans X, où $\mathcal{A}$ est un opérateur d'agrégation conjonctif ou un opérateur d'agrégation strictement monotone de compromis satisfaisant (1.2) à (1.4). Alors $\mathcal{I}$ est une mesure d'inclusion.

Démonstration.

(P1) ($\Leftarrow$) si $A \subseteq B$, alors $f_A(x_i) \leq f_B(x_i)$ pour tout $x_i \in X$. Puisque si $x \leq y$ alors $I_\top(x,y) = 1$, on obtient $I_\top(f_A(x_i),f_B(x_i)) = 1$ pour tout $x_i \in X$, ce qui donne $\mathcal{I}(A,B) = 1$ par conditions aux bornes de $\mathcal{A}$.
($\Rightarrow$) Deux cas sont à considérer :
- si $\mathcal{A}$ est conjonctif, alors on a trivialement
$f_A(x_i) \leq f_B(x_i)$ pour tout $x_i \in X$, puisque la valeur minimum de l'ensemble des implications sur X vaut 1.
- si $\mathcal{A}$ est un opérateur d'agrégation strictement monotone et de compromis, alors en utilisant le Théorème 2.1, on a
$I_\top\left(f_A(x_i),f_B(x_i)\right) = 1$ pour tout $x_i \in X$, ce qui donne $A \subseteq B$ par le principe de confinement $x \leq y$ *ssi* $I(x,y) = 1$ (voir *(P5)* en p. 49).

(P2) ($\Leftarrow$) si $A = X$, alors $I_\top(f_X(x_i),\emptyset(x_i)) = 0$ pour tout $x_i \in X$. Par conditions aux bornes de $\mathcal{A}$, on obtient $\mathcal{I}(A,A^c) = 0$.
($\Rightarrow$) si $\mathcal{I}(A,A^c) = 0$, alors $I_\top(f_A(x_i),1-f_A(x_i)) = 0$ pour tout $x_i \in X$ par le Théorème 2.1. Supposons que $A \neq X$, alors il existe x_i tel que
$\frac{1}{2} \leq f_A(x_i) < 1$, i.e. $0 < 1 - f_A(x_i) \leq \frac{1}{2}$.
Par non croissance avec la première variable de $I_\top$, on a $I_\top(f_A(x_i),1 - f_A(x_i)) \geq I_\top(1,1 - f_A(x_i)) = 1 - f_A(x_i) \neq 0$, puisque $I_\top$ satisfait le principe de bord : $I(1,x) = x, \forall x \in [0,1]$ (voir *(P6)* en p. 49). C'est en contradiction avec $I_\top(f_A(x_i),1-f_A(x_i)) = 0$, et donc $A = X$.

(P3) si $A \subseteq B \subseteq C$, $f_A(x_i) \leq f_B(x_i) \leq f_C(x_i)$, pour tout $x_i \in X$. Par non croissance avec la première variable et la non décroissance avec la seconde variable, nous obtenons que $I_\top(f_C(x_i),f_A(x_i) \leq I_\top(f_B(x_i),f_A(x_i))$ et $I_\top(f_C(x_i),f_A(x_i)) \leq I_\top(f_C(x_i),f_B(x_i))$ pour tout x_i. Par monotonie de $\mathcal{A}$, on a $\mathcal{I}(C,A) \leq \mathcal{I}(B,A)$ et $\mathcal{I}(C,A) \leq \mathcal{I}(C,B)$, ce qui conclut la preuve.

□

Une condition nécessaire pour définir une mesure d'inclusion forte $\mathcal{I}$ est que l'implication I satisfasse le principe de confinement et le principe de bord. Il est facile de montrer que les quatre implications usuelles $I_\top$, $I_\bot$, I_{QL} et I_D satisfont le principe de bord, mais

seule l'implication résiduelle $I_\top$ satisfait le principe de confinement. Ainsi, les S, QL et $D-$implications définissent des inclusions faibles, tandis que les R-implications donnent lieu à des mesures d'inclusion fortes pourvu que $\mathcal{A}$ soit conjonctif ou strictement monotone et de compensation.

Comme nous l'avons déjà observé, les notions d'inclusion et de similarité sont voisines. Les mesures de similarité sont également définies par un ensemble d'axiomes, nous prenons ici les axiomes proposés par [Liu, 1992].

Définition 2.7 ([Liu, 1992])**.** Une fonction $\mathcal{S} : \mathcal{F}(X) \times \mathcal{F}(X) \to [0,1]$ est appelée une mesure de similarité si elle satisfait

(P1) $\mathcal{S}(A,B) = \mathcal{S}(B,A), \quad \forall A,B \in \mathcal{F}(X),$

(P2) $\mathcal{S}(A,A) = 1, \forall A \in \mathcal{F}(X),$

(P3) $\mathcal{S}(D,D^c) = 0, \quad \forall D \in \mathcal{C}(X),$

(P4) $\forall A,B,C \in \mathcal{F}(X),$ si $A \subseteq B \subseteq C$, alors $\mathcal{S}(A,C) \leq \mathcal{S}(A,B) \wedge \mathcal{S}(B,C)$
ou, de manière équivalente,
$\forall A,B,C,D \in \mathcal{F}(X)$, si $A \subseteq B \subseteq C \subseteq D$, alors $\mathcal{S}(A,D) \leq \mathcal{S}(B,C)$

Soulignons que la propriété *(P3)* est proche de la propriété *(P2)* modifiée de l'inclusion : $\mathcal{I}(X,\emptyset) = 0$.
Nous proposons dans le théorème suivant une mesure de similarité générique construit à partir d'implications résiduelles.

Théorème 2.3 ([Le Capitaine and Frélicot, 2009b])**.** *Soit $I_\top$ un opérateur d'implication résiduelle. Soient $A,B \in \mathcal{F}(X)$, posons*

$$\mathcal{S}(A,B) = \underset{i=1}{\overset{n}{\mathcal{A}}} I_\top \left(f_{(1)}(x_i), f_{(2)}(x_i) \right) \tag{2.63}$$

pour tout x_i dans X, où $f_{(.)}$ est une permutation de f_A et f_B telle que $f_{(1)}(x_i) = (f_A \cup f_B)(x_i)$, $f_{(2)}(x_i) = (f_A \cap f_B)(x_i)$, et $\mathcal{A}$ un opérateur d'agrégation satisfaisant les équations (1.2-1.3) et (1.4). Alors $\mathcal{S}$ est une mesure de similarité.

Démonstration.

(P1) on a $I_\top(x,x) = 1$, pour tout $x \in [0,1]$. Par conditions aux bornes de $\mathcal{A}$, voir équations (1.2-1.3), on obtient $\mathcal{S}(A,A) = 1$.

(P2) par commutativité des opérations d'union et d'intersection d'ensembles flous, on a

$$\begin{aligned} \mathcal{S}(A,B) &= \underset{i=1}{\overset{n}{\mathcal{A}}} I_\top \left((f_A \cup f_B)(x_i), (f_A \cap f_B)(x_i) \right) \\ &= \underset{i=1}{\overset{n}{\mathcal{A}}} I_\top \left((f_B \cup f_A)(x_i), (f_B \cap f_A)(x_i) \right) \\ &= \mathcal{S}(B,A) \end{aligned}$$

(P3) par définition, $I(1,0) = 0$. Par conditions aux bornes de $\mathcal{A}$, voir équation (1.2-1.3), on obtient $\mathcal{S}(D,D^c) = 0$.

(P4) puisque $A \subseteq B \subseteq C \subseteq D$, on a pour tout $x_i \in X$

$$f_D(x_i) \geq f_C(x_i) \tag{2.64}$$
$$f_B(x_i) \geq f_A(x_i) \tag{2.65}$$

Par non croissance avec la première variable et non décroissance avec la seconde de $I_\top$, on obtient pour tout $x_i \in X$

$$I_\top(f_D(x_i),f_A(x_i)) \leq I_\top(f_C(x_i),f_A(x_i)) \quad \text{par l'équation (2.64)}$$

$$I_\top(f_C(x_i),f_A(x_i)) \leq I_\top(f_C(x_i),f_B(x_i)) \quad \text{par l'équation (2.65)}$$

Enfin, la monotonie de $\mathcal{A}$, (1.4), assure que $\mathcal{S}(A,D) \leq \mathcal{S}(B,C)$, ce qui termine la preuve.

□

À la différence des mesures d'inclusion, aucune restriction n'est imposée sur l'opérateur d'agrégation $\mathcal{A}$. Il peut donc être choisi librement, pourvu que les conditions classiques aux bornes et la monotonie soient respectées. En revanche, le principe de confinement est nécessaire pour obtenir la propriété *(P1)* des mesures de similarité.

Comme $\mathcal{S}(A,B)$ est réflexive et symétrique, c'est à dire $\mathcal{S}(A,A) = 1$ et $\mathcal{S}(A,B) = \mathcal{S}(B,A)$ satisfait pour $A,B \in \mathcal{F}(X)$, alors $\mathcal{S}$ est une relation de proximité sur $\mathcal{F}(X)$.

Une distance entre deux ensembles flous étant une mesure évaluant leurs différences, on voit facilement que les concepts de similarité et de distance sont duals.

Définition 2.8 ([Liu, 1992])**.** Une fonction $\mathcal{D} : \mathcal{F}(X) \times \mathcal{F}(X) \to [0,1]$ est appelée une mesure de distance si $\mathcal{D}$ vérifie les propriétés suivantes

(P1) $\mathcal{D}(A,B) = \mathcal{D}(B,A)\,, \quad \forall A,B \in \mathcal{F}(X),$
(P2) $\mathcal{D}(A,A) = 0\,, \forall A \in \mathcal{F}(X),$
(P3) $\mathcal{D}(D,D^c) = 1\,, \quad \forall D \in \mathcal{C}(X),$
(P4) $\forall A,B,C \in \mathcal{F}(X)$, si $A \subseteq B \subseteq C$,
alors $\mathcal{D}(A,B) \leq \mathcal{D}(A,C)$ et $\mathcal{D}(B,C) \leq \mathcal{D}(A,C)$

Nous proposons donc, en se fondant sur les mesures de similarité précédemment introduites, le théorème suivant.

Théorème 2.4 ([Le Capitaine and Frélicot, 2009b])**.** *Soient $A,B \in \mathcal{F}(X)$, et $\mathcal{S}(A,B)$ une mesure de similarité définie par l'équation (2.63), alors la fonction $\mathcal{D} : \mathcal{F}(X)\times\mathcal{F}(X) \to [0,1]$ définie par $\mathcal{D}(A,B) = \overline{\mathcal{S}(A,B)}$ est une mesure de distance entre A et B.*

Démonstration.
La preuve est immédiate, puisque pour tout mesure de similarité $\mathcal{S}(A,B)$ satisfaisant les quatre propriétés de la Définition 2.7, alors $\overline{\mathcal{S}(A,B)}$ satisfait les quatre propriétés de la Définition 2.8. C'est évident pour les trois premières, et nous donnons la preuve pour *(P4)*.

Soient $A \subseteq B \subseteq C$, alors,

$$\mathcal{S}(A,B) \geq \mathcal{S}(A,C) \text{ et } \mathcal{S}(B,C) \geq \mathcal{S}(A,C),$$

d'où

$$\mathcal{D}(A,B) = \overline{\mathcal{S}(A,B)} \leq \overline{\mathcal{S}(A,C)} = \mathcal{D}(A,C), \text{ et}$$
$$\mathcal{D}(B,C) = \overline{\mathcal{S}(B,C)} \leq \overline{\mathcal{S}(A,C)} = \mathcal{D}(A,C).$$

□

La proposition $\mathcal{D}(A,B) = \mathcal{S}(\overline{A},\overline{B})$ ne peut être prouvée à cause de *(P2)* de la DÉFINITION 2.8, puisque $\mathcal{D}(A,A) = \mathcal{S}(\overline{A},\overline{A}) = 1$ quel que soit $\overline{A}$, ce qui est une contradiction.

Nous donnons des exemples de mesures d'inclusion (TAB. 2.10) et de mesures de similarité (TAB. 2.11) de la littérature, ainsi que les nouvelles mesures paramétriques de Hamacher, Dombi, Yager et Frank. Toutes ces mesures sont obtenues avec le cadre général proposé, en se donnant certaines implications et certains opérateurs d'agrégation. Les tableaux ne reportent évidemment pas l'ensemble des mesures, puisqu'il en existe une infinité. Grâce à ce tableau, on voit qu'il est extrêmement aisé de vérifier si une mesure d'inclusion est forte ($\mathcal{A}$ non disjonctif, et $I_\top$) ou faible ($\mathcal{A}$ non disjonctif), voir DÉFINITION 2.6, ainsi que la définition correcte d'une mesure de similarité (tout $\mathcal{A}$ et $I_\top$, voir Théorème 2.3). On notera cependant qu'un bon nombre de mesures proposées jusqu'alors utilisent $I_\bot$, et par conséquent ne respecte pas les conditions imposées ici. A titre d'exemple, nous donnons dans les FIG. 2.5-2.6-2.7-2.8-2.9-2.10 les visualisations de la similarité d'un ensemble flou $A = \{0.4/x_1, 0.4/x_2\}$ avec tous les autres ensembles flous possibles de dimension 2, et ce pour différentes implications $I_\top$. Comme on pouvait s'y attendre, la similarité est d'autant plus forte que les ensembles sont proches de x_1 et x_2 de A. L'utilisation de différentes implications illustre en revanche le changement de comportement selon la t-norme. Ainsi, la similarité peut très rapidement diminuer lorsque l'on s'éloigne un tant soit peu de A (par exemple Dombi pour $\gamma = 2$, Yager pour $\gamma = 2$). Inversement, la similarité peut rester très élevée même si l'ensemble comparé est loin dans le carré unité (par exemple Dombi pour $\gamma = 0.25$, Yager pour $\gamma = 0.5$). On remarque aussi un autre comportement. Certaines implications favorisent les ensembles de cardinalité floue[7] (par exemple Yager, Frank) inférieures : pour deux ensembles B et C dont la distance euclidienne à A est égale, si $|B| \leq |C|$, alors $\mathcal{S}(A,B) \geq \mathcal{S}(A,C)$. D'autres favoriseront les cardinalités supérieures (par exemple Dombi, Hamacher) : si $|B| \leq |C|$, alors $\mathcal{S}(A,B) \leq \mathcal{S}(A,C)$.

Les mesures que l'on a proposées dépendent donc d'une implication I, qui doit satisfaire certaines conditions que nous donnons (en particulier les principes de bord et confinement), et d'un opérateur d'agrégation $\mathcal{A}$ (pour les mesures de similarité et de distance), ou d'un opérateur d'agrégation conjonctif ou strictement monotone de compromis (pour les mesures d'inclusion). Le choix d'un couple spécifique$(\mathcal{A}, I)$ permet de retrouver de nombreuses mesures de la littérature.

7. Nous prenons comme définition de la cardinalité floue la définition usuelle, c'est à dire la somme sur X de $f(x)$.

Mesure d'inclusion $\mathcal{I}$	**Opérateur $\mathcal{A}$**	**Implication**
$\mathcal{I}(A,B) = \min_{x\in X}(\min(1,1-f_A(x)+f_B(x)))$, [Sinha and Dougherty, 1993]	min	$I_{\top_L}$
$\mathcal{I}(A,B) = \frac{1}{n}\sum_{x\in X}\min(1,1-f_A(x)+f_B(x))$, [Goguen, 1969]	moyenne arith.	$I_{\top_L}$
$\mathcal{I}(A,B) = \frac{1}{n}\sum_{x\in X}\max(1-f_A(x),f_B(x))$, [Dubois and Prade, 1980]	moyenne arith.	$I_{\perp_M}$
$\mathcal{I}(A,B) = \frac{1}{n}\sum_{x\in X} 1-f_A(x)+f_A(x)\,f_B(x)$, [Young, 1996]	moyenne arith.	$I_{\perp_P}$
$\mathcal{I}(A,B) = \frac{1}{n}\sum_{x\in X}\max\left(\frac{f_B(x)(\gamma+f_A(x)-\gamma f_A(x))}{f_B(x)(\gamma+f_A(x)-\gamma f_A(x))+f_A(x)-f_B(x)},1_{(f_B(x)\geq f_A(x))}\right)$,[Le Capitaine and Frélicot, 2009b]	moyenne arith.	$I_{\top_{H_\gamma}}$
$\mathcal{I}(A,B) = \frac{1}{n}\sum_{x\in X}\max\left(\left(1+\left(\left(\frac{1-f_B(x)}{f_B(x)}\right)^\gamma-\left(\frac{1-f_A(x)}{f_A(x)}\right)^\gamma\right)^{1/\gamma}\right)^{-1},1_{(f_B(x)\geq f_A(x))}\right)$, [Le Capitaine and Frélicot, 2009b]	moyenne arith.	$I_{\top_{D_\gamma}}$
$\mathcal{I}(A,B) = \frac{1}{n}\sum_{x\in X}\max\left(1-\left((1-f_B(x))^\gamma-(1-f_A(x))^\gamma\right)^{1/\gamma},1_{(f_B(x)\geq f_A(x))}\right)$	moyenne arith.	$I_{\top_{Y_\gamma}}$
$\mathcal{I}(A,B) = \frac{1}{n}\sum_{x\in X}\max\left(\log_\gamma\left(1+\frac{(\gamma^{f_B(x)}-1)(\gamma-1)}{\gamma^{f_A(x)}-1}\right),1_{(f_B(x)\geq f_A(x))}\right)$	moyenne arith.	$I_{\top_{F_\gamma}}$
$\mathcal{I}(A,B) = \frac{1}{n}\sum_{x\in X}\max\left(\max(\gamma,f_A(x))\frac{f_B(x)}{f_A(x)},1_{(f_B(x)\geq f_A(x))}\right)$	moyenne arith.	$I_{\top_{DP_\gamma}}$
...		

TAB. 2.10: Mesures d'inclusion de la littérature, ainsi que de nouvelles mesures, toutes obtenues à partir du cadré général proposé.

Mesure de similarité $\mathcal{S}$	**Opérateur $\mathcal{A}$**	**Implication**
$\mathcal{S}(A,B) = \frac{1}{n}\sum_{x\in X}\frac{\min(f_A(x),f_B(x))}{\max(f_A(x),f_B(x))}$, [Wang, 1997]	moyenne arith.	$I_{\top_P}$
$\mathcal{S}(A,B) = \frac{1}{n}\sum_{x\in X} 1-\lvert f_A(x)-f_B(x)\rvert$, [Wang, 1997]	moyenne arith.	$I_{\top_L}$
$\mathcal{S}(A,B) = \max_{x\in X}\min(f_A(x),f_B(x))$, [Chen et al., 1995]	max	$I_{\top_M}$
$\mathcal{S}(A,B) = 1-\max_{x\in X}\lvert f_A(x)-f_B(x)\rvert$, [Pappis and Karacapilidis, 1993]	min	$I_{\top_L}$
$\mathcal{S}(A,B) = \frac{1}{n}\sum_{x\in X}\frac{f_{(2)}(x)(\gamma+f_{(1)}(x)-\gamma f_{(1)}(x))}{f_{(2)}(x)(\gamma+f_{(1)}(x)-\gamma f_{(1)}(x))+f_{(1)}(x)-f_{(2)}(x)}$, [Le Capitaine and Frélicot, 2009b]	moyenne arith.	$I_{\top_{H_\gamma}}$
$\mathcal{S}(A,B) = \frac{1}{n}\sum_{x\in X}\left(1+\left(\left(\frac{1-f_{(2)}(x)}{f_{(2)}(x)}\right)^\gamma-\left(\frac{1-f_{(1)}(x)}{f_{(1)}(x)}\right)^\gamma\right)^{1/\gamma}\right)^{-1}$, [Le Capitaine and Frélicot, 2009b]	moyenne arith.	$I_{\top_{D_\gamma}}$
$\mathcal{S}(A,B) = \frac{1}{n}\sum_{x\in X} 1-\left((1-f_{(2)}(x))^\gamma-(1-f_{(1)}(x))^\gamma\right)^{1/\gamma}$	moyenne arith.	$I_{\top_{Y_\gamma}}$
$\mathcal{S}(A,B) = \frac{1}{n}\sum_{x\in X}\log_\gamma\left(1+\frac{(\gamma^{f_{(2)}(x)}-1)(\gamma-1)}{\gamma^{f_{(1)}(x)}-1}\right)$	moyenne arith.	$I_{\top_{F_\gamma}}$
$\mathcal{S}(A,B) = \frac{1}{n}\sum_{x\in X}\max(\gamma,f_{(1)}(x))\frac{f_{(2)}(x)}{f_{(1)}(x)}$	moyenne arith.	$I_{\top_{DP_\gamma}}$
...		

TAB. 2.11: Mesures de similarité de la littérature, ainsi que de nouvelles mesures, toutes obtenues à partir du cadré général proposé.

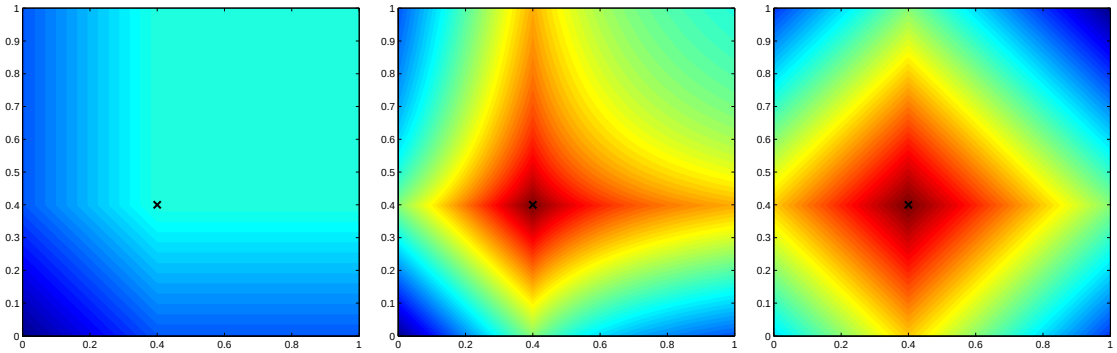

FIG. 2.5: Exemples de mesures de similarité entre $A = \{0.4/x_1, 0.4/x_2\}$ (dénoté par $\times$) et l'ensemble des ensembles flous de $\mathcal{F}(X)$, où $n = 2$. Comme à l'habitude le rouge correspond aux grandes valeurs, tandis que le bleu correspond aux petites. De gauche à droite, $\mathcal{A}$ est la moyenne arithmétique usuelle, et nous utilisons $I_{\top_M}$, $I_{\top_P}$ et $I_{\top_L}$, respectivement.

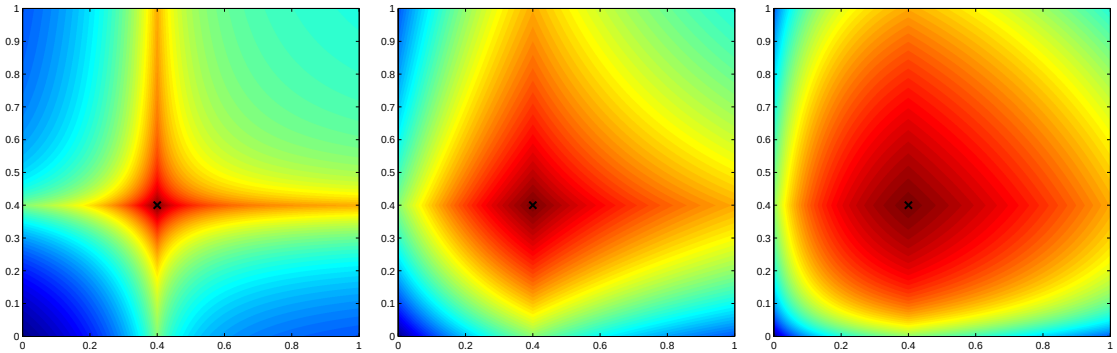

FIG. 2.6: Exemples de mesures de similarité entre $A = \{0.4/x_1, 0.4/x_2\}$ et l'ensemble des ensembles flous de $\mathcal{F}(X)$, où $n = 2$. De gauche à droite, $\mathcal{A}$ est la moyenne arithmétique usuelle, et nous utilisons $I_{\top_{H_\gamma}}$, pour $\gamma = 0, 2, 5$, respectivement.

2.3.4.1 Travaux voisins

Dans cette section, nous essayons, autant que possible, de relier de manière chronologique le cadre proposé aux différentes approches de la littérature utilisant des implications pour la définition de mesures d'inclusions. Dans leur article [Bandler and Kohout, 1980] Bandler et Kohout proposent d'utiliser des fonctions d'implication afin de quantifier l'inclusion de chaque élément dans un autre, et agrègent ensuite ces mesures individuelles par un opérateur conjonctif, le minimum. Quelques année plus tard, Hirota et Pedrycz proposent les implications pour l'appariement de quantités floues [Hirota and Pedrycz, 1991]. Ils fusionnent les différentes valeurs d'implication sur X par une intégrale de Choquet, en calculant la mesure floue μ grâce à une famille d'ensembles flous pris comme prototypes flous. Ils introduisent également une mesure d'entropie fondée sur cette mesure, qui permet de donner un aperçu de l'incertitude d'appariement. Dans [Kosko, 1992], Kosko critique la définition originelle de l'inclusion d'ensembles flous de Zadeh : B contient A si et seulement si $f_A(x) \leq f_B(x)$ pour tout x dans X, soulignant que si cette inégalité est fausse pour seulement quelques x de X, on peut toujours considérer que A est un sous-ensemble de B à un degré inférieur à 1. Il propose alors une seconde définition basée sur l'intersection

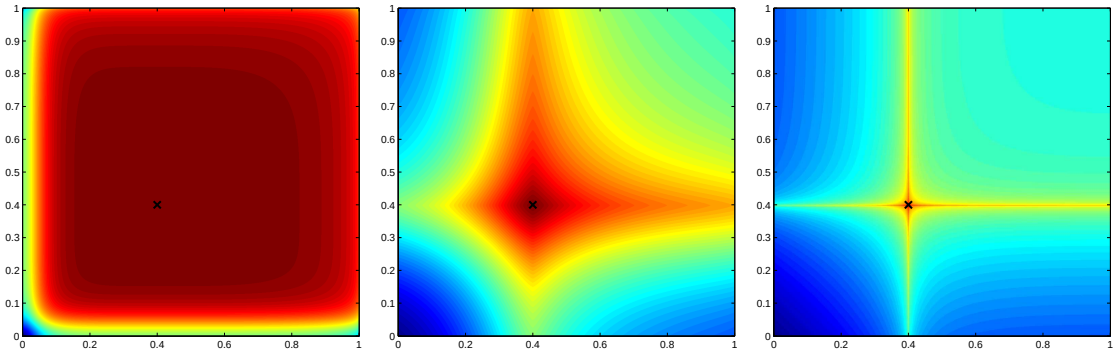

FIG. 2.7: Exemples de mesures de similarité entre $A = \{0.4/x_1, 0.4/x_2\}$ et l'ensemble des ensembles flous de $\mathcal{F}(X)$, où $n = 2$. De gauche à droite, $\mathcal{A}$ est la moyenne arithmétique usuelle, et nous utilisons $I_{\top_{D_\gamma}}$, pour $\gamma = 0.25, 0.75, 2$, respectivement.

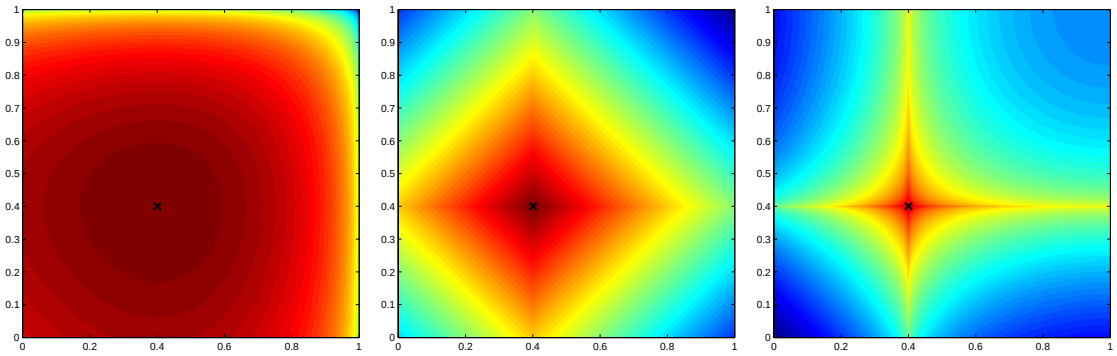

FIG. 2.8: Exemples de mesures de similarité entre $A = \{0.4/x_1, 0.4/x_2\}$ et l'ensemble des ensembles flous de $\mathcal{F}(X)$, où $n = 2$. De gauche à droite, $\mathcal{A}$ est la moyenne arithmétique usuelle, et nous utilisons $I_{\top_{Y_\gamma}}$, pour $\gamma = 0.5, 1, 2$, respectivement.

floue des deux ensembles, qui peut être vue comme la probabilité conditionnelle $P(B|A)$ dans certaines circonstances. De plus, il définit l'incertitude d'un ensemble flou A comme l'inclusion de $A \cup \overline{A}$ dans $A \cap \overline{A}$, qui satisfait les axiomes de l'entropie floue proposés par De Luca et Termini [De Luca and Termini, 1972].

Les mesures d'inclusion et de similarité d'un point de vue ensembliste, venant principalement de la proposition de Tversky [Tversky, 1977], sont décrites par Bouchon-Meunier *et al.* dans [Bouchon-Meunier et al., 1996]. Les auteurs introduisent un cadre général de mesures de comparaison, mais prennent seulement deux exemples d'opérateurs d'agrégation $\mathcal{A}$: une t-norme et l'opérateur OWA. Ici, nous proposons une étude plus détaillée du choix de $\mathcal{A}$ et des propriétés qu'il doit satisfaire. La même remarque s'applique aussi pour le travail de Young [Young, 1996]. L'auteur propose une axiomatique de mesures d'inclusion et leurs connections aux opérateurs d'implication, mais se restreint au minimum et à la moyenne arithmétique pour $\mathcal{A}$. De plus, ce travail ne donne pas de conditions nécessaires sur les implications pour la définition d'inclusions fortes ou faibles. Wang [Wang, 1997] présente deux mesures de similarité qui peuvent être vues comme des cas particuliers du cadre que l'on propose. Il élargit ensuite sa définition pour évaluer la similarité d'éléments flous appartenant à des ensembles flous. Encore une fois, le cadre que l'on propose permet d'obtenir une similarité entre éléments flous, puisque nous calculons une mesure de similarité individuelle

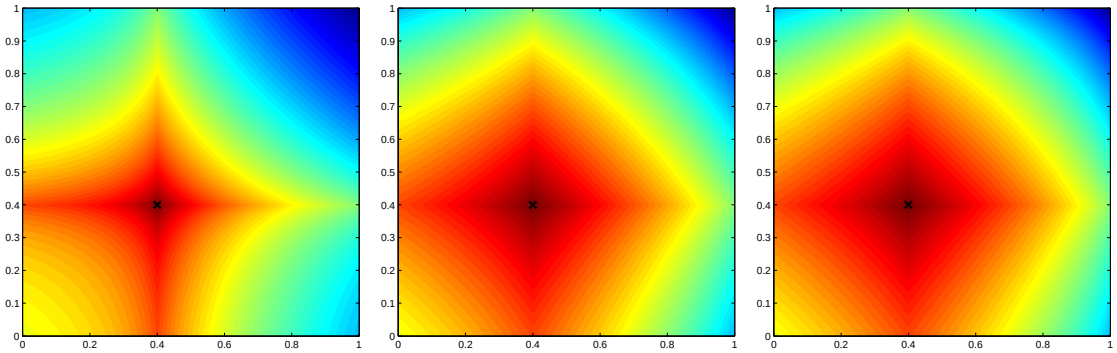

FIG. 2.9: Exemples de mesures de similarité entre $A = \{0.4/x_1, 0.4/x_2\}$ et l'ensemble des ensembles flous de $\mathcal{F}(X)$, où $n = 2$. De gauche à droite, $\mathcal{A}$ est la moyenne arithmétique usuelle, et nous utilisons $I_{\top_{F_\gamma}}$, pour $\gamma = 0.1, 5, 10$, respectivement.

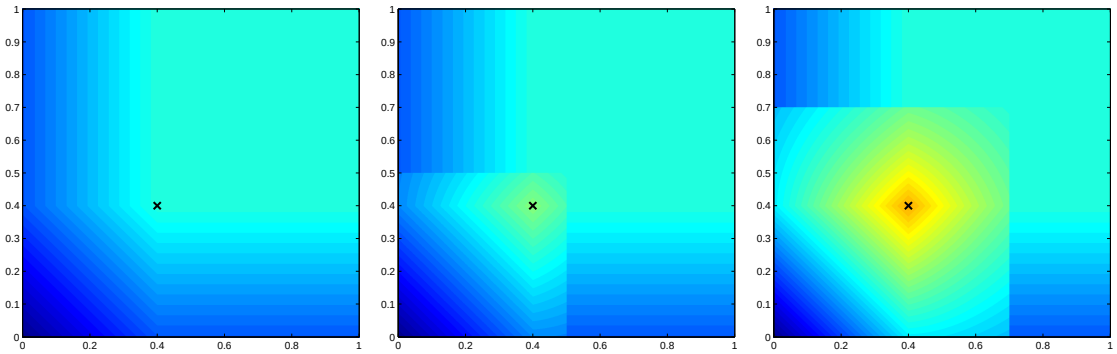

FIG. 2.10: Exemples de mesures de similarité entre $A = \{0.4/x_1, 0.4/x_2\}$ et l'ensemble des ensembles flous de $\mathcal{F}(X)$, où $n = 2$. De gauche à droite, $\mathcal{A}$ est la moyenne arithmétique usuelle, et nous utilisons $I_{\top_{DP_\gamma}}$, pour $\gamma = 0.3, 0.5, 0.7$, respectivement.

pour chaque élément (voir section 2.2.4)

À partir des observations de Kosko et Young, Botana [Botana, 1998] présente un ensemble de nouvelles mesures obtenues à partir d'implications. Il vérifie que les mesures introduites satisfont les axiomes de Young, et/ou les axiomes de Sinha et Dougherty [Sinha and Dougherty, 1993], lorsqu'il utilise les implications de Wu, Goguen, la modification de Goguen et Gödel. Il donne aussi la formulation de l'entropie correspondante au sens de [De Luca and Termini, 1972; Kosko, 1992]. Une autre approche concernant les opérateurs d'agrégation est dûe à [Fonck et al., 1998], où les valeurs d'implications sont combinées par des opérateurs conjonctifs et disjonctifs, modélisant ainsi une agrégation pessimiste et optimiste, respectivement. Fan et al. discutent des liens entre inclusion, entropie et implications floues, et proposent de nouveaux axiomes pour ces mesures [Fan et al., 1999]. Burillo et al. présentent une famille d'opérateurs d'implication dérivés de l'implication de Lukasiewicz afin de définir une famille d'opérateurs d'inclusion [Burillo et al., 2000], où l'opérateur minimum est utilisé pour l'agrégation finale.

Dans [Kehagias and Konstantinidou, 2003], Kehagias et Konstantinidou introduisent des L-mesures d'inclusion, similarité, et distance de valeur floues, c'est à dire des fonctions $\mathcal{I}$,$\mathcal{S}$ et$\mathcal{D} : \mathcal{F}(X) \times \mathcal{F}(X) \rightarrow [0,1]^n$, mais contraignent la sortie à des valeurs strictes. Comme il est souligné par les auteurs, cette sortie vectorielle peut mener à des difficultés d'interprétation.

De plus, comme les vecteurs sont partiellement ordonnés, ils sont d'autant plus difficiles à comparer, puisque cela exige une nouvelle mesure de comparaison des sorties. Le cadre proposé peut donner une sortie vectorielle, puisque chaque implication est calculée sur les éléments de X. Par contraste avec [Kehagias and Konstantinidou, 2003], ce seraient des L-mesures floues. Dans [Bustince et al., 2006], les auteurs proposent une généralisation analogue à celle-ci, mais ne traitent que les mesures d'inclusion et d'entropie, et ne donnent pas de mesures obtenues à partir de t-normes paramétriques.

Plus récemment, une distance entre opérateurs flous, en particulier les implications floues, a été proposée [Papadopoulos et al., 2007; Balopoulos et al., 2007]. Elle conduit à des normes tensorielles normalisées permettant de définir une similarité entre ensembles flous. Les auteurs présentent ensuite une heuristique pour choisir l'implication la plus adaptée à un système d'inférence floue. Zhang et Zhang introduisent une approche hybride dans [Zhang and Zhang, 2009]. Ils se restreignent à la moyenne pondérée pour $\mathcal{A}$, qui peut être considérée comme un cas particulier de notre proposition. Contrairement à Young, les implications utilisées satisfont les principes de confinement et de bord. On mentionnera enfin les travaux de Fono et al. [Fono et al., 2007], où des opérations de différence sont définies à l'aide d'implications paramétriques, permettant de proposer des mesures de comparaison pouvant être classées dans les approches ensemblistes.

2.4 Conclusion

Dans ce chapitre, après avoir introduit les mesures de comparaisons entre ensembles flous puis les différents concepts de caractérisation de vecteur à valeurs dans $[0, 1]$ (entropie, spécificité), nous avons formulé trois propositions d'opérateurs dédiés à des situations différentes. La première proposition vise à déterminer s'il existe des blocs de valeurs similaires dans un vecteur. L'optique visée est ici sensiblement la même que l'opérateur précédent, à la différence que l'on souhaite localiser ces valeurs, et qu'elles ne sont pas forcément les plus fortes. L'autre apport par rapport à l'approche précédente est l'ajout de noyaux permettant une prise en compte progressive des valeurs au sein du bloc analysé. Ici encore, les applications sont nombreuses, comme nous l'avons détaillé en section 2.2.3.

Les deux dernières propositions sont différentes. On part cette fois-ci d'opérateurs logiques (ici des implications), et, utilisant le fait qu'une similarité peut être vue comme l'implication réciproque de valeurs, une mesure de similarité est construite. Dans un premier temps utilisée sur des valeurs numériques dans $[0, 1]$, nous étendons ensuite l'opérateur aux ensembles flous. La possibilité d'utiliser des normes triangulaires pour la génération de mesures d'implication autorise en outre l'obtention de familles paramétriques de mesures. Le comportement de chacune de ces familles est analysé en fonction de l'évolution du paramètre, et permet ainsi de choisir plus facilement sa mesure parmi les nombreuses (infinies) possibles. Une fois de plus, l'utilisation de ces opérateurs dans la pratique permet d'aborder de nombreux domaines. Ces résultats peuvent être utilisés dans des domaines où des fonctions de comparaison sont nécessaires. Nous citerons, entre autres, les applications qui nous intéressent : la morphologie mathématique floue [Burillo et al., 2000], la validation de partitions [Fan et al., 1999], ou encore la sélection de variables en classification, l'apparie-

ment d'images dans des bases de données, la combinaison de classifieurs [Kuncheva, 2001; Kuncheva et al., 2001].

Deuxième partie

Applications

Chapitre 3

Reconnaissance de formes

Si l'activité inconsciente de l'esprit consiste à imposer des formes à un contenu, et si ces formes sont fondamentalement les mêmes pour tous les esprits, anciens et modernes, primitifs et civilisés, comme l'étude de la fonction symbolique, il faut et il suffit d'atteindre la structure inconsciente, sous jacente à chaque institution et à chaque coutume, pour obtenir un principe d'interprétation valide pour d'autres institutions et d'autres coutumes.

— Claude Lévi-Strauss
Anthropologie Structurale (1954)

Résumé : *Dans ce chapitre, nous abordons la problématique générale de la reconnaissance de formes. Nous essaierons autant que possible de décrire le fonctionnement global d'un système de reconnaissance automatique. Nous ne présentons évidemment pas l'ensemble des techniques pouvant être mises en œuvre, tant le sujet est vaste. Nous nous concentrons plutôt sur des méthodes pouvant se rapprocher du sujet d'étude principal de ce mémoire et pour lesquelles des liens peuvent être établis.*

3.1 Introduction

L'augmentation phénoménale de données numériques due à l'automatisation de systèmes opérationnels dans des domaines d'activités tels que le commerce, l'industrie, ou encore l'environnement, ainsi que les avancées technologiques en matière de stockage d'information durant ces dernières années a abouti à un nombre gigantesque d'informations, parfois hétérogènes, stockées dans des bases de données. En extraire de l'information pertinente et l'analyser est devenu une problématique importante tant dans les recherches récentes que dans l'industrie. Trouver des relations, des corrélations, des régularités ou encore des tendances dans ces données est donc capital à de nombreux points de vue. La reconnaissance de formes est un domaine de recherche visant à fournir des méthodes permettant la fouille de données et ainsi aider les humains à analyser les structures de données complexes que l'on peut trouver, et ce de manière automatique. Bien que la reconnaissance d'un objet, d'une forme soit une action opérée presque instantanément et de manière multiple par l'être

humain, confier cette tâche à un système automatique s'avère très difficile. Le principe général de la reconnaissance automatique passe par la mise en place d'algorithmes analysant la structure des données et prenant des décisions concernant la *classification* de chaque observation parmi des catégories, ou *classes*.

3.1.1 Fonctionnement général

Ainsi, la reconnaissance de formes (rdf) est une discipline scientifique ayant pour objectif le classement d'observations, ou d'individus en un certain nombre de classes de manière automatique. Le terme automatique a son importance, puisque il contraint à l'utilisation de machine, et l'essor de cette discipline a vraiment démarré dans les années soixante, avec l'apparition de l'informatique. De manière générale l'observation que l'on souhaite classer est décrite par un ensemble de mesures prises avec un système physique (nous ne considérons pas le cas où les observations sont qualitatives). Ce problème de reconnaissance peut être formalisé par une fonction $\mathbf{y}$ qui à une entrée $\mathbf{x}$ décrite dans un espace d'attributs E de dimension p, associe une sortie s dans S, représentant l'ensemble des affectations possibles :

$$\mathbf{y} : E \rightarrow S, \tag{3.1}$$

$$\mathbf{y}(\mathbf{x}) \mapsto s \tag{3.2}$$

Cette fonction $\mathbf{y}$ sera appelée le *classifieur*. Généralement, afin de simplifier le problème, on formule l'hypothèse que les observations appartenant à une même classe sont regroupées dans l'espace d'attributs, et qu'elles forment ce qu'on appelle un *cluster* de taille et de forme à déterminer. On séparera ainsi chaque cluster par des *frontières de décision* afin de discriminer chaque nouvelle observation. Celles-ci sont obtenues à partir de connaissances que l'on possède du modèle, et peuvent être de deux natures. La première est la connaissance a priori que l'on obtient généralement grâce à un expert du domaine. La seconde est obtenue de manière automatique par un processus d'apprentissage. Elle consiste, à partir d'exemples composant la base d'apprentissage, à évaluer, par exemple, la taille et la forme de chaque cluster. Une phase de test, impliquant des observations pour lesquelles leurs groupe d'appartenance n'est pas connu, est ensuite réalisée, afin de vérifier la faculté de *généralisation* du classifieur $\mathbf{y}$, [Duda et al., 2001; Theodoridis and Koutroumbas, 2006].

Afin de clarifier la lecture future de ce chapitre, un point sur les notations utilisées est nécessaire.

- on considère n observations appartenant à un ensemble χ.
- chacune de ces observations est décrite par p attributs.
- le vecteur de caractérisation de chaque observation sera écrit $\mathbf{x}_k \in \chi$, appartenant donc à $\mathbb{R}^p$, où k correspond à l'indice de l'observation, et nécessairement $k \in [1, n]$.
- chaque observation $\mathbf{x}_k$ appartient à une classe ω_i. L'ensemble des classes forme $\Omega = \{\omega_1, \cdots, \omega_c\}$, où c dénote le nombre de classes en présence. On notera que l'on apportera quelques modifications à ce point de vue dans la suite du manuscrit, puisque d'une part on introduira la possibilité pour une observation d'appartenir à un nombre entier quelconque de classes (Chapitre 4), et que d'autre part on considèrera le cas où le nombre c de classes n'est pas toujours connu (Chapitre 5).

Une fois défini l'objectif général de la classification, nous pouvons maintenant décrire les différentes étapes qui constituent le processus de classification, et les questions qu'elles engendrent.

- Comment obtenir des variables descriptives? Ceci correspond à la génération de variables (3.1.1.2).
- Parmi ces descripteurs, combien et lesquels utiliser? C'est la sélection de variables (3.1.1.3).
- Comment construire le classifieur (3.1.1.4)?
- Comment déterminer si le classifieur construit est performant (3.1.1.5)?

La Fig. 3.1 représente le schéma général de fonctionnement d'un système de reconnaissance de formes [Theodoridis and Koutroumbas, 2006].

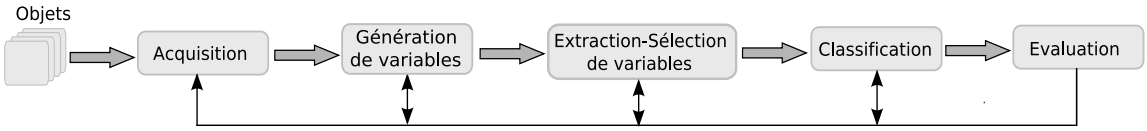

Fig. 3.1: Schéma de fonctionnement d'un système de reconnaissance de formes.

3.1.1.1 Acquisition des données

Bien que le type de classifieur utilisé soit d'une grande importance, obtenir des données adéquates et représentatives du problème à résoudre l'est tout autant. Le terme adéquation suppose qu'il existe des données en quantité suffisante pour construire les frontières de décision entre les classes. Il n'existe pas de règles fixant le nombre d'observations, mais il semble, de manière expérimentale, que le nombre d'observations doive être supérieur à 10 fois le nombre d'attributs (i.e. $n > 10 * p$) [Jain et al., 2000]. Bien que l'importance de chacune des caractéristiques ne soit pas prise en compte lors de cette étape, il faut tout de même veiller aux incertitudes d'acquisition, puisque des incohérences peuvent apparaître. De plus, il faudra être attentif au fait de capturer les variations possibles au sein de l'ensemble de l'apprentissage, en particulier faire en sorte que l'ensemble des sorties (ou classes) soit représentées.

3.1.1.2 Génération de variables

Lorsque les données ont été acquises, elles peuvent ne pas être directement exploitables. Un exemple typique de ce genre de situation est le traitement d'image. Un algorithme considérant l'ensemble des pixels d'une image n'est pas très efficace, dans la mesure où la dimension des données explose rapidement, d'autant plus que les informations véhiculées par les pixels voisins sont souvent redondantes. On procède alors à un pré-traitement permettant d'extraire par exemple les contours des objets. Une fois ces contours obtenus, ils seront ensuite décrits par des caractéristiques qui permettront la classification.

Les nouvelles caractéristiques sont souvent obtenues grâce à un expert du domaine. Certaines sont ainsi conservées tandis que d'autres sont purement supprimées car elles ne

permettent pas de discriminer les classes et sont donc jugées non pertinentes. Comme il été précisé dans l'introduction, une fois les variables engendrées, nous disposons de *vecteurs-forme* $\mathbf{x}$ faisant partie d'un ensemble d'observations. Une manière de représenter ces données est donc une matrice X de dimension $n \times p$.

3.1.1.3 Sélection et extraction de variables

Le nombre de variables descriptives peut être important, et il convient de procéder à une réduction de la dimension du problème. Cette réduction présente deux avantages distincts. Le premier est évidemment le gain de temps, puisque le volume de données traitées est moins important. Le second avantage réside dans le fait que lorsque l'on possède un échantillon réduit, (n faible), réduire le nombre d'attributs permet de s'affranchir de la malédiction de la dimension (*curse of dimensionality*), voir section 3.4.1.

On distingue deux cas parmi les méthodes de réduction de dimension. Soit on cherche à sélectionner parmi les p attributs un nombre q strictement inférieur à p de telle manière à optimiser un critère que l'on se fixe, soit on procède à une transformation (linéaire ou non) des données en les projetant dans un espace de dimension inférieure. La première solution correspond au processus de *sélection de variables*, tandis que la seconde est appelée *extraction de variables*.

- Sélection de variables : celle-ci permet d'identifier les variables ne contribuant pas à la séparation des classes, et qui sont donc jugées non pertinentes. Il s'agit ici de sélectionner un sous-ensemble S_q de q variables parmi les p originales (S_p), selon un certain critère J à optimiser que l'on fixe. Dans de nombreuses applications, diminuer le nombre de variables descriptives conduit à de meilleures performances en terme de classification. Les deux problématiques de cette approche sont : dans la formulation de critères de sélection et la méthode de recherche, ou parcours, de l'ensemble des sous-ensembles possibles. S'il semble logique d'utiliser le taux de bonne classification comme critère, cela rend la méthode dépendante du classifieur et de la taille des ensembles d'apprentissage et de test [Jain et al., 2000]. Le lecteur pourra consulter [Semani, 2004] pour un panorama des critères utilisés dans la littérature. Pour la méthode de recherche, l'approche naïve consiste à explorer l'ensemble de tous les sous-ensembles possibles, mais cela conduit à une explosion des calculs puisque pour q variables désirées, il faut évaluer C_p^q combinaisons. Des méthodes sous-optimales permettant tout de même d'évaluer les interactions entre attributs, citons les *Sequential Forward Selection* (SFS) [Devijver and Kittler, 1982], ou plus récentes et performantes les *Sequential Forward Floating Search* (SFFS) [Pudil et al., 1994].
- Extraction de variables : brièvement, l'extraction de variables consiste à trouver un nombre restreint de variables particulièrement informative, et invariantes à diverses transformations. Une méthode classiquement utilisée pour ce genre d'opération est la transformée de Karhunen-Loève, aussi connue sous le nom d'Analyse en Composantes Principales (ACP). Il s'agit de projeter les données de manière à obtenir les axes principaux où la variance des données est la plus importante. Une autre méthode, appelée Analyse Factorielle Discriminante (AFD), consiste à projeter les données afin de minimiser l'inertie intra-classes et de maximiser l'inertie inter-classes. Cette dernière

présente l'avantage de s'assurer de la séparation des classes dans l'espace de projection, à la différence de l'ACP. À noter que l'on a vu ces dernières années de nombreuses variantes apparaitre, parmi elles, on citera l'*Independent Component Analysis* [Hyvärinen et al., 2001], ou encore le *Kernel PCA* [Schölkopf et al., 1998]. Bien que ces transformations soient communément utilisées, il peut arriver qu'elles ne soient pas les meilleures pour toutes les applications. On pourra en effet trouver des situations où l'information discriminante se situe dans le domaine spectral, auquel cas la transformée de Fourier, cosinus ou en ondelettes, seront plus performantes [Akay, 1998].

3.1.1.4 Construction du classifieur

Dès lors que des caractéristiques communes à chaque objet observé sont disponibles, on peut alors procéder à la classification. On distingue deux grandes familles de méthodes de classification : les méthodes supervisées, et les méthodes non supervisées. Tandis que l'on dispose d'un ensemble de points étiquetés dans le premier cas, on ne connait absolument pas l'appartenance aux classes dans le second. Dans le domaine supervisé, une grande difficulté lors de la conception d'un classifieur consiste à choisir la modélisation sur laquelle il repose. On distingue deux approches duales :

- **la modélisation discriminante** : définition des classes par la recherche de frontières à l'aide de fonctions discriminantes.
- **la modélisation intrinsèque, ou par partition** : définition des classes par leur propriétés intrinsèques, souvent à l'aide de **prototypes**.

Ces approches différentes sont illustrées dans la Fig. 3.2 dans le cas simple de trois classes bi-dimensionnelles linéairement séparables.

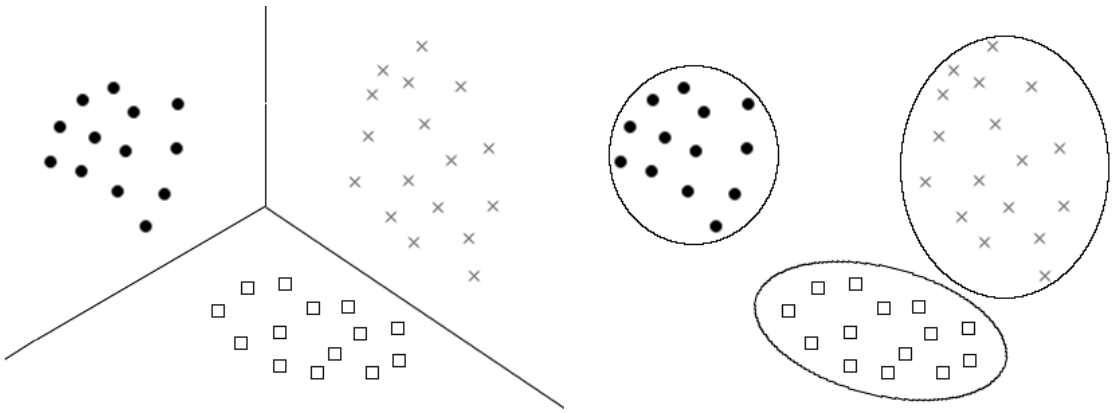

Fig. 3.2: Deux approches différentes de la classification : discriminante (*gauche*) et par partition (*droite*).

On cherche dans la modélisation discriminante à trouver les frontières (généralement des hypersurfaces) séparant au mieux les classes dans l'espace d'attributs. On trouve dans ce genre d'approche des méthodes de types AFD, ou encore les Séparateurs à Vaste Marge (SVM – Support Vector Machines [Cortes and Vapnik., 1995]) et les réseaux de neurones. Il suffit alors, pour la classification, de déterminer de quel côté se trouve la forme.

Dans le cas de la modélisation par partition, on essaie autant que possible de décrire les classes grâce à des propriétés les caractérisants dans l'espace d'attributs. On utilise généra-

lement des prototypes de classes dans ce cas. Par exemple, une classe peut être caractérisée par un centre et une matrice de covariance. Pour procéder à la classification, il suffira alors de comparer, pour la forme à classer, sa probabilité, ou son degré d'appartenance par rapport à l'ensemble des prototypes, et sélectionner le plus raisonnable (généralement en terme d'erreur de classification, voir section 3.1.1.5). On trouvera aussi le terme de modélisation générative pour ce type de classifieur, puisque on pourra générer de nouvelles formes correspondant au modèle à partir de la description initiale.

Dans ce travail, nous proposons de nouveaux outils pour deux problèmes particuliers qui apparaissent dans les méthodes supervisées pour l'un, et dans les méthodes non supervisées pour l'autre. C'est pourquoi nous consacrerons ce chapitre à l'étude et la description de ces deux approches.

3.1.1.5 Évaluation des performances

Lorsque le classifieur a été mis en place, il est nécessaire de vérifier ses performances avec une base de test. En effet, un classifieur peut être très performant sur la base d'apprentissage, mais donner des performances médiocres sur de nouvelles données. On dira que ce type de classifieur aura une capacité de généralisation faible. À l'inverse, un classifieur peut être formulé de manière trop simple, et ne permettra pas une discrimination suffisante des données.

On introduit ainsi une base de test n'ayant pas servi à l'apprentissage pour estimer l'erreur réelle de classement. Cette estimation peut se faire selon plusieurs procédures, où l'on divise la base à disposition en deux bases d'apprentissage et de test :

- **Resubstitution** : ici le même ensemble est considéré pour l'apprentissage, puis pour la phase de test. Évidemment, utiliser les mêmes données pour l'apprentissage et le test conduit à une évaluation optimiste des erreurs commises par le classifieur. Cette méthode d'estimation introduit un biais qui est fonction du rapport n/p, et la variance de l'estimation est inversement proportionnelle à n. Il faudra donc une valeur de n importante et une valeur de p assez faible pour obtenir une bonne estimation. Le taux d'erreur se calcul enfin de la manière suivante

$$E = \frac{n_e}{n} \tag{3.3}$$

  où n_e est le nombre d'erreurs commises par le classifieur sur les n observations.
- **Holdout** : dans ce cas, on partitionne l'ensemble χ des observations en deux sous-ensembles d'apprentissage et de test. L'inconvénient de cette méthode est évidemment que l'on dispose de moins d'exemples pour l'apprentissage, et de moins de données à tester. Un autre problème consiste à décider du rapport de la taille de l'ensemble d'apprentissage par rapport à celle de l'ensemble de test. Ne disposant d'aucun résultat théorique à ce sujet, la coutume a voulu que l'ensemble soit divisé en deux, ou d'un rapport 2/3-1/3 pour les ensembles d'apprentissage-test, respectivement. Le taux d'erreur est calculé sur l'ensemble de test par

$$E = \frac{n_e}{n_t} \tag{3.4}$$

  où n_t est la taille de l'ensemble de test.

– **Leave-one-out (LOO)** : cette méthode d'estimation d'erreur permet de s'affranchir des limitations des deux précédentes. Il s'agit ici de considérer n ensembles d'apprentissages constitués par $n-1$ observations, et de tester l'observation restante. On répète cette opération n fois, en excluant à chaque fois une observation différente. Le nombre total d'erreur permet donc de calculer le taux par l'équation (3.3). L'avantage principal de cette technique réside dans l'utilisation d'un nombre maximal de données d'apprentissage tout en conservant le principe d'indépendance entre les ensembles d'apprentissage et de test, mais au prix d'une complexité importante. C'est donc la méthode à privilégier pour des bases de petite taille.

– **Validation croisée** : la validation croisée est en fait une généralisation de la méthode précédente. Lorsque la base est de taille importante le temps de calcul requis par la méthode LOO est tel qu'il fut suggéré de sélectionner plusieurs observations en même temps au lieu d'une seule. Ainsi, l'ensemble est partitionné en D parties de tailles sensiblement égales. L'ensemble d'apprentissage est ensuite constitué de $D-1$ parties, et l'ensemble de test est la D-ème partie. On répète cette procédure de manière à ce que chacune des D parties soit un ensemble test. L'estimation de l'erreur s'obtient alors par

$$E = \frac{\sum_{i=1}^{D} n_e(i)}{n} \tag{3.5}$$

où $n_e(i)$ est le nombre d'erreurs de classement commises sur la i-ème partie. Évidemment, si l'on fixe $D = n$, on retrouve la méthode LOO, et la méthode de resubstitution si $D = 1$. Cette méthode est privilégiée dans le cas de grandes bases, et la valeur la plus couramment utilisée pour D est 10.

– **Bootstrap** [Efron, 1983] : avec cette technique, au lieu de répéter l'analyse de sous-ensembles de données, on répète l'analyse de sous-échantillons des données. En d'autres termes, on tire avec remise n observations dans l'ensemble de départ, et l'ensemble de test est χ. On répète alors cette procédure plusieurs fois, l'estimation d'erreur étant obtenue par la moyenne des taux d'erreurs successifs.

3.1.2 Les différentes approches

Parmi l'ensemble des approches que l'on pourra trouver en reconnaissance de formes, on en distinguera quatre :

3.1.2.1 Approches structurelles

Le mot structurel est ici employé dans le sens où chaque observation est définie comme une structure composée de différentes sous-structures, elle-mêmes composées de structures, jusqu'à parvenir à la description complète de l'objet par un ensemble restreint de primitives. Cette approche hiérarchique est intéressante et permet ainsi de décrire la construction de chaque forme. Le fait d'utiliser des primitives communes à l'ensemble du problème permet en outre de relier cette approche à la syntaxe du langage. Une grammaire est alors mise en place et conduit au traitement de la forme pour sa classification. Cette vision grammaticale a d'ailleurs conduit à dénommer ces méthodes sous le nom de *syntaxiques*. Pour plus d'informations on pourra consulter le livre [Bunke and Sanfeliu, 1990].

3.1.2.2 Approches statistiques

Les méthodes statistiques, contrairement aux méthodes syntaxiques et structurelles, utilisent la représentation des observations sous la forme d'un vecteur de p attributs les décrivant le plus précisément possible. Ce choix de description la plus "précise" possible permet ensuite de pouvoir séparer les formes de différentes classes dans cet espace de représentation. Des notions de décision statistique sont ensuite utilisées afin de déterminer les frontières de décision entre les classes. Au sein de ces méthodes statistiques, on distinguera les méthodes paramétriques et non paramétriques (voir section 3.2). Nos travaux de recherche s'inscrivent dans ce cadre. La littérature à ce sujet étant très abondante, nous suggérons au lecteur de se référer à la suite du manuscrit, et à [Jain et al., 2000; Webb, 2002; Bishop, 2006] pour de plus amples détails.

3.1.2.3 Approches connexionnistes

Le terme connexionniste est employé pour caractériser les méthodes se basant sur des réseaux de neurones. On retrouvera en reconnaissance de formes beaucoup de propositions utilisant ce concept. Brièvement, il s'agit d'un grand nombre de processeurs mis en place de manière parallèle, mais disposant de nombreuses inter connections. Il s'agit en fait d'un graphe orienté pondéré où les nœuds possèdent des fonctions d'activations. On citera parmi les propositions les perceptrons multi-couches, les fonctions à base radiale, ou encore les cartes auto-organisatrices, voir [Bishop, 1995; Ripley, 1996]. On trouve dans les méthodes statistiques des liens avec les réseaux de neurones. De nombreux auteurs considèrent d'ailleurs les réseaux de neurones comme faisant partie des méthodes statistiques.

3.1.2.4 Mise en correspondance

Alors que les autres approches visent à apparier une forme à l'une des classes du problème, la mise en correspondance de formes repose sur un concept différent. On dispose dans ce cas de formes de référence, et l'on doit décider, parmi cet ensemble de formes référentes, laquelle est la plus proche de la forme à classer. La première chose à faire dans ce type d'approche est de mettre en place une mesure de similarité entre les formes. Une simple distance euclidienne ne suffisant bien souvent pas à obtenir de bons résultats, des méthodes fondées sur la recherche de chemins optimaux ou sur la corrélation des formes ont été proposées. Cette approche est couramment utilisée en traitement d'images, voir [Brunelli, 2009].

Nous distinguerons dans ce chapitre deux catégories de méthodes, qui sont en fait deux situations différentes. La première correspond à la classification *supervisée*. On dispose ici d'exemples, où la classe d'assignation est connue, et l'on pourra donc baser l'apprentissage sur cette connaissance a priori. Ceci fera l'objet de la section 3.2. Alternativement, lorsque nous ne disposons pas de cette connaissance, la classification est dite *non supervisée*. Dans ce cas, il s'agit de découvrir des similarités entre les observations afin de les regrouper dans des sous-ensembles appelés *clusters*. Cette approche et les méthodes associées seront décrites dans la section 3.3.

3.2 Classification supervisée

Dans un cas simple et standard, l'objectif de la classification supervisée est de mettre en place des règles de décision pour une observation $\mathbf{x}$. Ces règles doivent être optimales au sens de la séparation des classes mutuellement exclusives de Ω. La décision peut donc être vue comme une application $d : \mathbb{R}^p \to \Omega, \mathbf{x} \mapsto \omega_i$, c'est à dire un cas particulier de l'équation (3.1).

3.2.1 Règle de Bayes

En supposant que l'on a une observation $\mathbf{x}$, et que l'on souhaite déterminer à quelle classe ω_i elle appartient. On s'intéresse donc à la probabilité conditionnelle $P(\omega_i|\mathbf{x})$. Celle-ci peut s'obtenir grâce au théorème de Bayes :

$$P(\omega_i|\mathbf{x}) = \frac{P(\omega_i)P(\mathbf{x}|\omega_i)}{P(\mathbf{x})} \tag{3.6}$$

où

- $P(\omega_i|\mathbf{x})$ est la probabilité a posteriori de la classe ω_i conditionnellement à l'observation de $\mathbf{x}$.
- $P(\omega_i)$ est la probabilité a priori de la classe ω_i
- $P(\mathbf{x}|\omega_i)$ est la densité de probabilité en x conditionnée par la classe ω_i
- $P(\mathbf{x})$ est la densité mélange de $\mathbf{x}$ définie par :

$$P(\mathbf{x}) = \sum_{i=1}^{c} P(\omega_i)P(\mathbf{x}|\omega_i) \tag{3.7}$$

Dans ce cadre, les classe sont supposées mutuellement exclusives, de sorte que

$$\sum_{i=1}^{c} P(\omega_i) = 1 \tag{3.8}$$

d'où le fait que la somme des probabilités a posteriori soit égale à 1. Si l'on se définit une fonction de coût (ou perte) associée à chaque décision $\mathbf{x} \mapsto \omega_i$, le but est alors de minimiser la perte totale engendrée. Certains auteurs définissent une fonction d'utilité et la maximisent, mais cela reste un concept équivalent. Cette fonction peut être vue comme une matrice de coût C dont le terme général C_{ij} désigne le coût associé à la décision j alors que la décision i devrait être prise. La convention est de choisir des coûts tels que $C_{ij} \geq C_{ii}$, et même le plus souvent $C_{ii} = 0$, signifiant ainsi que lorsque la décision prise est correcte, et le coût est nul, et qu'il coûte plus cher de se tromper. Dans ce cadre, la solution optimale consiste à minimiser la fonction de coût. Pourtant, celle-ci dépend de la classe de l'objet, qui est inconnue. On cherche donc à minimiser le coût moyen :

$$\mathbb{E}[C] = \sum_i \sum_j \int_{\omega_j} C_{ij} P(\mathbf{x},\omega_i) \mathrm{d}\mathbf{x} \tag{3.9}$$

Comme on cherche à trouver une décision ω_j la minimisation de (3.9) revient à la minimisation de

$$\sum_i C_{ij} P(\mathbf{x},\omega_i) \tag{3.10}$$

On sait que $P(\mathbf{x},\omega_i) = P(\mathbf{x})P(\omega_i|\mathbf{x})$, et puisque $P(\mathbf{x})$ est un facteur commun, on peut le supprimer lors de la minimisation, donnant

$$R(\omega_j|\mathbf{x}) = \sum_i C_{ij} P(\omega_i|\mathbf{x}) \tag{3.11}$$

qui est le risque conditionnel associé à la règle de Bayes. On choisira donc la classe ω_j telle que

$$j = \text{argmin}_i R(\omega_i|\mathbf{x}) \tag{3.12}$$

Si la distribution des données ainsi que la matrice de coût sont connues, la règle de Bayes ainsi énoncée est optimale, c'est à dire que l'erreur de Bayes est la plus faible que l'on puisse obtenir. Pourtant, dans des problèmes pratiques de reconnaissance de formes, la distribution des données est toujours inconnue et doit donc être estimée. Parmi les méthodes utilisées pour cela, on distinguera deux approches. L'approche paramétrique, où l'on émet des hypothèses sur le modèle que suivent les données, et inversement les méthodes non paramétriques, où aucune hypothèse n'est posée.

3.2.2 Estimations paramétriques

L'approche paramétrique consiste à choisir un modèle de distribution, puis d'en estimer les paramètres grâce aux données. Par souci de simplicité, et comme on le suppose souvent dans la littérature, les données suivent une distribution normale (gaussienne) définie sur un vecteur $\mathbf{x}$ de dimension p par

$$P(\mathbf{x}|\mathbf{v},\mathbf{\Sigma}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{v})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{v}) \right) \tag{3.13}$$

où le vecteur $\mathbf{v}$ est la moyenne, la matrice $\mathbf{\Sigma}$ de taille $p \times p$ est la covariance, et $|\mathbf{\Sigma}|$ dénote son déterminant, et enfin $(\mathbf{x}-\mathbf{v})^T$ le transposé de $(\mathbf{x}-\mathbf{v})$.

3.2.2.1 Maximum de vraisemblance

Une interprétation du théorème de Bayes peut nous conduire à l'écrire de la façon *a posteriori* $\propto$ *vraisemblance*$\times$*a priori*. Un critère généralement adopté consiste à maximiser la vraisemblance. Ceci peut paraître étrange, car on aurait plutôt tendance à maximiser la probabilité des paramètres conditionnellement aux données, plutôt que la probabilité des données conditionnellement aux paramètres. Pourtant ces deux critères sont liés. Si l'on a $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$, n réalisations indépendantes de la distribution donnée par (3.13), la loi de vraisemblance s'écrit donc

$$\mathcal{L}(X|\mathbf{v},\mathbf{\Sigma}) = P(X|\mathbf{v},\mathbf{\Sigma}) = \prod_{k=1}^{n} \mathcal{N}(\mathbf{x}_k|\mathbf{v},\mathbf{\Sigma}) \tag{3.14}$$

Afin de trouver les valeurs des paramètres $\mathbf{v}$ et $\mathbf{\Sigma}$, on maximise cette vraisemblance. En pratique, il est plus simple de maximiser le logarithme, fonction croissante monotone permettant de passer du produit à la somme :

$$\log P(X|\mathbf{v},\mathbf{\Sigma}) = \sum_{k=1}^{n} \log P(\mathbf{x}_k|\mathbf{v},\mathbf{\Sigma}) \tag{3.15}$$

La maximisation de (3.15) par rapport à $\mathbf{v}$ et $\mathbf{\Sigma}$ donne, respectivement pour la moyenne et la covariance :

$$\hat{\mathbf{v}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \tag{3.16}$$

et

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\mathbf{v}})(\mathbf{x}_k - \hat{\mathbf{v}})^T \tag{3.17}$$

ou

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\mathbf{v}})(\mathbf{x}_k - \hat{\mathbf{v}})^T \tag{3.18}$$

pour une version non biaisée. L'estimation par maximum de vraisemblance est ainsi non biaisée, présente une distribution normale et une variance minimale, à condition toutefois que n soit assez grand.

3.2.2.2 Maximum a posteriori

Alors que l'on maximisait $P(X|\mathbf{v},\mathbf{\Sigma})$ pour la méthode précédente, l'estimateur du *Maximum A Posteriori* (MAP) cherche à maximiser $P(X|\mathbf{v}_i,\mathbf{\Sigma}_i)P(\omega_i)$, et cherche donc un pic de la densité a posteriori. C'est un cas plus général que le maximum de vraisemblance, où les classes sont supposées équiprobables. On estime dans un premier temps la probabilité a priori de chacune des classes ω_i par

$$\hat{P}(\omega_i) = \frac{n_i}{n} \tag{3.19}$$

où n_i est l'effectif de la classe ω_i. Les vecteurs moyennes et les matrices de covariance sont ensuite donnés par

$$\hat{\mathbf{v}}_i = \frac{1}{n_i}\sum_{\mathbf{x}_k\in\omega_i}\mathbf{x}_k \tag{3.20}$$

$$\hat{\mathbf{\Sigma}}_i = \frac{1}{n_i-1}\sum_{\mathbf{x}_k\in\omega_i}(\mathbf{x}_k - \hat{\mathbf{v}}_\mathbf{i})(\mathbf{x}_k - \hat{\mathbf{v}}_\mathbf{i})^T \tag{3.21}$$

respectivement. La règle MAP consiste enfin à prendre la classe ω_i qui maximise la probabilité a posteriori (3.6) où $P(\omega_i|\mathbf{x}) = \mathcal{N}(\mathbf{x}|\hat{\mathbf{v}}_i,\hat{\mathbf{\Sigma}}_i)$ dans le cas gaussien.

3.2.3 Approches non paramétriques

3.2.3.1 Estimation de densité

Jusqu'à présent, nous avons traité le cas de la classification supervisée, et nous supposions connue la fonction de densité. Il s'agissait ensuite d'estimer les paramètres des modèles qui avaient été choisis. Un défaut important de cette approche est que la densité choisie peut être un mauvais modèle de la distribution réelle des données, impliquant ainsi de mauvaises performances. En particulier, la plupart des modèles de densités sont unimodaux, alors qu'en pratique, on observe souvent des densités multimodales. La plupart des techniques d'estimation de densité se basent sur le fait que la probabilité P qu'une forme $\mathbf{x}$ soit dans une région $\mathcal{R}$ est donnée par

$$P = \int_{\mathcal{R}} P(\mathbf{x})\mathrm{d}\mathbf{x} \tag{3.22}$$

Si l'on estime maintenant que $P(\mathbf{x})$ est continue et que la région $\mathcal{R}$ est suffisamment petite, on peut écrire

$$P \simeq P(\mathbf{x})V \tag{3.23}$$

où V est le volume de la région $\mathcal{R}$ donné par

$$V = \int_{\mathcal{R}} \mathrm{d}\mathbf{x} \tag{3.24}$$

En utilisant l'approximation, pour un grand n, que $P \simeq k/n$, où k est le nombre de points dans la région $\mathcal{R}$, on obtient l'estimation de la densité $P(\mathbf{x})$:

$$P(\mathbf{x}) \simeq \frac{k}{nV} \tag{3.25}$$

On pourra constater que cette estimation tend vers la vraie valeur de $P(\mathbf{x})$ lorsque n tend vers l'infini. On peut exploiter le résultat (3.25) de deux manières. Soit on fixe la volume V, et on détermine k grâce aux données, c'est la méthode des noyaux de Parzen (voir 3.2.3.1), soit on fixe k, et on détermine V grâce aux données, c'est la méthode des plus proches voisins (voir 3.2.3.1).

Estimation par noyaux de Parzen Nous voulons ici, à partir d'une région $\mathcal{R}$ centrée en $\mathbf{x}$, déterminer le nombre de points appartenant au volume de dimension p associé à cette région. Une fonction noyau φ est donc introduite; celle-ci vaut 1 lorsque si le point $\mathbf{x}$ appartient au volume, 0 sinon. La région $\mathcal{R}$ étant généralement définie comme un hypercube de côté h, on trouve ainsi

$$k = \sum_{k=1}^{n} \varphi\Big(\frac{\mathbf{x} - \mathbf{x}_k}{h}\Big) \tag{3.26}$$

qui permet d'obtenir $P(\mathbf{x})$ grâce à (3.25). Le paramètre h joue un rôle de lissage : une faible valeur impliquera une grande sensibilité au bruit, tandis qu'une grande valeur lissera trop la densité estimée, au risque de perdre des modes. D'autre part, dans des régions de forte densité, une forte valeur de h fera perdre la structure, tandis que la même valeur dans des régions à faible densité sera inutile. Idéalement, il faudrait donc une valeur de h dépendant de l'espace des données. La méthode décrite dans le paragraphe suivant permet de s'affranchir de cette limitation.

Les k plus proches voisins À l'inverse de la méthode précédente, on fixe ici k dans (3.25). On considère donc une hypersphère centrée en $\mathbf{x}$ sur laquelle on va estimer la densité $P(\mathbf{x})$ en augmentant le volume de l'hypersphère jusqu'à ce qu'elle contienne exactement k points. Ces k points constituent ainsi les k plus proches voisins (k-ppv) de $\mathbf{x}$ donnant le nom de cette méthode. Ici encore, la fixation du paramètre k détermine les performances de la méthode. Un nombre k faible produit des frontières de décision souvent non linéaires, et donc un classifieur peu généralisable, tandis que pour k grand, la faculté de discrimination se réduit, produisant un classifieur potentiellement moins performant.

Pour la méthode des noyaux de Parzen et des k-ppv, on pourra en pratique déterminer V et k par validation croisée (section 3.1.1.5).

On notera enfin qu'une troisième manière de faire consiste à transformer l'espace d'attributs de manière à utiliser des méthodes paramétriques : c'est typiquement le cas de l'analyse discriminante linéaire (ou Fisher).

3.2.4 Modèle de distance

Dans ce type d'approche, un vecteur $\mathbf{v}_i$ est considéré comme le représentant de la classe ω_i, en d'autres termes, l'ensemble des points appartenant à la classe ω_i. Nous utilisons ici le terme *prototype*, mais il faut savoir que l'on retrouvera la même signification pour des termes comme *centroïdes*, *vector quantizer*, *codevector*, *template*, etc ... Une méthode de classification consiste donc à calculer, pour une forme $\mathbf{x}$, la distance à l'ensemble des prototypes, et de l'associer à la classe du plus proche prototype :

$$\mathbf{x} \mapsto \omega_i \text{ tel que } i = argmin_{j=1,\cdots,c} d(\mathbf{x},\mathbf{v}_j)$$

Les performances de ce classifieur seront liées à la définition de la fonction de distance d. Une forme générale de distance est donnée par la métrique de Minkowski :

$$d(\mathbf{x},\mathbf{v}_i) = \Big(\sum_{j=1}^{p} |x_j - v_j|^r \Big)^{1/r} \tag{3.27}$$

où l'on retrouve la distance euclidienne pour $r = 2$, et la distance de Manhattan (City Block) pour $r = 1$. Évidemment, si l'on utilise la distance euclidienne, les classes recherchées sont de forme hyper-sphèrique. On pourra se référer au Chapitre 2 pour de plus amples définitions sur les distances (ou métriques). Une autre distance, permettant, outre d'utiliser le vecteur prototype $\mathbf{v}_i$, de modéliser la forme de la classe grâce à sa matrice de covariance Σ_i, est la distance de Mahalanobis :

$$d^2(\mathbf{x},\mathbf{v}_i) = (\mathbf{x} - \mathbf{v}_i)^T \Sigma_i^{-1} (\mathbf{x} - \mathbf{v}_i) \tag{3.28}$$

Nous verrons dans le Chapitre 4 comment transformer ces distances de manière à obtenir des degrés d'appartenance, ou de typicalité, plus aisés à manipuler pour des opérateurs d'agrégation.

3.2.5 Systèmes d'inférence floue

Les Systèmes d'Inférence Floue (SIF) sont aujourd'hui largement utilisés dans le domaine de l'aide à la décision et de la classification [Kuncheva, 2000]. Ils sont fondés sur des règles *si-alors* et un mécanisme d'inférence qui permettent de modéliser un raisonnement expert sur un problème donné. Un classifieur si-alors flou pour un objet $\mathbf{x} = (x_1 \;\; x_2 \;\; ... \;\; x_p)^T$ est un SIF constitué de règles floues de la forme:

$$\text{Règle r : } si\ A_r^1\ ET\ A_r^2\ ET \cdots ET\ A_r^p\ alors\ B_r$$

où A_r^j ($r = 1,m$ et $j = 1,p$) est un ensemble flou ayant pour fonction d'appartenance $a_r^j : \mathbb{R} \to [0,1]$ définie pour tout attribut x_j de l'objet $\mathbf{x}$ à classer, et B_r une conclusion ou sortie. On pourra par exemple prendre pour la fonction d'appartenance a_r^j la fonction modélisant x_j *est à peu près* k par

$$a_r(x_j) = \exp(-(x_j - k)^2/2) \tag{3.29}$$

Parmi les modèles les plus populaires, citons le modèle de Mamdani et celui de Takagi-Sugeno-Kang [Takagi and Sugeno, 1985], dont la différence principale est la nature de la

conclusion (ensemble flou pour le premier, sortie nette pour le second). Un système TSK se caractérise alors par :

1. un ensemble de m règles,
2. un connecteur $\mathcal{A}$ dont la sortie évalue la satisfaction τ_r de la règle r:

$$\tau_r(\mathbf{x}) = \mathcal{A}(a_r^1(x_1), \cdots, a_r^p(x_p)),$$

3. un mécanisme de défuzzification permettant de retourner une valeur réelle à partir de l'ensemble flou de sortie.

Par exemple, en choisissant comme connecteur l'opérateur produit, et la méthode de defuzzification *COA* (*Center of Area*), on obtient un classifieur de type TSK2 pour l'objet $\mathbf{x} \in \mathbb{R}^p$:

$$\mathbf{y}(\mathbf{x}) = \frac{\sum_{r=1}^{m} B_r \prod_{j=1}^{p} a_r^j(x_j)}{\sum_{r=1}^{m} \prod_{j=1}^{p} a_r^j(x_j)} \tag{3.30}$$

où $\mathbf{y}(\mathbf{x})$ est la sortie du classifieur. Il est intéressant de noter que selon le type de fonctions d'appartenance et le type de modèle, le classifieur ainsi modélisé peut être équivalent aux méthodes non paramétriques évoquées dans la section précédente : les noyaux de parzen et les k-ppv, voir [Kuncheva, 2000] pour les preuves et [Le Capitaine and Frélicot, 2008c] pour des exemples.

3.2.6 Intégrales floues

Nous avons montré dans le cadre de la classification basée sur la théorie bayésienne que les densités de probabilités de chaque attribut étaient combinées par une opération de produit, sous l'hypothèse que ceux-ci étaient indépendants (Eq. 3.14). L'approche Fuzzy Pattern Matching [Dubois et al., 1988] peut être vue comme l'alternative possibiliste de l'approche bayésienne pour la classification [Grabisch et al., 1997]. L'approche par intégrale floue [Tahani and Keller, 1990; Grabisch, 2000b] consiste à utiliser comme opérateur d'agrégation des intégrales floues. Dans un premier temps, on évalue le degré d'appartenance $u_i(x_j)$ de l'attribut j à la classe ω_i grâce à une méthode (paramétrique ou non) standard d'estimation de densité. L'ensemble des degrés d'appartenance correspondant aux attributs d'une forme $\mathbf{x}$ sont ensuite agrégés par le biais d'une intégrale floue pour obtenir le degré d'appartenance de $\mathbf{x}$ à ω_i, pour $i = 1, \cdots, c$:

$$\Phi_i(\mathbf{x}) = \mathcal{C}_\mu(u_i(x_1), \cdots, u_i(x_p)) \tag{3.31}$$

où l'on peut remplacer $\mathcal{C}$ par $\mathcal{S}$, et où μ est une mesure floue, voir section 1.4. On assigne enfin à $\mathbf{x}$ la classe pour laquelle $\Phi_i(\mathbf{x})$ est maximum. Comme nous l'avions évoqué dans le chapitre précédent, fixer les valeurs des mesures floues est fastidieux, mais permet de modéliser les interactions entre attributs, ce qui n'est pas le cas dans l'approche probabiliste. Afin de déterminer ces valeurs, plusieurs algorithmes ont été proposés : critère quadratique, quadratique généralisé, et enfin une heuristique sur les moindres carrés qui est une méthode sous-optimale sans grande perte de performances, voir [Grabisch, 1995]. Une autre manière consiste, à partir des données de l'ensemble d'apprentissage, à déterminer à quel point chaque attribut sépare les classes, et d'associer le poids correspondant à la mesure floue [Tahani and Keller, 1990].

3.3 Classification non supervisée

Jusqu'à présent, nous avons considéré que chaque forme qui est présentée au système possède une étiquette, et que l'on connaissait ainsi sa classe, ce qui permettait de procéder à l'apprentissage. Dans le cas de la classification non supervisée, on ne dispose pas de ces informations, ce qui rend plus complexe le processus de classification.

3.3.1 Modèles de mélange

Le principe des modèles de mélange consiste à décomposer la densité $P(\mathbf{x})$ en une somme de c composantes correspondant aux c classes, paramétrées par Θ_i, $i = 1, \cdots, c$. Chacune de ces composantes se voit affectée d'un coefficient π_i représentant la probabilité a priori des différentes classes :

$$P(\mathbf{x}|\Theta) = \sum_{i=1}^{c} \pi_i P(\mathbf{x}|\Theta_i). \tag{3.32}$$

Ces coefficients respectent les contraintes suivantes :

$$\pi_i \in]0, 1[\text{ et } \sum_{i=1}^{c} \pi_i = 1 \tag{3.33}$$

L'ensemble des paramètres à estimer Θ est donc composé des coefficients de mélange et des paramètres des distributions propres à chaque classe. (Ici encore, par souci de simplicité, nous considérerons le cas où les distributions sont gaussiennes.) Dans le cas où les distributions sont normales, les classes sont définies par une moyenne et une matrice de covariance, donnant

$$P(\mathbf{x}|\Theta_i) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\mathbf{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{v}_i)^T \mathbf{\Sigma_i}^{-1} (\mathbf{x} - \mathbf{v}_i) \right) \tag{3.34}$$

Afin de trouver ces paramètres, on cherchera à maximiser la vraisemblance.À la différence du cas supervisé, nous ne connaissons pas les classes, nous sommes donc dans une situation ou les données sont incomplètes. L'algorithme *Expectation-Maximization* (EM) est utilisé dans cette situation. Une variable aléatoire Z correspondant aux données manquantes est ajoutée au modèle. Les réalisations $\mathbf{z}_k = \{z_{1k}, \cdots, z_{ck}\}$ de Z sont en fait les vecteurs d'appartenance de $\mathbf{x}$ aux classes :

$$z_{ik} = \begin{cases} 1 & \text{si } \mathbf{x}_k \text{ appartient à } \omega_i \\ 0 & \text{sinon} \end{cases} \tag{3.35}$$

Ainsi lorsque nous maximisions $\mathcal{L}(X|\mathbf{v},\mathbf{\Sigma})$ par la technique du maximum de vraisemblance, nous cherchons maintenant à maximiser $\mathcal{L}(X,Z|\mathbf{v},\mathbf{\Sigma})$. L'approche EM consiste, de manière alternative, à calculer l'espérance conditionnelle de la vraisemblance complète (E-step)

$$Q(\Theta|\hat{\Theta}^{(t)}) = \mathbb{E}[\mathcal{L}(\Theta)|(X,\hat{\Theta}^{(t)})] \tag{3.36}$$

puis de mettre à jour les paramètres des modèles (M-step)

$$\hat{\Theta}^{(t+1)} = \text{argmax}_{\Theta} Q(\Theta|\hat{\Theta}^{(t)}) \tag{3.37}$$

jusqu'à ce que $||Q(\Theta|\hat{\Theta}^{(t)}) - Q(\Theta|\hat{\Theta}^{(t-1)})||$ soit inférieur à un seuil ε.

3.3.2 Partitions

Dans cette section, nous rappelons brièvement quelques principes généraux des méthodes de classification par partitionnement.

3.3.2.1 Voronoï

La région $\mathcal{R}_i$ de Voronoï du prototype $\mathbf{v}_i$ est l'ensemble des vecteurs de $\mathbb{R}^p$ pour lesquels $\mathbf{v}_i$ est le vecteur le plus proche :

$$\mathcal{R}_i = \{\mathbf{x} \in X : i = \text{argmin}_j ||\mathbf{x} - \mathbf{v}_j||^2\} \tag{3.38}$$

Ces régions forment des cellules, et l'ensemble constitue un diagramme de Voronoï. Chaque point contenu dans une cellule, c'est à dire une région $\mathcal{R}_i$, appartient ainsi à la classe du prototype correspondant, voir Fig. 3.3. Cette notion de partition de l'espace en cellules est

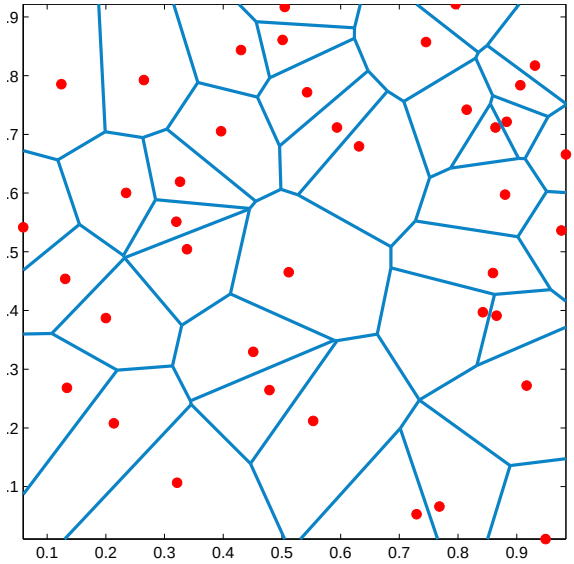

Fig. 3.3: Diagramme de Voronoï - Les prototypes sont représentés par des points rouge, et les frontières par des traits bleus

primordiale, puisque, selon que l'on soit en classification supervisée ou non supervisée, on aboutira aux méthodes des k plus proches voisins et des c-moyennes, respectivement.

3.3.2.2 C-moyennes

Lloyd propose dans [Lloyd, 1982] un algorithme simple permettant de construire un diagramme de Voronoï à partir des données X. Celui-ci est plus connu sous le nom de *K-means*, ou HCM, ou en français, et en gardant la notation c pour le nombre de classes, *C-moyennes*. Brièvement, cet algorithme, à partir d'une initialisation des c vecteurs prototypes, cherche à minimiser la distance intra-classes

$$J = \sum_{i=1}^{c} \sum_{\mathbf{x} \in \omega_i} ||\mathbf{x} - \mathbf{v}_i||^2 \tag{3.39}$$

À chaque itération, les prototypes $\mathbf{v}_i$ sont mis à jour selon

$$\mathbf{v}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \omega_i} \mathbf{x} \tag{3.40}$$

jusqu'à convergence des prototypes. En fin d'algorithme, on obtient donc un ensemble de prototypes, et un diagramme de Voronoï associé. Un problème de cette approche est sa sensibilité à l'initialisation des prototypes, c'est pourquoi il est usuel d'initialiser de manière aléatoire les prototypes, et de relancer plusieurs fois l'algorithme.

3.3.2.3 C-moyennes floues

Dans l'algorithme précédent, les éléments appartenaient ou non à des ensembles, et ce de manière stricte. Grâce à l'approche floue, cette appartenance est dorénavant une valeur comprise entre 0 et 1, et permet ainsi de modéliser l'appartenance de $\mathbf{x}$ à plusieurs ensembles. L'algorithme des C-moyennes floues (FCM, [Bezdek, 1981]) est l'extension immédiate des C-moyennes. On cherche à minimiser la fonctionnelle suivante

$$J_m(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m ||\mathbf{x}_k - \mathbf{v}_i||^2 \tag{3.41}$$

où les degrés u_{ik} d'appartenance de $\mathbf{x}_k$ au cluster i sont calculés sous les contraintes

$$\sum_{i=1}^{c} u_{ik} = 1 \tag{3.42}$$

et

$$0 < \sum_{k=1}^{n} u_{ik} < n (\forall i = 1,c) \tag{3.43}$$

pour tout $\mathbf{x}_k$ dans X, et sont éléments de la matrice de c-partition floue $U = [\mathbf{u}_1,...,\mathbf{u}_c]$ de taille $(c \times n)$. Le paramètre de fuzzification $m > 1$ permet de rendre la partition obtenue plus ou moins floue. Plus m est grand, plus les frontières sont douces, plus m est petit, plus la partition obtenue sera stricte (c'est à dire qu'elle contiendra uniquement des 0 et des 1). La minimisation de (3.41) s'obtient par itérations successives de (U,V) :

$$u_{ik} = 1 \Big/ \sum_{j=1}^{c} \left(\frac{||\mathbf{x}_k - \mathbf{v}_i||}{||\mathbf{x}_k - \mathbf{v}_j||} \right)^{2/(m-1)} \tag{3.44}$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{n} u_{ik}^m} \tag{3.45}$$

L'algorithme continue jusqu'à ce que la valeur de la fonctionnelle $J_m(U,V)$ converge. Comme pour l'algorithme précédent, il faut aussi choisir le nombre de clusters, ainsi qu'un paramètre supplémentaire, m.

Nous ne développons pas plus l'étude de cet algorithme dans cette partie, dans la mesure ou celle-ci sera fournie dans un autre chapitre, où nous nous concentrerons sur le choix de la valeur de c pour le clustering flou.

3.3.2.4 C-moyennes possibilistes

En tant que nouvelle modification des C-moyennes, l'approche possibiliste relâche la contrainte (3.42). Ainsi, on pourra trouver une forme $\mathbf{x}$ ayant un faible degré d'appartenance à l'ensemble des classes, comme une forme possédant un fort degré d'appartenance à plusieurs classes. Ces deux situations correspondront aux points atypiques et aux classes se chevauchant, respectivement. Dans ce contexte, nous appellerons les degrés d'appartenance des *degrés de typicalité*. La première formulation des C-moyennes possibilistes (PCM, [Krishnapuram and Keller, 1993]) inclut un terme de pénalité évitant les solutions triviales, en introduisant un biais de 1 pour tous les degrés.

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m ||\mathbf{x}_k - \mathbf{v}_i||^2 + \sum_{i=1}^{c}\eta_i \sum_{k=1}^{n}(1-u_{ik})^m \tag{3.46}$$

Dans la seconde approche [Krishnapuram and Keller, 1996], le terme de pénalité est l'entropie des groupes de données, on voudra ainsi minimiser le désordre. La fonctionnelle s'écrit

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik} ||\mathbf{x}_k - \mathbf{v}_i||^2 + \sum_{i=1}^{c}\eta_i \sum_{k=1}^{n}(u_{ik}\log u_{ik} - u_{ik})^m \tag{3.47}$$

où le paramètre η_i est un compromis lié à la taille des groupes. Les auteurs suggèrent de l'estimer de la manière suivante

$$\eta_i = \gamma \frac{\sum_{k=1}^{n} u_{ik}^m ||\mathbf{x}_k - \mathbf{v}_i||^2}{\sum_{k=1}^{n} u_{ik}^m} \tag{3.48}$$

lors de chaque itération, ou de manière fixe, ce qui est plus en adéquation avec le calcul lors de la minimisation, où il est considéré comme tel. Un nouveau paramètre γ est introduit dans ce calcul, mais celui-ci est généralement fixé à 1. Selon la fonctionnelle considérée, la mise à jour des degrés d'appartenance se fait par

$$u_{ik} = 1 \Big/ 1 + \left(\frac{||\mathbf{x}_k - \mathbf{v}_i||^2}{\eta_i} \right)^{\frac{1}{m-1}} \tag{3.49}$$

$$u_{ik} = \exp\left(- \frac{||\mathbf{x}_k - \mathbf{v}_i||^2}{\eta_i} \right) \tag{3.50}$$

et celle des centres en utilisant (3.45), et en fixant $m = 1$ pour la deuxième approche. L'avantage de cette méthode réside en plusieurs points. Le degré d'appartenance d'un point à une classe ne dépend plus de celui aux autres classes, on peut en fait le voir comme une distance particulière (induisant des difficultés d'interprétation des degrés d'appartenance obtenus). Cette indépendance est particulièrement intéressante dans le cas de données bruitées ou se chevauchant car ce sont des situations que l'on pourra reconnaître en étudiant les degrés d'appartenance. Un deuxième point, que certains auteurs voient comme un inconvénient, est qu'à la fin de l'algorithme, deux clusters peuvent être identiques. Dans le cas où le nombre de clusters n'est pas connu, il peut être pratique de fixer celui-ci à une valeur assez élevé, et laisser l'algorithme PCM trouver un nombre *correct* de clusters, quitte à fusionner les clusters par la suite. On remarquera que ce nombre de clusters *correct* est obtenu de manière indirecte par une mesure d'entropie dans (3.47), qui est aussi une mesure de validité de partition, voir Chapitre 5.

Les trois méthodes HCM, FCM et PCM font parties de la famille des C-moyennes, et chacune d'elles impose des contraintes différentes sur les degrés d'appartenance. Ainsi les u_{ik} appartiennent aux ensembles suivants

– HCM :

$$\mathbf{u}_k \in \mathcal{L}_{hc} = \{\mathbf{u}_k \in [0,1]^c : \sum_{i=1}^{c} u_{ik} = 1, u_{ik} \in \{0,1\}\} \tag{3.51}$$

– FCM :

$$\mathbf{u}_k \in \mathcal{L}_{fc} = \{\mathbf{u}_k \in [0,1]^c : \sum_{i=1}^{c} u_{ik} = 1\} \tag{3.52}$$

– PCM :

$$\mathbf{u}_k \in \mathcal{L}_{pc} = \{\mathbf{u}_k \in [0,1]^c\} \tag{3.53}$$

Ainsi défini, on remarque assez rapidement que

$$HCM \subset FCM \subset PCM$$

3.3.2.5 Variantes des C-moyennes

Les trois algorithmes des C-moyennes qui viennent d'être présentés sont basés sur une représentation par points des prototypes. D'autre part, l'utilisation de distance euclidienne ne permet d'obtenir que des clusters hyper-sphériques, ne donnant pas toujours de bons résultats. C'est ainsi que Gustafsson et Kessel [Gustafson and Kessel, 1979] proposent d'utiliser des matrices de covariances A_i estimées à partir des degrés d'appartenance u_{ik}. Cela ajoute donc un descripteur supplémentaire à chaque classe, décrite maintenant par le couple $(\mathbf{v}_i, A_i)$. La fonctionnelle à minimiser s'écrit donc

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m ||\mathbf{x}_k - \mathbf{v}_i||_{A_i}^2 \tag{3.54}$$

où $\det(A_i) = \rho_i > 0$, garantissant que A_i est définie positive. Les auteurs montrent en particulier que la minimisation de (3.54) conduit à la condition

$$A_i = \Big(\rho_i \det(C_i)\Big)^{1/p} C_i^{-1} \tag{3.55}$$

où C_i est la matrice de covariance floue usuelle :

$$C_i = \frac{\sum_{k=1}^{n} u_{ik}^m (\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T}{\sum_{k=1}^{n} u_{ik}^m} \tag{3.56}$$

Cette idée d'adapter les distances à la forme de chaque cluster a mené à l'idée d'introduire des prototypes de lignes [Gunderson et al., 1981] et de cercles [Davé, 1992]. On pourra enfin trouver une généralisation avec le modèle FCQS (Fuzzy C-Quadric Shell) [Krishnapuram et al., 1995]. Le lecteur pourra consulter [Bezdek et al., 1999b] pour plus de détails. À noter que l'algorithme des C-moyennes floues peut être adapté à la décomposition de mélanges suivant le principe de maximum de vraisemblance précédemment introduit, produisant l'algorithme FMLE [Gath and Geva, 1989].

À la fin de l'algorithme de classification, et quelle que soit la métrique et les prototypes employés, les méthodes des C-moyennes ont en commun de produire une **matrice de partition** (stricte, floue, possibiliste). Afin de prendre une décision, on choisit de la

transformer en une partition stricte. Cette opération s'effectue généralement en choisissant pour un élément $\mathbf{x}_k$, la classe ω_i pour laquelle le degré d'appartenance u_{ik} est le plus fort, de manière analogue à la règle MAP utilisée dans le cadre probabiliste et supervisé. Dans certains cas, la supériorité d'un degré d'appartenance par rapport aux autres n'est pas évidente; c'est typiquement le cas lorsque les classes se chevauchent. C'est le problème que nous aborderons dans cette thèse : d'une part en classification supervisée pour le rejet en ambiguïté (section 3.4.2), et d'autre part en classification non supervisée pour la validation de partition (section 3.4.3).

3.4 Problèmes – ouverts ou non

3.4.1 Dimension, absence et bruit des données

Plusieurs types de problème apparaissent en classification. Le premier d'entre eux est la dimension des données. En classification supervisée, nous avons vu que nous estimions des fonctions de densité, et que celles-ci étaient de meilleure qualité si le nombre d'observations n était important. De la même manière, augmenter p peut faire espérer une amélioration des performances. C'est effectivement le cas si l'ensemble d'apprentissage est de taille infinie, ce qui est impossible en pratique. De manière paradoxale, augmenter le nombre de caractéristiques décrivant chaque observation ne conduit pas nécessairement à améliorer les performances, et peut même conduire au contraire. Dans cette dernière situation, on dit que l'on fait face au phénomène de Hughes, ou la malédiction de la dimension [Bellman, 1961]. Ce phénomène, que l'on retrouve dans tous les algorithmes d'apprentissage, provient du fait que le nombre de données, n, nécessaire pour apprendre un concept en dimension p croît exponentiellement en fonction de p. D'autre part, la complexité des algorithmes augmente évidemment lorsque la dimension de p augmente. De ce fait, même si la relation exacte entre la probabilité d'erreur, le nombre de variables p et la taille n de l'ensemble d'apprentissage et la complexité du classifieur n'est pas établie, il est conseillé de disposer d'un nombre d'observations dix fois supérieur au nombre de variables (3.1.1.1, [Jain et al., 2000]). Puisqu'une contrainte importante est de rassembler un ensemble d'apprentissage de grande dimension, on essaie souvent de diminuer l'espace de représentation des données (voir section 3.1.1.3), tout en conservant les performances du modèle.

Dans de nombreuses situations, certaines variables décrivant les observations peuvent être absentes. Nous sommes alors en présence de données incomplètes. Il est important de différencier la situation où la valeur manquante apparaît dans les données d'apprentissage ou de test, puisque cela mènera à des stratégies différentes. Parmi celles-ci, nous noterons les techniques d'imputation, consistant à dire que si une variable est manquante, alors on peut l'estimer par les données que l'on possède déjà (soit par estimation de la distribution conditionnelle de la valeur, soit en prédisant la valeur manquante). Une alternative à l'approche par imputation consiste à n'utiliser qu'un sous-ensemble des variables descriptives pour lesquels il ne manque pas d'information. Le lecteur pourra se reporter à [Saar-Tsechansky and Provost, 2007] pour un panorama récent complet, ainsi que diverses comparaisons.

Enfin, il est possible de trouver des données non conformes au comportement espéré. Ces données sont généralement qualifiées d'exceptions (ou d'*outliers*). Parmi ces données, on distinguera les valeurs extrêmes dans une distribution des données réellement aberrantes.

Bien que les distinguer de manière théorique est immédiat, en pratique, il est très difficile de le faire. Généralement, les travaux sur la détection de points aberrants s'appuient sur la distance (ou densité) du point aux classes existantes. Le problème principal consiste ainsi à définir une distance (densité) limite au delà de laquelle un point sera considéré comme aberrant. Cela soulève aussi la question de savoir à partir de quel moment un regroupement de point est considéré comme une classe, ou comme un groupe de points aberrants. On pourra consulter un état de l'art récent pour de plus amples détails [Chandola et al., 2007].

3.4.2 Option de rejet

Ce chapitre se voulant une brève introduction à la reconnaissance de formes et aux différents concepts que nous allons utiliser par la suite, nous présentons ici rapidement l'option de rejet, pour la décrire plus amplement dans le Chapitre 4.

Le classement d'un objet peut être problématique si la probabilité (ou risque) d'erreur associée est trop élevée. Typiquement, dans le diagnostic médical automatique, un faux-négatif est bien plus couteux qu'un faux positif. Dans ce genre de situation, il est préférable de refuser le classement à l'objet, qui sera appelé un point rejeté. D'autre part, le classement exclusif d'un objet repose sur deux hypothèses rarement vérifiées en pratique :

1. les classes ne se chevauchent pas dans l'espace d'attributs $\mathbb{R}^p$,
2. une description exhaustive des classes est disponible (hypothèse du monde fermé).

C'est ainsi que les options de rejet ont été proposées, afin de pallier à ces problèmes. Au sein des options de rejet, on trouve deux types de rejet [Dubuisson and Masson, 1993]. Le premier d'entre eux, le **rejet d'ambiguïté**, consiste à donner la possibilité à un objet d'être associé à plusieurs classes (rejet sélectif), voir toutes (rejet total), voir Fig. 3.4. Ce genre de décision est nécessaire lorsque l'objet se situe au niveau des frontières de décision des classes. Le deuxième type de rejet est le **rejet de distance**, qui concerne des points loin de la quasi-totalité des points observés. On voudra à ce moment-là ne pas associer de classe à l'objet, qui sera considéré comme du bruit, voir Fig. 3.5. Plus de détails, ainsi que les différentes stratégies employées pour mettre en place le rejet, seront donnés dans le Chapitre 4.

3.4.3 Validation de partitions

À l'instar de la section précédente, nous présentons brièvement le problème de la validation de partitions, et nous le détaillerons dans le Chapitre 5.

L'apport des approches floues dans la classification non supervisée (section 3.41) permet d'obtenir des méthodes moins sensibles aux minimum locaux, grâce à l'itération floue à chaque étape. Pourtant, comme les méthodes strictes, les approches floues présentent certains défauts. Elles possèdent aussi le défaut, outre leur non robustesse au bruit, de devoir fixer le nombre c de classes du problème. Lorsque celui-ci est connu, il n'y a pas de difficulté, mais dans le cas contraire, la partition obtenue pour un nombre quelconque de classes doit être validée. On voit dans les Fig. 3.6-3.7 que ce choix peut avoir une grande incidence. Trouver la valeur optimale de c est encore aujourd'hui un problème ouvert en analyse de données, et est habituellement appelé Validation de partitions (Cluster Validity) dans la littérature. Le principe général est de lancer l'algorithme pour différentes valeurs de c, et de comparer les partitions obtenues. Cette comparaison permettra de distinguer une partition

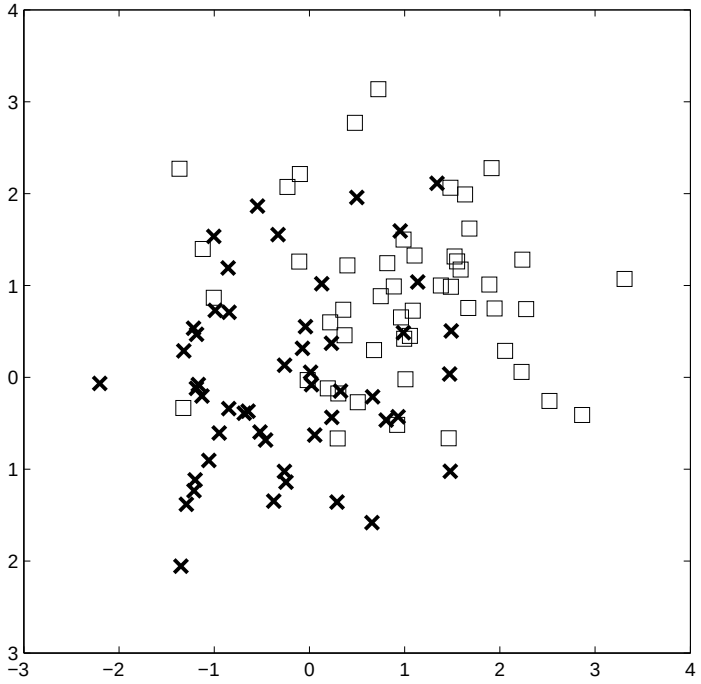

FIG. 3.4: Deux classes (□ et ×) se chevauchant dans $\mathbb{R}^2$.

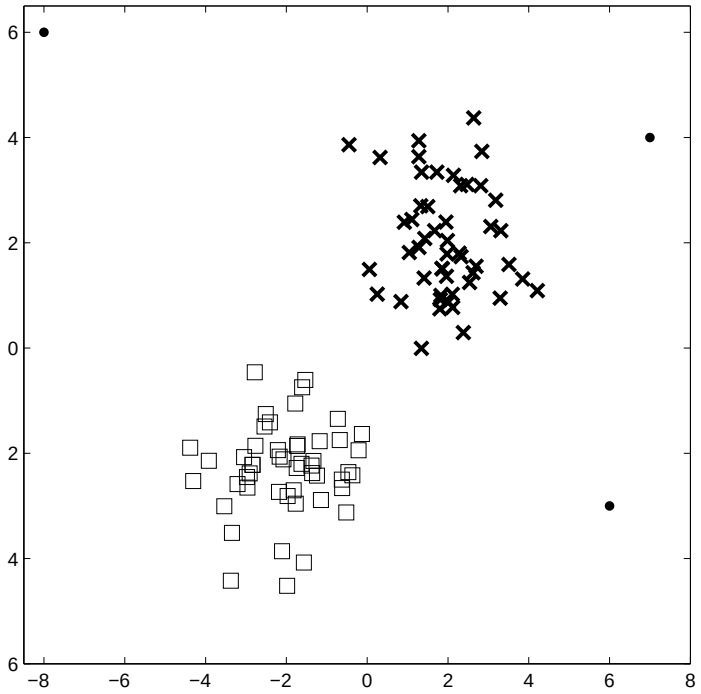

FIG. 3.5: Présence de points atypiques (•), loin des deux classes bien définies (□ et ×).

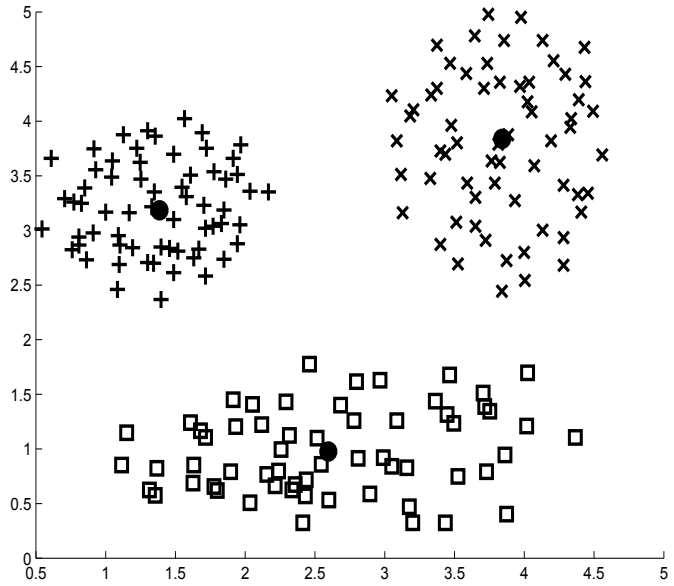

FIG. 3.6: Bonne ($c = 3$) partition stricte obtenue à partir de FCM, et les centres associés (•).

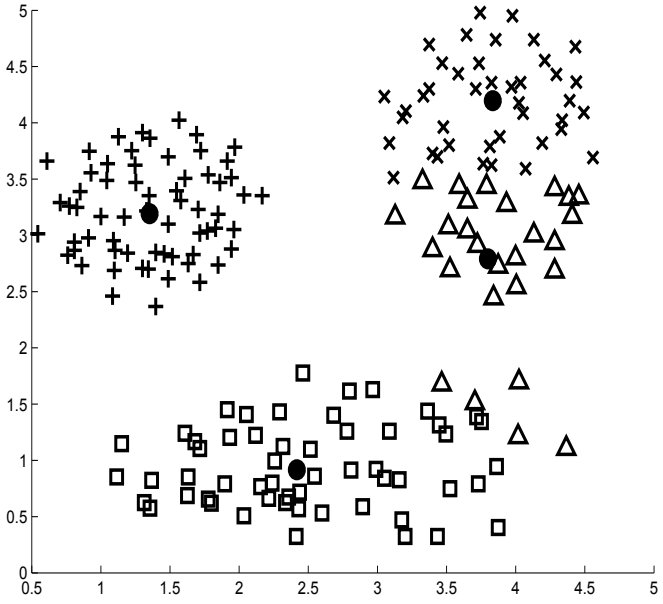

FIG. 3.7: Mauvaise ($c = 4$) partition stricte obtenue à partir de FCM, et les centres associés (•).

meilleure (au sens du critère de mesure) que les autres, déterminant ainsi le nombre optimal c de classes, et la classification associée.

3.5 Conclusion

Dans ce chapitre, après avoir introduit les concepts et la démarche générale de la reconnaissance de formes, nous avons proposé un bref aperçu des différentes techniques employées dans le domaine. Cet aperçu ne se veut bien sûr absolument pas exhaustif, tant la littérature à ce sujet est abondante. Tout au long de ce chapitre, des références ont été données, et nous suggérons aux lecteurs désireux d'en connaître plus de se référer à celles-ci. Nous avons par conséquent donné les bases sur lesquelles nous allons travailler dans le Chapitre 4 et le Chapitre 5, dédiés à l'application pratique des opérateurs proposés dans le Chapitre 2. En particulier, nous nous attaquerons aux deux problèmes précédemment évoqués, le rejet en classification supervisée, et la validation de partitions en classification non supervisée.

Chapitre 4

Options de rejet en classification supervisée

Rejetez le noir, et ce mélange de blanc et de noir qu'on nomme le gris. Rien n'est noir, rien n'est gris. Ce qui semble gris est un composé de nuances claires qu'un œil exercé devine.

— Paul Gauguin
(date inconnue)

Résumé : *Dans ce chapitre, une première application des mesures proposées au* Chapitre 2 *est proposée. Il s'agit du rejet en classification supervisée, où l'on cherche dans l'espace d'attribut des zones de faible densité (rejet en distance), ou des zones de forte densité impliquant plusieurs classes (rejet en ambiguïté). Nous proposons de nouvelles règles de sélection des classes d'affectation, ainsi qu'une nouvelle méthodologie d'évaluation des résultats obtenus par ces règles.*

4.1 Introduction, motivations et formalisme

Lors de processus classiques de classification, l'objectif est d'associer une seule classe à une observation. Pourtant, plusieurs problèmes apparaissent. Deux d'entre eux sont d'une importance majeure et peuvent dégrader de manière significative les performances d'un classifieur. D'une part, les classes peuvent présenter un chevauchement non négligeable dans l'espace d'attributs, le caractère exclusif de la décision est donc remis en cause. D'autre part, en classification supervisée, l'hypothèse d'un monde fermé est souvent adoptée. En d'autres termes, on suppose que l'on a une description complète des classes, ce qui est rarement le cas en pratique. Le principe général de l'option de rejet consiste à introduire la possibilité au classifieur de ne pas classer une observation s'il est jugé qu'il y a un doute quant à la décision à prendre. L'intérêt de ce rejet réside dans le fait qu'il permet de réduire le taux d'erreur de classification. En particulier, pour des applications où une erreur de classification est extrêmement critique (par exemple en diagnostic médical, en contrôle de machines) il est opportun de préférer un faux positif à un faux négatif[1]. Même si elle ne permet pas

1. Un faux positif signifie que l'on se trompe en estimant que le test est positif, alors que le faux négatif signifie que l'on se trompe en disant que le test est négatif

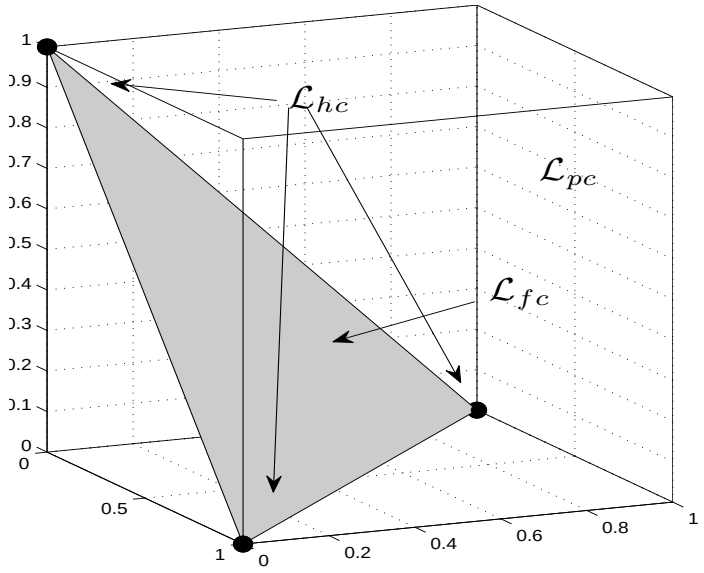

FIG. 4.1: Différents ensembles $\mathcal{L}_{\bullet c}$ de classifieurs selon leurs restrictions

d'augmenter le taux de classification correcte, l'option de rejet permet de diminuer les risques tant qu'un classifieur idéal ne sera pas proposé.

On distingue deux types de rejet [Dubuisson and Masson, 1993]. Le premier, appelé *rejet d'ambiguïté*, consiste à donner la possibilité d'affecter un objet à plusieurs classes. On distingue le rejet total, où l'ensemble des classes sont concernées, du rejet sélectif, où un nombre inférieur ou égal à c classes sont impliquées. On notera cependant que le rejet sélectif n'exclut pas de procéder à du rejet total, tandis que l'inverse est impossible. Ce genre de situation apparaît lorsque les classes se chevauchent dans $\mathbb{R}^p$. Le second type de rejet est le *rejet en distance*, situation où l'objet à classer n'est caractéristique d'aucune classe connue à ce jour. C'est typiquement le cas de points de bruit, ou de points aberrants.

Lors du processus de classification, nous distinguons deux étapes distinctes : l'estimation des degrés d'appartenance (probabilité a posteriori, degrés possibilistes, flous), et la prise de décision en elle-même, où le classifieur associe l'objet à la classe. Jusqu'à présent, nous avions considéré que cette association, ou appartenance, était exclusive, de telle manière que l'on pouvait décrire un processus de classification par les deux fonctions d'étiquetage ($\mathcal{L}$abeling) et d'assignation ($\mathcal{H}$ardening) :

- $\mathcal{L} : \mathbf{x} \mapsto \mathbf{u}(\mathbf{x}) = {}^t(u_1(\mathbf{x}), \ldots, u_c(\mathbf{x})) \in \mathcal{L}_{\bullet c}$
- $\mathcal{H} : \mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x}) = {}^t(h_1(\mathbf{x}), \ldots, h_c(\mathbf{x})) \in \mathcal{L}_{hc}$.

où, selon les classifieurs, $\mathcal{L}_{\bullet c}$ représente l'un des espaces suivants, voir aussi FIG. 4.1.

- $\mathcal{L}_{pc} = [0,1]^c$ pour des degrés de typicalité,
- $\mathcal{L}_{fc} = \left\{ \mathbf{u}(\mathbf{x}) \in \mathcal{L}_{pc} : \sum_{i=1}^{c} u_i(\mathbf{x}) = 1 \right\}$ pour des probabilités a posteriori ou des degrés d'appartenance,
- $\mathcal{L}_{hc} = \left\{ \mathbf{u}(\mathbf{x}) \in \mathcal{L}_{fc} : u_i(\mathbf{x}) \in \{0,1\} \right\}$ pour des étiquettes exclusives.

La classification exclusive consiste à déterminer $\mathcal{H}$ de manière à ce que $\mathbf{h}(\mathbf{x}) \in \mathcal{L}_{hc}$. Introduire une option de rejet consiste à modifier $\mathcal{H}$, et en particulier l'espace $\mathcal{L}_{hc}$, de la manière

suivante :

$$\mathcal{L}_{hc}^{c} = \left\{ \mathbf{h}(\mathbf{x}) \in \mathcal{L}_{pc} : h_i(\mathbf{x}) \in \{0,1\} \right\} \tag{4.1}$$

Cet espace est l'ensemble des sommets de l'hypercube unité, et est de ce fait composé de 2^c valeurs possibles, représentant l'ensemble des sous-ensembles de classes possibles parmi les c classes. Dans ce cadre, les trois (quatre si l'on différencie rejet sélectif et rejet total) options de classement seront retranscrites par des situations particulières de $\mathbf{h}(\mathbf{x}) \in \mathcal{L}_{hc}^{c}$

- si $\sum_{i=1}^{c} h_i(\mathbf{x}) = 0$, la décision de rejet en distance est prise,
- si $1 < \sum_{i=1}^{c} h_i(\mathbf{x}) < c$, la décision de rejet en ambiguïté sélectif est prise,
- si $\sum_{i=1}^{c} h_i(\mathbf{x}) = c$, la décision de rejet total est prise,
- si $\sum_{i=1}^{c} h_i(\mathbf{x}) = 1$, la décision de classification exclusive est prise.

L'option de rejet se concentre ainsi exclusivement sur l'étape $\mathcal{H}$, et l'étape d'estimation des degrés d'appartenance est faite indépendamment de celle-ci.

Dans le cas probabiliste, les $u_i(\mathbf{x})$ sont des probabilités a posteriori $P(\omega_i|\mathbf{x})$, qui peuvent être obtenues à partir des densités conditionnelles des classes, dont les paramètres sont estimés à partir d'un ensemble d'apprentissage, voir section 3.2. Si le classifieur est flou, on peut utiliser la fonction de mise à jour des degrés d'appartenance (3.44) des Fuzzy c-means. Enfin, et nous nous concentrerons sur cette dernière situation, on peut estimer les degrés d'appartenance dits possibilistes (c'est à dire sans contrainte de somme à 1), au sens du degré de typicalité de la forme $\mathbf{x}$ par rapport à la classe ω_i, [Krishnapuram and Keller, 1993]. Puisqu'elle utilise la moyenne et la covariance et qu'elle possède un pouvoir discriminant intéressant, la distance de Mahalanobis est souvent utilisée

$$d^2(\mathbf{x},\mathbf{v}_i) = (\mathbf{x} - \mathbf{v}_i)^T \Sigma_i^{-1} (\mathbf{x} - \mathbf{v}_i) \tag{4.2}$$

Afin de ramener cette distance à un étiquetage possibiliste, plusieurs fonctions sont utilisées, parmi elles, citons les fonctions de type exponentielles et de type Cauchy :

$$u_i(\mathbf{x}) = \exp\left(-\alpha_i d^2(\mathbf{x},\mathbf{v}_i) \right) \tag{4.3}$$

et

$$u_i(\mathbf{x}) = \frac{\alpha_i}{\alpha_i + d^2(\mathbf{x},\mathbf{v}_i)} \tag{4.4}$$

où les α_i sont des paramètres à fixer dans $[1, +\infty[$. On pourra en particulier faire le lien entre ces fonctions et les fonctions noyaux utilisées dans le Chapitre 2. Dans la suite de ce chapitre, nous utiliserons, sauf si le contraire est explicitement mentionné, l'équation (4.4) où l'on fixe les α_i à 1, pour déterminer les degrés d'appartenance. Notons que pour obtenir des degrés d'appartenance possibilistes, une alternative consiste à utiliser l'étiquetage possibiliste de l'algorithme PCM défini par l'équation (3.49). Le choix de (4.4) n'a que peu d'influence sur les résultats obtenus, nous avons opté pour celui-ci car des résultats empiriques montrent qu'il est un bon choix pour les fonctions d'appartenance modélisant des concepts de classe, voir [Zimmermann and Zysno, 1985].

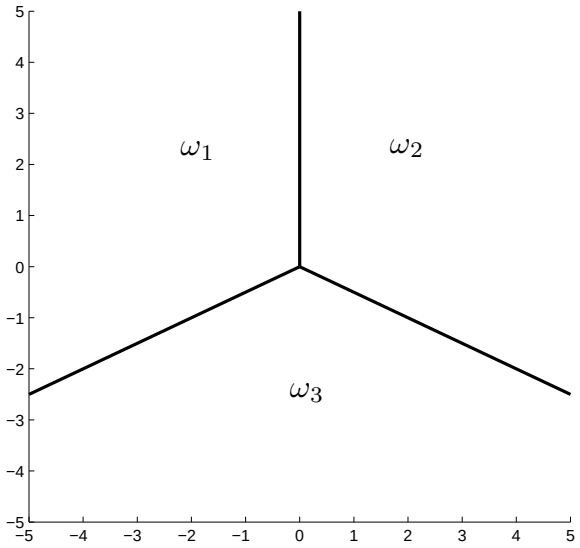

FIG. 4.2: Partitionnement de l'espace d'attributs sans rejet pour trois classes ω_1, ω_2 et ω_3

Nous nous concentrons dans cette thèse sur l'apport des degrés d'appartenance pour la mise en place du rejet, et nous utilisons pour cela des classifieurs possibilistes ou fondés sur une distance. On notera cependant que récemment, plusieurs propositions visant à introduire du rejet dans des classifieurs SVM ont été introduites, par exemple [Fumera and Roli, 2002; Grandvalet et al., 2008; Bartlett and Wegkamp, 2008]. Ces approches ne font pas partie de cette étude, mais pourront faire l'objet de futures recherches. Nous présentons dans un premier temps les différentes approches ayant trait au rejet que l'on peut trouver dans la littérature.

4.2 Stratégies standard

Le rejet d'ambiguïté a tout d'abord été introduit par Chow [Chow, 1957; 1970]. Cette première proposition se fonde sur la règle optimale de Bayes, et les valeurs u_i sont donc des probabilités a posteriori $P(\omega_i|\mathbf{x})$. Comme les valeurs u_i sont des probabilités a posteriori, cette règle affecte à la forme $\mathbf{x}$ la classe la plus probable, et est connue pour être optimale par rapport à la probabilité d'erreur si les paramètres du modèle sont exacts. La FIG. 4.2 illustre la partition de l'espace ainsi opérée. La règle de Chow consiste à rejeter la forme $\mathbf{x}$ si

$$u_{(1)}(\mathbf{x}) = \max_i u_i(\mathbf{x}) < 1 - t \tag{4.5}$$

Celle-ci minimise l'erreur de probabilité pour une probabilité de rejet donnée par un seuil $t \in [0, (c-1)/c]$. Chow montre que le taux d'erreur et le taux de rejet sont des fonctions monotones de t, le taux d'erreur est une fonction convexe du taux de rejet, et que pour des taux différentiables,

$$\frac{dE(t)}{dR(t)} = -t$$

où

$$E(t) = \int P(\mathbf{x})H(t - (1 - u_{(1)}(\mathbf{x})))(1 - u_{(1)}(\mathbf{x}))\mathrm{d}\mathbf{x}$$

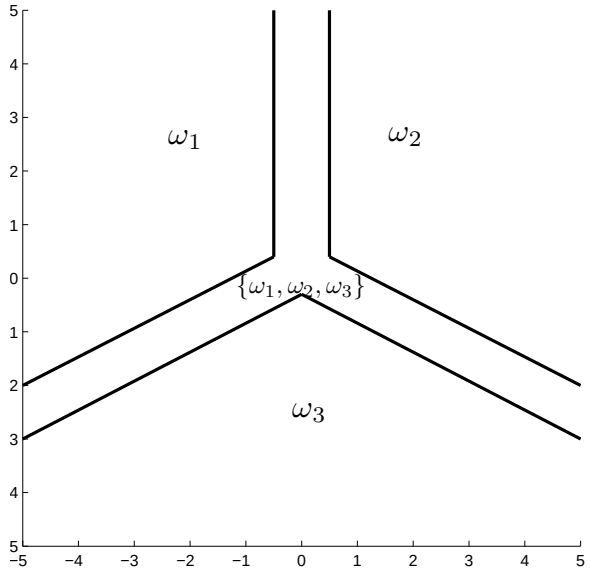

FIG. 4.3: Partitionnement de l'espace d'attributs avec rejet total.

et

$$R(t) = \int P(\mathbf{x})H((1 - u_{(1)}(\mathbf{x})) - t)\mathrm{d}\mathbf{x}$$

avec $H(\cdot)$ la fonction Heaviside.

Si $t > (c-1)/c$, il n'y a pas de rejet. Dans le cas contraire, le rejet est qualifié de *total*, puisque une seule région d'ambiguïté est définie pour l'ensemble des classes, voir FIG. 4.3. Nous verrons dans la section 4.2.3 que des variantes ont été introduites afin de permettre de sélectionner les classes pertinentes pour chaque forme. Comme précisé auparavant, la règle de Chow n'est optimale que si les probabilités a posteriori estimées sont exactes, ce qui n'est pas le cas en pratique. Dans [Fumera et al., 2000], les auteurs proposent donc d'utiliser un seuil par classe, modifiant donc la règle de Chow et proposent la règle de rejet

$$\max_{i=1,\cdots,c} u_i(\mathbf{x}) = u_j(\mathbf{x}) < 1 - t_j \tag{4.6}$$

À ce rejet en ambiguïté proposé par Chow, Dubuisson et Masson opposent le rejet en distance, [Dubuisson and Masson, 1993]. Ici, les valeurs u_i sont des probabilités a posteriori, et on utilise la densité mélange $P(\mathbf{x})$ pour la comparaison avec un seuil $\mathcal{C}_d$. La règle de Dubuisson (FIG. 4.4) consiste à rejeter la forme $\mathbf{x}$ si

$$P(\mathbf{x}) < \mathcal{C}_d \tag{4.7}$$

Fixer la valeur $\mathcal{C}_d$ est problématique, trop faible aucune forme ne sera rejetée, trop élevée, et le rejet sera presque systématique. A cela s'ajoute le fait que si les classes ont des distributions variées, l'utilisation d'un seul seuil $\mathcal{C}_d$ pour l'ensemble des classes n'est pas adaptée, poussant les auteurs de [Muzzolini et al., 1998] à proposer un seuil pour chaque classe. Cette généralisation est analogue à celle proposée pour le rejet d'ambiguïté ou plusieurs seuils étaient utilisés.

Pour ces deux extensions à l'utilisation de plusieurs seuils, un nouveau problème vient s'ajouter : fixer les c valeurs de seuil, alors qu'une seule valeur suffisait jusqu'alors. Divers

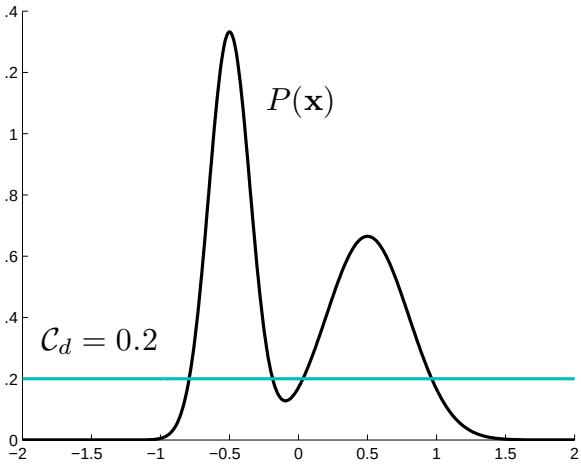

FIG. 4.4: Rejet en distance avec seuillage de la densité mélange $P(\mathbf{x})$ par $\mathcal{C}_d$. Les formes situées sous la ligne sont rejetées en distance.

algorithmes ont été proposés pour cela, nous citerons, sans être exhaustif, des méthodes fondées sur des nuées de particules [Oliveira et al., 2005], ou encore sur la maximisation de performance sous contraintes [Fumera et al., 2000]. Si les valeurs u_i ne sont pas de nature probabiliste, on pourra simplement seuiller ces valeurs, et procéder à un *rejet d'appartenance* [Frélicot, 1996] : le point sera rejeté si le plus grand de ses degrés d'appartenance est inférieur à un seuil t. De manière identique aux deux précédentes extensions, on pourra considérer un seuil d'appartenance t_j pour chaque classe ω_j.

4.2.1 Mesures de rejet d'ambiguïté et de distance

La plupart des stratégies de rejet se fondent sur l'utilisation de seuils. Il existe aussi des variantes où le rejet est vu comme un problème de classification, et une classe supplémentaire est ajoutée lors de l'apprentissage du classifieur. On trouve également des propositions où un classifieur à deux classes (acceptation et rejet) est utilisé [Landgrebe et al., 2006]. Enfin, on peut entrainer un classifieur sur des mesures d'ambiguïté et de distance, de manière à discriminer les situations. Si les mesures sont discriminantes, les classes formées sont linéairement séparables, et la classification est facilitée, voir [Pitrelli et al., 2006]. Nous étudions ici les mesures de rejet d'ambiguïté et de distance (également appelées fonctions de confiance lorsqu'elles sont complémentées) dans le cas où une stratégie fondée sur des seuils est utilisée, ce qui représente la majorité des propositions de la littérature. La règle de Chow est souvent utilisée pour les classifieurs produisant des estimations de probabilités a posteriori. Pourtant, ces estimations peuvent être incorrectes, ou le classifieur ne produit pas de probabilités a posteriori, mais plutôt des degrés d'appartenance. C'est ainsi que d'autres règles ont été proposées, principalement fondées sur de nouvelles mesures d'ambiguïté, ou de confiance par rapport au classifieur.

Mesures de rejet en ambiguïté Φ :

Dans [Chow, 1970], la première mesure d'ambiguïté est introduite. Elle correspond au risque pris lorsque la classe correspondant au maximum de la probabilité a posteriori est

choisie :

$$\Phi_{Chow}(\mathbf{u}) = 1 - u_{(1)}(\mathbf{x}) \tag{4.8}$$

C'est, et de loin, la mesure d'ambiguïté la plus utilisée. On notera que d'autres auteurs reprendront exactement la même mesure, la différence de leur proposition se situant dans la stratégie, puisqu'ils utilisent des seuils de rejet différents pour chaque classe, [Fumera et al., 2000].

Une seconde mesure, fondée sur l'idée qu'il faut affecter à $\mathbf{x}$ toutes les classes dont la probabilité a posteriori est supérieure à un seuil t est proposée par Ha, [Ha, 1997]. Si l'on restreint à la mesure d'ambiguïté en tant que telle, sans se soucier de la stratégie (voir pour cela la section 4.2.3 sur la sélection du nombre optimal de classes), alors la mesure d'ambiguïté de Ha est définie par

$$\Phi_{Ha}(\mathbf{u}) = u_{(2)}(\mathbf{x}) \tag{4.9}$$

En réalité, cette mesure d'ambiguïté devrait s'écrire $u_{(k+1)}(\mathbf{x})$, où k varie de 1 à c, mais le rejet (non sélectif, voir section 4.2.3) intervient à partir de $k = 1$, d'où la notation de l'équation (4.9).

Une autre approche, proposée initialement par Horiuchi [Horiuchi, 1998], puis utilisée dans [Ishibuchi and Nakashima, 1998; De Stefano et al., 2000], consiste à considérer les relations entre probabilités a posteriori, et en particulier leurs distance. Ils comparent ainsi les deux plus grandes valeurs de $\mathbf{u}$:

$$\Phi_{Horiuchi}(\mathbf{u}) = 1 - (u_{(1)}(\mathbf{x}) - u_{(2)}(\mathbf{x})) \tag{4.10}$$

Ici encore, la notation introduite par l'auteur est $u_{(k)}(\mathbf{x}) - u_{(k+1)}(\mathbf{x})$, mais dans la mesure où le rejet intervient à partir de $k = 1$, nous dérivons de cette formulation la mesure d'ambiguïté définie par l'équation (4.10) [2].

Dans [Frélicot and Dubuisson, 1992], les auteurs proposent le rapport de la deuxième plus grande valeur sur la plus grande valeur :

$$\Phi_{FD}(\mathbf{u}) = \frac{u_{(2)}(\mathbf{x})}{u_{(1)}(\mathbf{x})} \tag{4.11}$$

L'avantage par rapport à la proposition de Horiuchi est que tout en conservant la notion de relations entre valeurs, le rapport permet de traiter les fortes valeurs semblables d'une autre manière que les faibles valeurs similaires, correspondant au rejet en distance. Cette mesure d'ambiguïté est réintroduite de nombreuses fois dans la littérature, par exemple dans [Foggia et al., 1999; De Stefano et al., 2000; Sansone et al., 2001; Mouchere and Anquetil, 2006] sous la forme d'une fonction de confiance $\psi(\mathbf{u}) = 1 - \Phi_{FD}(\mathbf{u})$. Cette mesure est étendue à $k \in \{1, \cdots, c\}$ dans [Frélicot and Mascarilla, 2002] par

$$\Phi_{FM}(\mathbf{u}) = \frac{u_{(k+1)}(\mathbf{x})}{u_{(1)}(\mathbf{x})}, \tag{4.12}$$

et on trouvera sa forme générale définie par

$$\Phi_{LF}(\mathbf{u}) = \frac{u_{(k+1)}(\mathbf{x})}{u_{(k)}(\mathbf{x})}, \tag{4.13}$$

2. Notons que l'auteur a introduit une distance entre probabilités, c'est à dire une mesure de non-ambiguïté, ne mentionnant à aucun moment une quelconque mesure d'ambiguïté. Ce terme est adopté ici pour la cohérence du propos.

dans [Frélicot and Le Capitaine, 2009]. Récemment, dans [Mascarilla et al., 2008], les auteurs utilisent la notion de k plus grandes valeurs de $\mathbf{u}$ pour dériver une nouvelle mesure d'ambiguïté associée à l'ordre k :

$$\Phi_{k,\top}(\mathbf{u}) = \overset{k}{\perp}(\mathbf{u}) \tag{4.14}$$

où $\overset{k}{\perp}(\mathbf{u})$ est défini par l'équation (1.57). Si l'on désire faire du rejet simple, alors $\overset{2}{\perp}(\mathbf{u})$ est utilisé. Dans le cas de l'utilisation de t-normes min, cette mesure est une généralisation de la mesure de Ha. Nous proposerons une manière de procéder à du rejet sélectif grâce à cet opérateur en section 4.3. Notons que dans [Tax and Duin, 2008], les auteurs proposent l'utilisation de la règle de Chow multi-seuils ([Fumera et al., 2000]) pour le rejet, mais procèdent à une normalisation des degrés d'appartenance avec l'introduction dans la modification des seuils t_i. Cette modification prend deux formes selon que l'on utilise un classifieur probabiliste :

$$u_i(\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i) - t_i}{\frac{1}{n_i}\sum_{j=1}^{n_i} P(\mathbf{x}_j|\omega_j) - t_i} \tag{4.15}$$

ou un classifieur utilisant une distance aux prototypes :

$$u_i(\mathbf{x}) = \frac{t_i - d(\mathbf{x},\mathbf{v}_i)}{t_i - \frac{1}{n_i}\sum_{j=1}^{n_i} d(\mathbf{x}_j,\mathbf{v}_i)} \tag{4.16}$$

où les t_i sont les seuils fixés par l'utilisateur.
Enfin, dans [Frélicot and Le Capitaine, 2009] nous proposons deux nouvelles mesures d'ambiguïté. La première est une généralisation de la minimisation du risque de Chow par l'utilisation de connecteurs disjonctifs :

$$\Phi_{\overline{1},\top}(\mathbf{u}) = 1 - \perp(\mathbf{u}) \tag{4.17}$$

Ceci permet de mesurer à quel point la plus grande valeur de $\mathbf{u}$ (par complément) est forte. Cette mesure est une généralisation de la mesure Φ_{Chow}, puisque dans le cas particulier ou l'on choisit le couple de normes triangulaires $(\top,\perp)_M$, on a $\Phi_{\overline{1},\top_M}(\mathbf{u}) = 1 - u_{(1)}(\mathbf{x})$. La seconde mesure que nous proposons repose sur la notion de spécificité, en particulier sur l'opérateur ou exclusif (2.10).

$$\Phi_{\underline{1},\top}(\mathbf{u}) = 1 - \underline{\perp}(\mathbf{u}) \tag{4.18}$$

Comme $\underline{\perp}(\mathbf{u})$ est grand si la plus grande valeur est forte comparée à la seconde plus grande valeur (et donc nécessairement grande par rapport aux autres), le complément de $\underline{\perp}(\mathbf{u})$ définit une mesure d'ambiguïté.

Mesures d'acceptation (rejet en distance) Ψ :

Dans le premier article de Chow [Chow, 1957] au sujet du rejet, aucune distinction n'était faite entre le rejet en ambiguïté et le rejet en distance : la forme était simplement rejetée. Clairement, dans le cas probabiliste, la mesure Φ_{Chow} est insuffisante puisque la contrainte de somme à 1 empêche la distinction entre ambiguïté et distance, distinction introduite dans [Dubuisson and Masson, 1993]. Une mesure d'acceptation fonction de $\mathbf{u}$ est définie comme

$$\Psi_{Chow}(\mathbf{u}) = u_{(1)}(\mathbf{x}) \tag{4.19}$$

Cette solution est adoptée dans de nombreuses situations [Fumera et al., 2000; Landgrebe et al., 2006; Le Capitaine and Frélicot, 2009a]. Selon la nature des degrés d'appartenance, Ψ_{Chow} permet ou non de procéder à du rejet en distance. Dans le cas de degrés possibilistes, Ψ_{Chow} peut être utilisée, mais dans ce cas, la mesure d'ambiguïté associée Φ_{Chow} ne permettra plus de détecter les régions d'ambiguïté. Contrairement à Ψ_{Chow}, la mesure d'acceptation introduite dans [Dubuisson and Masson, 1993] n'est pas une fonction de $\mathbf{u}$, mais directement de $\mathbf{x}$:

$$\Psi_{DM}(\mathbf{x}) = P(\mathbf{x}) \tag{4.20}$$

Dans [De Stefano et al., 2000], les auteurs proposent d'introduire dans le calcul le degré d'appartenance maximum sur l'ensemble des formes pour une classe donnée

$$\Psi_{max} = 1 - \frac{u_{(1)}(\mathbf{x})}{u^{\star}_{(1)}(\mathbf{x})} \tag{4.21}$$

où $u^{\star}_{(1)}(\mathbf{x}) = \max_{k=1,\cdots,n} u_{(1)}(\mathbf{x}_k)$. Ceci permet de s'assurer de la validité de la mesure, puisque l'exemple le plus typique de chaque classe est pris comme référence.

Enfin, notons que des tentatives de proposer un seul opérateur combinant les deux types de rejet. En particulier, dans [Foggia et al., 1999], les deux mesures $1 - \Phi_{FD}$ et Ψ_{Chow} sont combinées par les opérateurs d'agrégation min, max et la moyenne arithmétique.

4.2.2 Double option de rejet : deux étapes pour une règle

Une stratégie est dite de double option de rejet si elle intègre de manière séquentielle dans une règle de classement les deux types de rejet évoqués précédemment. Plusieurs stratégies ont été proposées. Elles diffèrent principalement par l'ordre dans lequel les décisions de classement exclusif, rejet de distance et d'ambiguïté sont considérées. On distinguera ainsi les approches suivantes

- *Monde fermé*, où le nombre de classes est artificiellement incrémenté d'une classe représentant la classe de rejet de distance. Le degré d'appartenance de $\mathbf{x}$ à la $c+1$ ème classe représente le degré de non spécificité de $\mathbf{x}$ aux c autres classes, [Smyth, 1994].
- *Accepte d'abord*, où l'on regarde dans un premier temps si l'on rejette la forme en distance (en appartenance), puis si elle est acceptée, on décide de la classification exclusive ou du rejet par ambiguïté, [Dubuisson and Masson, 1993; Mascarilla and Frélicot, 2001].
- *Rejette d'abord*, où l'on décide d'abord si l'on rejette la forme que ce soit en distance ou en ambiguïté, puis, si c'est le cas, on détermine lequel des deux rejets est à opérer, [Mascarilla and Frélicot, 2001].
- *Mélange d'abord*, où l'ambiguïté de classification est en premier lieu inspectée. Si il n'y a pas d'ambiguïté, alors on décide de la classification exclusive ou du rejet en appartenance, [Semani et al., 2002].

Notons que l'utilisation de règles floues afin de faciliter la gestion conjointe des trois types de décision a également été proposée. Dans [Ishibuchi and Nakashima, 1998], les auteurs construisent un classifieur fondé sur des règles floues, et ajoutent la possibilité d'avoir une sortie avec de multiples classes. Dans [Le Capitaine and Frélicot, 2008c], nous

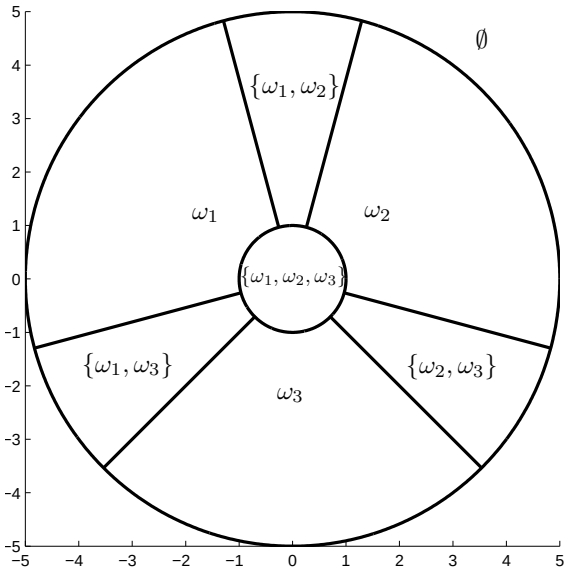

FIG. 4.5: Partitionnement de l'espace d'attributs avec rejet et sélection de classes.

proposons un nouveau système d'inférence flou fondé sur la mise en place de trois règles floues correspondant aux trois options possibles : rejet en ambiguïté, rejet en distance, classification exclusive. Chacune des trois situations est décrite par une mesure de confiance, donnant lieu à un degré de vérité pour chacune des règles. On trouvera l'article concernant ces règles en ANNEXE .

4.2.3 Sélection du nombre de classes

Les règles de décision diffèrent souvent dans le choix du type de sortie. Alors que pour la règle de Bayes, les sorties possibles sont des singletons (FIG. 4.2), Chow modifie l'espace de sortie en ajoutant Ω aux singletons (FIG. 4.3). Dans [Ha, 1997], l'auteur propose une nouvelle modification de l'espace des décisions. Cette fois-ci, les décisions possibles appartiennent à l'ensemble des sous-ensembles de Ω privé de $\emptyset$ (voir FIG. 4.5). Ce partitionnement de l'espace peut ainsi comprendre jusqu'à $2^c - 1$ sous-ensembles, ou régions. Plusieurs propositions de partitionnement ont été faites jusqu'à maintenant. La première, la plus utilisée, consiste à prendre les k plus grands degrés d'appartenance (*top-k ranking* rule), où k est un paramètre global. En d'autres termes, chaque observation $\mathbf{x}$ sera affectée au même nombre de classes, quelle que soit sa position dans $\mathbb{R}^p$. Une autre règle populaire est celle dite du risque constant (*constant risk* rule). Ici, on choisit le nombre minimum de classes pour chaque forme de manière à ce que le risque cumulé des probabilités a posteriori ne dépasse pas un seuil t [Gupta, 1965]. En adoptant le formalisme proposé par la suite, cette règle peut se réécrire comme la sélection de ce nombre

$$n(\mathbf{x},t) = \min_{k\in[1,c]} \left\{k : 1 - \sum_{i=1}^{k} u_{(i)}(\mathbf{x}) \leq t\right\} \tag{4.22}$$

Si l'on choisit pour chaque forme l'ensemble des classes, le taux d'erreur devient nul. Pour empêcher la sélection de cette partition triviale, le nombre moyen de classes est introduit

[Ha, 1997] :

$$\overline{n} = \int_X n(\mathbf{x}) P(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{4.23}$$

où $n(\mathbf{x})$ est le nombre de classes que l'on affecte à $\mathbf{x}$. La règle de décision optimale est alors la règle qui minimise le taux d'erreur pour un nombre moyen de classes donné. Le risque introduit par Ha devient alors

$$r_{Ha}(\mathbf{x}) = 1 - \sum_{i \text{ sélectionnés}} u_i(\mathbf{x}) \tag{4.24}$$

et la combinaison des k classes de plus grand degré d'appartenance parmi les c classes donne

$$r_{Ha}(\mathbf{x}) n(\mathbf{x}) = 1 - \sum_{i=1}^{k} u_{(i)}(\mathbf{x}) \tag{4.25}$$

L'auteur aboutit ainsi à la règle de sélection de classe suivante

$$n^\star(\mathbf{x},t) = \min_{k \in [1,c]} \{k : u_{(k+1)}(\mathbf{x}) \leq t\} \tag{4.26}$$

avec la convention $u_{(c+1)}(\mathbf{x}) = 0$. Le domaine de spécification pour t est $[0, 1/2]$, puisque si $t > 1/2$, on obtient la règle de Bayes sans rejet. A l'appui d'exemples où les décisions prises par la règle de Ha sont contre intuitives, Horiuchi [Horiuchi, 1998] propose une nouvelle règle :

$$n^\star(\mathbf{x},t) = \min_{k \in [1,c]} \{k : u_{(k)}(\mathbf{x}) - u_{(k+1)}(\mathbf{x}) \geq t\} \tag{4.27}$$

toujours avec la convention $u_{(c+1)}(\mathbf{x}) = 0$. Cette fois-ci le rejet peut intervenir pour un seuil compris dans $[0, 1]$. Si pour tout k dans $[1, c]$, on a $u_{(k)}(\mathbf{x}) - u_{(k+1)}(\mathbf{x}) < t$, alors on fixe $n^\star(\mathbf{x},t) = c$.

4.3 De nouvelles mesures de sélection

La définition de mesures d'ambiguïté est étroitement liée à la notion de similarité entre degrés d'appartenance. Nous proposons ainsi d'utiliser les mesures introduites dans le Chapitre 2 afin de définir de nouvelles règles (sélectives ou non) de rejet.

4.3.1 Sélection par blocs

Pour la sélection d'un (sous)-ensemble de classes, nous proposons d'utiliser la mesure $\Phi_{j,k}^{\mathcal{K}_\lambda}$ introduite dans la section 2.2.3 afin d'évaluer l'ambiguïté associée à $\mathbf{x}$ [Le Capitaine and Frélicot, 2008b]. Nous rappelons la définition de l'opérateur $\Phi_{j,k}(\mathbf{u})$, où $\mathbf{u}$ sera dans ce cadre associé aux degrés d'appartenance de $\mathbf{x}$ aux différentes classes de Ω

$$\Phi_{j,k}^{\mathcal{K}_\lambda}(\mathbf{u}(\mathbf{x})) = \begin{cases} \dfrac{\perp_{i=\frac{k+j}{2}}^{k} u_{(i)} \top \mathcal{K}_\lambda(i,k)}{\perp_{i=\frac{k+j}{2}}^{j} u_{(i)} \top \mathcal{K}_\lambda(i,j)} & \text{si } k-j \text{ est pair} \\ \dfrac{\perp_{i=\frac{k+j+1}{2}}^{k} u_{(i)} \top \mathcal{K}_\lambda(i,k)}{\perp_{i=\frac{k+j-1}{2}}^{j} u_{(i)} \top \mathcal{K}_\lambda(i,j)} & \text{si } k-j \text{ est impair} \end{cases} \tag{4.28}$$

Puisque une valeur élevée de $\Phi_{1,k}^{\mathcal{K}_\lambda}$ révèle que les k plus grandes valeurs de $\mathbf{u}(\mathbf{x})$ sont similaires, alors $\mathbf{x}$ peut être associé avec les classes correspondantes. Cet opérateur permet donc de définir une ambiguïté d'ordre k. Le schéma général de sélection de classes peut donc s'écrire

$$n(\mathbf{x},t) = \min_{k\in[1,c]} \{k : \Phi_{1,k+1}^{\mathcal{K}_\lambda}(\mathbf{u}(\mathbf{x})) \leq t\} \tag{4.29}$$

Nous avons proposé dans [Le Capitaine and Frélicot, 2008b] une deuxième stratégie, fondée sur un schéma itératif, afin de trouver k. Cet algorithme (voir ALGORITHME 2) est en fait la seconde étape $\mathcal{H}$ et se résume comme *pour i variant de 1 à c, on fixe $h_i(x) = 1$ si $\Phi_{1,i}(\mu(x)) \geq t$, où t est un seuil défini par l'utilisateur.*

Algorithme 2 : Étape d'assignation $\mathcal{H} : \mathcal{L}_{pc} \rightarrow \mathcal{L}_{hc}^c$.

Données : Un vecteur $\mathbf{u}$ de degrés d'appartenance, un seuil de rejet d'appartenance s, un seuil de rejet en ambiguïté t.
Résultat : Un vecteur $\mathbf{h}(\mathbf{x})$ d'affectation sélective.

```
début
    si u_(1)(x) < s alors
        h_i(x) ← 0 ∀i = 1,c
    fin
    si Σ_{i=1}^{c} h_i(x) > 0 alors
        pour i ← 1 à c faire
            si Φ_{1,i}(u(x)) ≥ t alors
                h_i(x) ← 1
            fin
            sinon
                h_i(x) ← 0
            fin
        fin
    fin
    retourner h(x)
fin
```

Par convention,

$$\Phi_{1,1}^{\mathcal{K}_\lambda}(\mathbf{u}(\mathbf{x})) = \frac{u_{(1)}(\mathbf{x})}{u_{(1)}(\mathbf{x})} = 1 \tag{4.30}$$

quel que soit le couple de t-normes. Ainsi, $\Phi_{1,1}^{\mathcal{K}_\lambda}(\mathbf{u}(\mathbf{x}))$ est toujours plus grande que t, pour tout t dans $[0,1]$. Cette propriété assure que au moins une classe sera sélectionnée, celle correspondant au degré d'appartenance maximum, c'est à dire celle sélectionnée par la règle de classification optimale de Bayes, mais cette fois-ci dans le sens de Chow [Chow, 1970]. En particulier, si l'on fixe t à 1, il n'y aura pas de rejet en ambiguïté[3]. Également par convention,

$$\Phi_{0,1}^{\mathcal{K}_\lambda}(\mathbf{u}(\mathbf{x})) = \frac{u_{(1)}(\mathbf{x})}{u_{(0)}(\mathbf{x})} = u_{(1)}(\mathbf{x}) \tag{4.31}$$

Les résultats expérimentaux seront donnés en section 4.5.

4.3.2 Approche logique et unification de règles usuelles

Dans [Le Capitaine and Frélicot, 2009a], nous proposons une nouvelle manière de définir une mesure d'ambiguïté à partir d'implications floues (section 2.2.4). Dans [Hirota and

3. Sauf dans le cas fortement improbable en pratique où $u_{(1)}(\mathbf{x}) = u_{(2)}(\mathbf{x})$. Dans ce cas, deux classes seront sélectionnées, ce qui reste cohérent par rapport à la pratique.

Pedrycz, 1991], les auteurs proposent de calculer le degré d'égalité entre deux quantités floues x et y par

$$(x \equiv y) = \frac{1}{2}\Big((x \to y) \wedge (y \to x) + (\overline{x} \to \overline{y}) \wedge (\overline{y} \to \overline{x})\Big) \tag{4.32}$$

où $\wedge$ est le minimum, $\to$ est une implication et $\overline{x}$ est la négation stricte $\overline{x} = 1 - x$.
Selon l'implication utilisée, certaines simplifications peuvent être opérées. En particulier, si $\to$ respecte le principe de confinement (voir section 2.2.4), et utilisant le fait que 1 est élément neutre de $\wedge$, on obtient

$$(x \equiv y) = \frac{1}{2}((x \to y) + (\overline{y} \to \overline{x})) \tag{4.33}$$

Comme dans le cas des mesures de comparaison, on retrouve encore une fois le rôle clé de la propriété de confinement.

Proposition 4.1 ([Le Capitaine and Frélicot, 2009a]). *À partir d'un ensemble de c degrés de vérité, triés de manière décroissante : $u_{(1)}(\mathbf{x}) \geq \cdots \geq u_{(c)}(\mathbf{x})$. Soient les deux prédicats ($\mathbf{x}$ est ω_i), de valeur $u_{(i)}(\mathbf{x})$, et ($\mathbf{x}$ est ω_k), de valeur $u_{(k)}(\mathbf{x})$. La valeur de l'implication* si l'observation $\mathbf{x}$ est ω_i, alors $\mathbf{x}$ est aussi ω_j, $\forall j$ variant de $i+1$ à k *est une mesure d'ambiguïté donnée par $I(u_{(i)}(\mathbf{x}),u_{(k)}(\mathbf{x}))$. Plus précisément, dans le contexte du rejet, on définit l'ambiguïté d'ordre k par*

$$\Phi_{k,I_\top}(\mathbf{u})) = I(u_{(k-1)}(\mathbf{x}),u_{(k)}(\mathbf{x})) \tag{4.34}$$

Puisque par convention $u_{(i)}(\mathbf{x}) \geq u_{(k)}(\mathbf{x})$, on suppose qu'il est plus probable d'associer $\mathbf{x}$ à la classe ω_i qu'à la classe ω_k, et on a évidemment $I(u_i(\mathbf{x}),u_{i+1}(\mathbf{x})) \geq I(u_{(i)}(\mathbf{x}),u_{(k)>i+1}(\mathbf{x}))$ par non décroissance avec la seconde variable. Introduit dans le schéma de sélection de classes, on obtient

$$n^\star(\mathbf{x},t) = \min_{k\in[1,c]} \Big\{k : I\Big(u_{(k)}(\mathbf{x}),u_{(k+1)}(\mathbf{x})\Big) \leq t\Big\} \tag{4.35}$$

Puisque $I(x,0) = 0$ si $x \neq 0$, c classes sont sélectionnées si $I\Big(u_{(k)}(\mathbf{x}),u_{(k+1)}(\mathbf{x})\Big) > t$ pour tout $k \in [1,c-1]$, et ce sans convention supplémentaire, contrairement aux règles de Ha et Horiuchi.

Proposition 4.2 ([Le Capitaine and Frélicot, 2009a]). *Soit $\top = \min$. Si on utilise la mesure d'ambiguïté engendrée par l'implication résiduelle $I_\top$ dans (4.35), on obtient la règle de Ha [Ha, 1997].*

Démonstration. On a $\top = \min$, l'implication résiduelle $I_\top$ associée est donc l'implication de Gödel définie par

$$I(x,y) = \begin{cases} 1 & \text{si } y \geq x \\ y & \text{si } y < x \end{cases} \tag{4.36}$$

Comme $u_{(k)}(\mathbf{x}) \geq u_{(k+1)}(\mathbf{x})$, l'implication $I_\top$ se réduit à $I\Big(u_{(k)}(\mathbf{x}),u_{(k+1)}(\mathbf{x})\Big) = u_{(k+1)}(\mathbf{x})$. Insérée dans (4.35), on obtient la règle de sélection

$$n^\star(\mathbf{x},t) = \min_{k\in[1,c]} \Big\{k : u_{(k+1)}(\mathbf{x}) \leq t\Big\} \tag{4.37}$$

ce qui termine la preuve. □

Proposition 4.3 ([Le Capitaine and Frélicot, 2009a]). *Prenons la t-norme $\top_L$ de Lukasiewicz. Si on utilise la mesure d'ambiguïté engendrée par l'implication résiduelle $I_\top$ dans (4.35), on obtient la règle de Horiuchi [Horiuchi, 1998].*

Démonstration. On a $\top = \top_L$, l'implication résiduelle $I_\top$ associée est donc l'implication de Lukasiewicz définie par

$$I(x,y) = \min(1, 1 - x + y) \tag{4.38}$$

Ici encore, puisque $u_{(k)}(\mathbf{x}) \geq u_{(k+1)}(\mathbf{x})$, on a $1 - u_{(k)}(\mathbf{x}) + u_{(k+1)}(\mathbf{x}) \leq 1$. L'implication se réduit donc à $I\Big(u_{(k)}(\mathbf{x}),u_{(k+1)}(\mathbf{x})\Big) = 1 - u_{(k)}(\mathbf{x}) + u_{(k+1)}(\mathbf{x})$ Insérée dans (4.35), on obtient la règle de sélection

$$n^\star(\mathbf{x},t) = \min_{k\in[1,c]} \Big\{k : 1 - u_{(k)}(\mathbf{x}) + u_{(k+1)}(\mathbf{x}) \leq t\Big\} \tag{4.39}$$

ce qui est équivalent à

$$n^\star(\mathbf{x},s) = \min_{k\in[1,c]} \Big\{k : 1 - u_{(k)}(\mathbf{x}) + u_{(k+1)}(\mathbf{x}) \leq 1 - s\Big\} \tag{4.40}$$

si l'on pose $s = 1 - t$. Comme le domaine de t dans la règle de Horiuchi est $[0, 1]$, ce changement de variable est valide, et on obtient enfin

$$n^\star(\mathbf{x},s) = \min_{k\in[1,c]} \Big\{k : u_{(k)}(\mathbf{x}) - u_{(k+1)}(\mathbf{x}) \geq s\Big\} \tag{4.41}$$

ce qui termine la preuve. □

Proposition 4.4 ([Le Capitaine and Frélicot, 2009a]). *Prenons la t-norme $\top_A$ produit. Si on utilise la mesure d'ambiguïté engendrée par l'implication résiduelle $I_\top$ dans (4.35), on obtient la mesure de Frélicot & Dubuisson [Frélicot and Dubuisson, 1992].*

Démonstration. On a $\top = \top_A$, l'implication résiduelle $I_\top$ associée est donc l'implication de Goguen définie par

$$I(x,y) = \begin{cases} 1 & \text{si } y \geq x \\ \frac{y}{x} & \text{si } y \leq x \end{cases} \tag{4.42}$$

Comme $u_{(k)}(\mathbf{x}) \geq u_{(k+1)}(\mathbf{x})$, l'implication $I_\top$ se réduit à

$$I\Big(u_{(k)}(\mathbf{x}),u_{(k+1)}(\mathbf{x})\Big) = u_{(k+1)}(\mathbf{x})/u_{(k)}(\mathbf{x}).$$

Insérée dans (4.35), on obtient la règle de sélection

$$n^\star(\mathbf{x},t) = \min_{k\in[1,c]} \left\{k : \frac{u_{(k+1)}(\mathbf{x})}{u_{(k)}(\mathbf{x})} \leq t\right\} \tag{4.43}$$

Si l'on pose $k = 1$, on obtient le rapport proposé dans [Frélicot and Dubuisson, 1992], ce qui termine la démonstration. □

L'utilisation d'autres implications résiduelles engendrées par d'autres t-normes, éventuellement paramétriques, permet de définir de nouvelles règles. Le schéma proposé représente donc une famille de règles permettant de retrouver les règles existantes, et d'en proposer de nouvelles (autant que de t-normes, c'est à dire une infinité). Comme le choix du couple

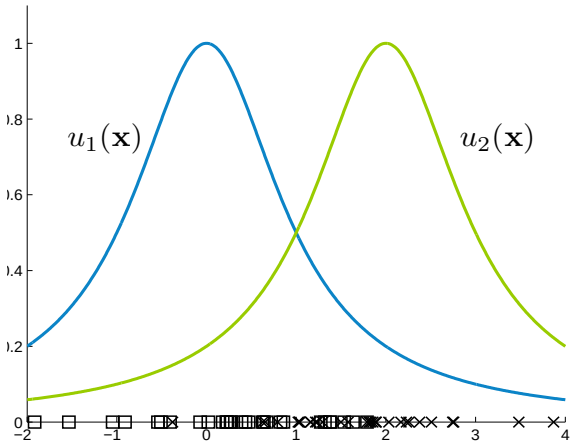

FIG. 4.6: Degrés d'appartenance aux classes ω_1 et ω_2 pour $\mathbf{x} \in \mathbb{R}$.

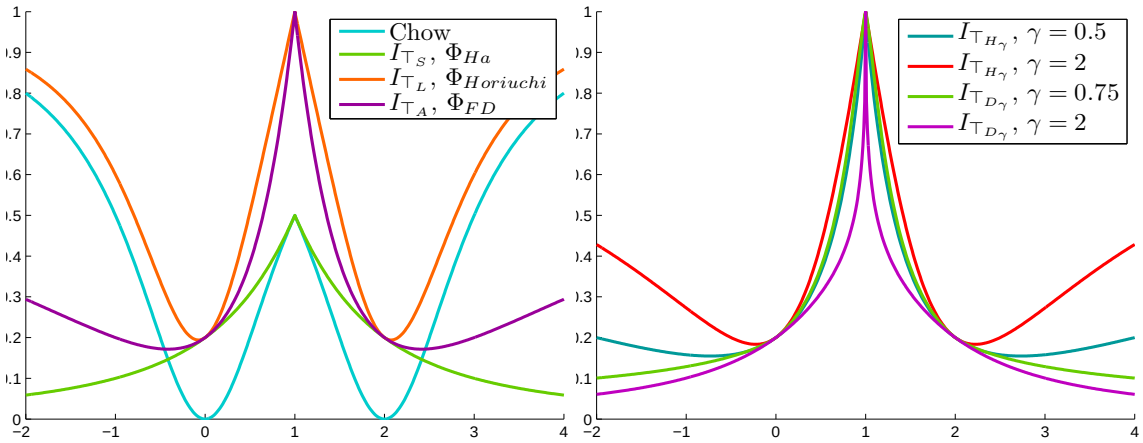

FIG. 4.7: Mesures d'ambiguïté de Chow, Ha, Horiuchi et Frélicot & Dubuisson. Les trois dernières sont également obtenues en prenant respectivement $I_{\top_S}$, $I_{\top_L}$ et $I_{\top_A}$ dans la règle proposée (*gauche*). Mesures d'ambiguïté engendrées par les implications de Hamacher et Dombi et pour différentes valeurs de γ (*droite*).

dual lors de l'utilisation de t-normes (et possiblement le paramètre γ dans le cas des familles paramétriques) est toujours une difficulté en pratique, nous proposons un exemple de comportement des mesures proposées dans le cas de données mono-dimensionnelles composées de deux classes (□ et ×) de distribution normale, et décrites par un modèle fondé sur une distance aux prototypes (équation (4.4)), voir FIG. 4.6. La stratégie de Chow et les valeurs des implications des $\mathbf{u}(\mathbf{x})$ fondées sur les t-normes min, produit et de Łukasiewicz, qui correspondent respectivement aux schémas de Ha, Frélicot et Horiuchi, sont tracées sur la FIG. 4.7 à gauche, et celles des implications paramétriques de Dombi et Hamacher le sont à droite. Nous donnons également les valeurs des implications des $\mathbf{u}(\mathbf{x})$ fondées sur les familles paramétriques de Yager et Frank en FIG. 4.8-(*gauche*), ainsi que celle obtenues par la famille de Dubois et Prade, FIG. 4.8-(*droite*).

Les graphiques montrent clairement que les schémas de Chow et Horiuchi (implication de Łukasiewicz) conduisent à rejeter de trop nombreux points, qu'ils soient ambigus (au centre) ou atypiques (à l'extérieur). Au contraire, les schémas fondés sur les implications produits (Frélicot) ou paramétriques (pour autant que le paramètre soit approprié) permettent de différencier les deux situations, et l'on peut donc s'attendre à de meilleurs performances

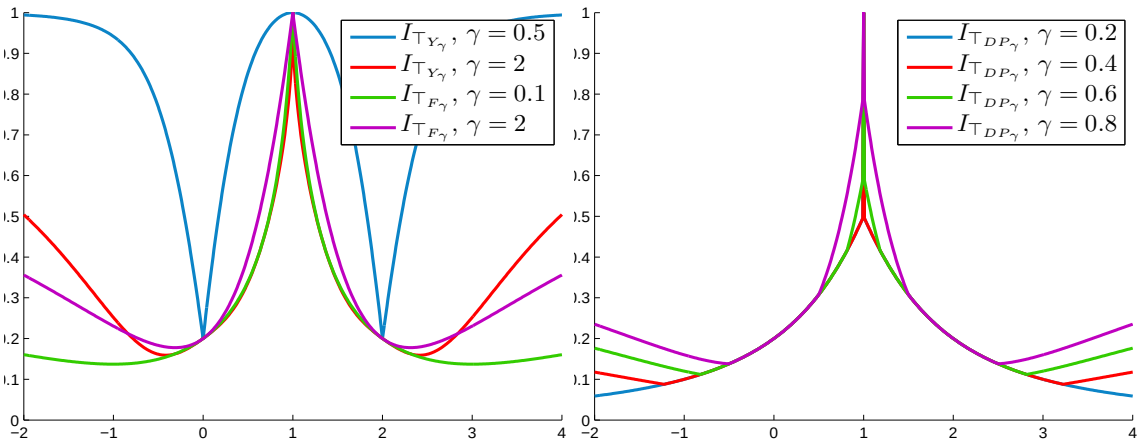

FIG. 4.8: Mesures d'ambiguïté engendrées par les implications de Yager et Frank (*gauche*) et Dubois-Prade pour différentes valeurs de γ (*droite*).

des schémas paramétriques. En particulier, les valeurs de γ pour les implications suivantes :

- Hamacher, avec γ plutôt fort (ici ≈ 2),
- Dombi, γ plutôt faible (ici ≈ 0.75),
- Yager, γ plutôt fort (ici ≈ 2),
- Frank, γ plutôt fort (ici ≈ 2),
- Dubois-Prade, γ plutôt fort (ici ≈ 0.8).

Remarquons enfin, pour ces dernières, comment le choix du paramètre γ permet de rejeter en ambiguïté et en distance (une grande valeur de γ pour la famille de Hamacher, et une faible valeur pour celle de Dombi), comme précisé dans la discussion. De manière plus générale, on peut définir les stratégies suivantes.

Proposition 4.5 ([Le Capitaine and Frélicot, 2009a])**.** *Soit $\Phi_k(\mathbf{u}(\mathbf{x}))$ une mesure d'ambiguïté quelconque de $\mathbf{u}$ à l'ordre k. Alors, le nombre optimal de classes $n^\star$ à sélectionner au sens du critère $\Phi_k(\mathbf{u})$ est donné par*

$$n^\star(\mathbf{x},t) = \min_{k\in[1,c]} \{k : \Phi_{k+1}(\mathbf{u}(\mathbf{x})) \leq t\} \tag{4.44}$$

où l'on prend comme convention $u_{(c+1)}(\mathbf{x}) = 0$.

4.3.3 Double option de rejet : une étape pour une règle

Afin de construire une règle permettant également de rejeter en distance, c'est à dire un partitionnement de l'espace en 2^c régions, il faut que la règle puisse sélectionner un nombre de classes égal à 0. Nous proposons donc l'extension de (4.44).

Proposition 4.6 ([Le Capitaine and Frélicot, 2009a])**.** *Soit $\Phi_k(\mathbf{u}(\mathbf{x}))$ une mesure d'ambiguïté quelconque de $\mathbf{u}$ à l'ordre k. Alors, le nombre optimal de classes $n^\star$ à sélectionner au sens du critère $\Phi_k(\mathbf{u})$ est donné par*

$$n^\star(\mathbf{x},t) = \min_{k\in[0,c]} \{k : \Phi_{k+1}(\mathbf{u}(\mathbf{x})) \leq t\} \tag{4.45}$$

où l'on prend comme convention $u_{(0)}(\mathbf{x}) = 1$ et $u_{(c+1)}(\mathbf{x}) = 0$.

On obtient ainsi une fonction $\mathcal{H}$ d'assignation unifiée, ou n'importe quelle mesure d'ambiguïté peut être employée. L'affectation de $\mathbf{x}$ en elle-même est décrite dans l'ALGORITHME 3.

Algorithme 3 : Étape d'assignation généralisée $\mathcal{H} : \mathcal{L}_{pc} \to \mathcal{L}_{hc}^{c}$.
Données : Un vecteur $\mathbf{u}$ de degrés d'appartenance dans $\mathcal{L}_{pc}$, un seuil de rejet t.
Résultat : Un vecteur $\mathbf{h}(\mathbf{x})$ dans $\mathcal{L}_{hc}^{c}$ d'affectation sélective.
début
  On fixe $\mathbf{h}(\mathbf{x}) \leftarrow 0$
  Avec une mesure d'ambiguïté $\Phi_k(\mathbf{u}(\mathbf{x}))$ donnée, on calcule $n^\star(\mathbf{x},t)$ avec (4.45).
  pour chaque $j \leftarrow 1$ **à** $n^\star(\mathbf{x},t)$ **faire**
    $h_j(\mathbf{x}) \leftarrow 1$ dans le sens décroissant des $u_{(j)}(\mathbf{x})$.
  fin
  retourner $\mathbf{h}(\mathbf{x})$
fin

Remarque 4.1. Dans le cas d'utilisation de mesures d'ambiguïté engendrées par des implications floues, le rejet de distance est simplifié. Par le principe de bord (satisfait par les quatre types d'implication), on peut procéder au rejet en distance, puisque $I(u_{(0)}(\mathbf{x}), u_{(1)}(\mathbf{x})) = I(1, u_{(1)}(\mathbf{x})) = u_{(1)}(\mathbf{x})$, c'est à dire un test de rejet en appartenance usuel. Lorsque $k = 0$ dans (4.45), la mesure d'ambiguïté Φ devient une mesure d'acceptation $\Psi(\mathbf{u}) = u_{(1)}(\mathbf{x})$.

Dans ce cadre, nous proposons une nouvelle définition d'opérateur de similarité par blocs compatible avec la règle de rejet sélectif (4.45).

Proposition 4.7. *Soit* $\Phi_{j,k}^{\mathcal{K}_\lambda}(\mathbf{u}(\mathbf{x}))$ *un opérateur de similarité du bloc de* $\mathbf{u}$ *indicé par* j *et* k*. Un nombre de classes à sélectionner pour* $\mathbf{x}$ *avec un* t *donné est trouvé par*

$$n^\star(\mathbf{x},t) = \min_{k \in [0,c]} \left\{ k : \Phi_{\mathbb{1}_k, k+1}^{\mathcal{K}_\lambda}(\mathbf{u}(\mathbf{x})) \leq t \right\} \tag{4.46}$$

où $\mathbb{1}_k$ *est la fonction indicatrice valant* 1 *si* $k > 0$, 0 *sinon.*

4.4 Une nouvelle méthode d'évaluation des règles sélectives

L'évaluation des options de rejet est un problème récurrent lorsque l'on souhaite mettre en place une stratégie de rejet. On distingue plusieurs approches pour cette évaluation. Si l'on classe n observations $\mathbf{x}$, alors on peut séparer celles-ci en trois : n_{corr} sont classées de manière correcte, n_{err} sont classées par erreur, et n_{rej} sont rejetées. On a évidemment

$$n_{corr} + n_{err} + n_{rej} = n \tag{4.47}$$

Selon l'ensemble des paramètres Θ de la stratégie (généralement des seuils), on obtient donc des taux de *performance* $P(\Theta)$, d'*erreur* $E(\Theta)$ et de *rejet* $R(\Theta)$ respectivement définis par n_{corr}/n, n_{err}/n et n_{rej}/n. Si l'on considère qu'à partir du moment où une forme est rejetée, elle ne rentre pas dans le calcul de la qualité de l'option de rejet, on peut aussi considérer un taux de *fiabilité* donné par $n_{corr}/(n_{corr} + n_{err})$. Cet indice de fiabilité est à utiliser avec précaution, puisque si l'on rejette beaucoup de formes, la fiabilité est élevée, mais beaucoup de décisions concernant les observations restent en suspend. Dans le cas limite, il suffit de rejeter assez de formes de sorte que le taux d'erreur soit nul, et la fiabilité sera de 100%. La manière la plus simple, et la moins complète, est donc de comparer pour un taux de rejet ou un taux d'erreur donné, le taux de performance $P(\Theta)$. Cette méthode peut servir

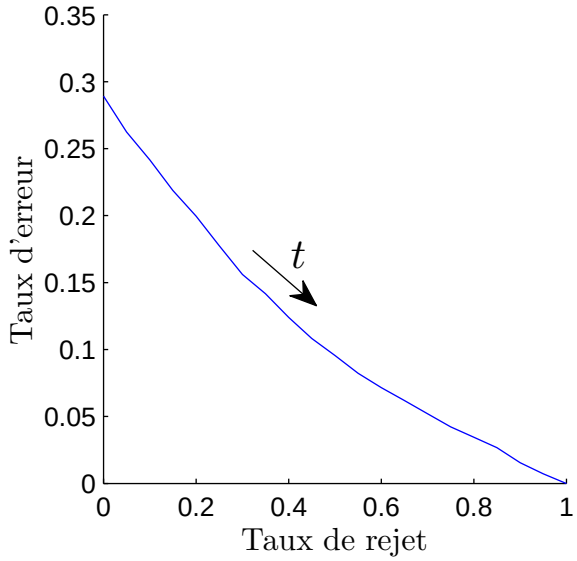

FIG. 4.9: Exemple d'une courbe Erreur Rejet.

dans les cas où l'application visée exige par exemple un taux d'erreur nul, quitte à rejeter des formes, ou encore que l'on accepte de rejeter jusqu'à 10% des données analysées. Cette situation correspond à un point de fonctionnement caractérisé par le couple $(E(\Theta),R(\Theta))$ pour un ensemble Θ spécifié.

L'objectif général de l'option de rejet étant de minimiser l'erreur et le rejet, l'évaluation précédente n'est pas assez complète. Une évaluation plus complète consiste à analyser le couple $(E(\Theta),R(\Theta))$ pour différentes valeurs de Θ. C'est le principe de compromis introduit par Chow [Chow, 1970], la courbe *Erreur-Rejet* (ER) étant dans son cas fonction de $\Theta = t$, voir FIG. 4.9. Dans le cas d'un classifieur Bayésien optimal, sa proposition de minimum de risque est également optimale par rapport à ce compromis. À partir de cette courbe, il convient ensuite de trouver un point de fonctionnement optimum. Dans notre cas, comme Θ se réduit le plus souvent à t, nous faisons varier le seuil sur le domaine de définition spécifié par la règle. Le point idéal se trouve à l'origine, mais comme ce point n'est en pratique pas atteignable, on cherche le point qui s'en rapproche le plus. Ceci peut se faire de deux manières. Soit en calculant une distance des points de fonctionnement disponibles à l'origine, puis en prenant le point le plus proche, soit en définissant des coûts de classement. En fonction de ces coûts, on peut tracer une *droite d'iso-coût* localisant le meilleur point de fonctionnement, voir [Golfarelli et al., 1997; Santo-Pereira and Pires, 2004]. Dans le cas optimal de Chow, cette droite a une pente de $-t$. Notons que des auteurs proposent une autre méthode de visualisation de ce compromis en observant $E(\Theta)/E_0$ en fonction de $R(\Theta)/E_0$, où E_0 est le taux d'erreur sans rejet [Hansen et al., 1997]. Comme on cherche à minimiser conjointement le taux d'erreur et de rejet, une manière plus générale d'évaluer une courbe ER que de considérer un seul point de fonctionnement consiste à calculer l'aire sous la courbe (*Aire sous la Courbe Erreur Rejet - AER*) :

$$AER = \int_0^1 E(R(t))\mathrm{d}t \tag{4.48}$$

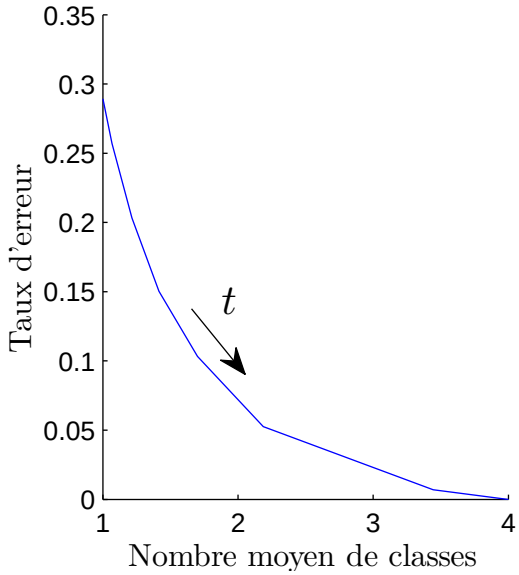

FIG. 4.10: Exemple d'une courbe Erreur Nombre moyen de classes.

Plus cette aire sera faible, meilleure sera la règle de rejet. Nous utiliserons cette mesure de performance pour évaluer nos règles de rejet (non-sélectif) et les comparer avec celles existantes sur des données synthétiques et réelles en section 4.5.

Une manière spécifique d'évaluer le rejet de distance est d'utiliser les courbes ROC (Receiver Operating Characteristic), [Fawcett, 2006]. En effet pour l'évaluation du rejet de distance, nous nous intéressons aux *vrais positifs* (VP) et aux *faux positifs* (FP). Pour différentes valeurs de t, la courbe $VP(t)$ en fonction de $FP(t)$ est une courbe ROC. Notons que des règles de rejet fondées sur des profits et leur maximisation (pour Chow, ce sont des coûts à minimiser) grâce aux courbes ROC ont été proposées pour des problèmes à deux classes, [Tortorella, 2005; Landgrebe et al., 2006]. La visualisation de courbes ROC pour des problèmes à plus de deux classes aboutit à une explosion combinatoire. Pour cette raison, Landgrebe & Duin introduisent des simplifications permettant de calculer la courbe ROC de dimension quelconque, [Landgrebe and Duin, 2008]. Appliquer cette proposition pour une nouvelle règle de rejet de la même manière que [Tortorella, 2005] peut faire l'objet de nouvelles recherches.

Dans le cas des règles de sélection de classes, on parle d'erreur lorsque la vraie classe de $\mathbf{x}$ n'est pas dans la liste des classes candidates. Clairement, plus le nombre de classes sélectionnées est élevé, plus le taux d'erreur sera faible. Le cas limite revient même à proposer l'ensemble des classes du problème pour chaque $\mathbf{x}$, le taux d'erreur est donc nul. Ce genre de décision ne sert à rien puisqu'aucune décision n'est prise. Une méthode de comparaison consiste à construire la courbe du taux d'erreur en fonction du nombre moyen de classes $\overline{n}$ défini par (4.23) en faisant varier t sur son domaine de définition, [Ha, 1997], voir FIG. 4.10 pour un exemple à 4 classes. De manière analogue aux courbes ER, comparer deux courbes *Erreur-Nombre Moyen de Classes* ($E\overline{n}$) peut se faire en calculant l'aire sous

la courbe (*Aire sous la Courbe Erreur Nombre Moyen de Classes - AE$\overline{n}$*).

Proposition 4.8. *L'aire sous la Courbe Erreur Nombre Moyen de Classes est donnée par*

$$AE\overline{n} = \int_0^1 E(\overline{n}(t))\mathrm{d}t \tag{4.49}$$

Plus cette aire sera faible, meilleure sera la règle de rejet sélectif. Nous utiliserons cette mesure de performance pour évaluer nos règles de rejet sélectif et les comparer avec celles existantes sur des données synthétiques et réelles en section 4.5.

4.5 Résultats expérimentaux

4.5.1 Les données

Pour présenter les résultats expérimentaux, nous avons choisi de proposer des données synthétiques et réelles représentant une variété en terme de

- nombre d'observations n,
- nombre d'attributs p,
- nombre de classes c,
- zones de chevauchement entre classes.

la plus large possible.
Pour ce faire, nous proposons trois jeux de données synthétiques :

- $D1$ contient $n = 2000$ points distribués selon un mélange de $c = 2$ distributions normales de dimension 7. Chaque classe comporte 1000 points de moyennes $\mathbf{v}_1 = (1\,0\,\cdots\,0)^T$ et $\mathbf{v}_2 = (-1\,0\,\cdots\,0)^T$, et de matrice de covariance $\Sigma_1 = \Sigma_2 = I$.
- $D2$ est constitué de $n = 4000$ points distribués selon un mélange de $c = 4$ distributions normales bi-dimensionnelle. Chaque classe comporte 1000 points de moyennes $\mathbf{v}_1 = (1\,1)^T$, $\mathbf{v}_2 = (1\ -1)^T$, $\mathbf{v}_3 = (-1\,1)^T$, $\mathbf{v}_4 = (-1\ -1)^T$, et de matrice de covariance $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = I$, voir Fig. 4.11-(*gauche*).
- DH est constitué de deux distributions normales de matrice de covariance différentes selon la distribution de Highleyman, [Highleyman, 1962]. Ces deux classes, chacune composées de 800 observations dans $\mathbb{R}^2$, se chevauchent, voir Fig. 4.11-(*droite*).

Les performances sur des jeux de données réelles de l'UCI, [Blake and Merz, 1998], sont également considérées. Leurs caractéristiques, ainsi que les degrés de chevauchement des classes sont donnés en Tab. 4.1. L'ensemble des taux d'erreur sont calculés par 10-validation croisée, voir section 3.1.1.5.

4.5.2 Étude comparative

Dans cette section, les mesures proposées seront comparées selon deux aspects. Dans un premier temps, nous analyserons la courbe ER, permettant de comparer entre elles les différentes mesures pour un taux de rejet variant de 0 à 1. Ensuite, afin de comparer les règles de sélection de classes, nous utiliserons et analyserons les courbes $E\overline{n}$. Dans chacun des tableaux récapitulatifs, les meilleurs scores seront indiqués en gras et rouge.
À titre d'exemple, nous donnons en Fig. 4.12 les courbes ER sur le jeu de données Pima pour l'ensemble des mesures d'ambiguïté Φ présentées. Par mesure de concision, les courbes

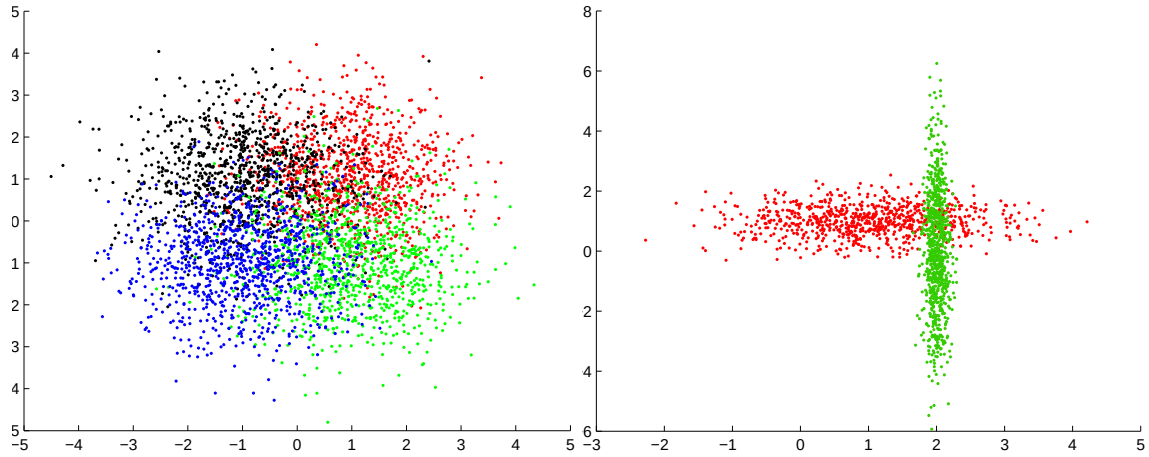

FIG. 4.11: Deux jeux de données synthétiques $D2$ (*gauche*), et DH (*droite*).

Données	n	p	c	**Chevauchement**
Ionosphere	351	34	2	très fort
Forest	495411	10	2	modéré
Vowel	528	10	11	léger
Digits	10992	16	10	léger
Thyroid	215	5	3	léger
Pima	768	9	2	fort
Statlog	6435	36	6	assez léger
Glass	214	9	6	modéré
Iris	150	4	3	léger
Cancer	699	9	2	modéré

TAB. 4.1: Les jeux de données réelles considérés et leurs caractéristiques : n, p, c et degré de chevauchement.

ER des autres jeux de données sont reportées en ANNEXE D.

À partir des résultats présentés en TAB. 4.2, plusieurs remarques peuvent être faites. Sur l'ensemble des jeux de données considérés, nos propositions donnent de meilleures performances que les mesures traditionnelles de Chow, Ha et Horiuchi. On note que selon les types de données, les performances particulièrement bonnes des mesures d'ambiguïté fondées sur les implication de Frank ($\gamma = 2$), Hamacher ($\gamma = 2$), Dombi ($\gamma = 0.75$) et Dubois-Prade, où la valeur de γ n'influe que très peu sur le résultat. Ces bonnes performances sont cependant dépendantes des données. En effet, la mesure de Ha sur les données artificielles donne des performances correctes relativement aux autres. Ceci s'explique par le fait que dans le cas

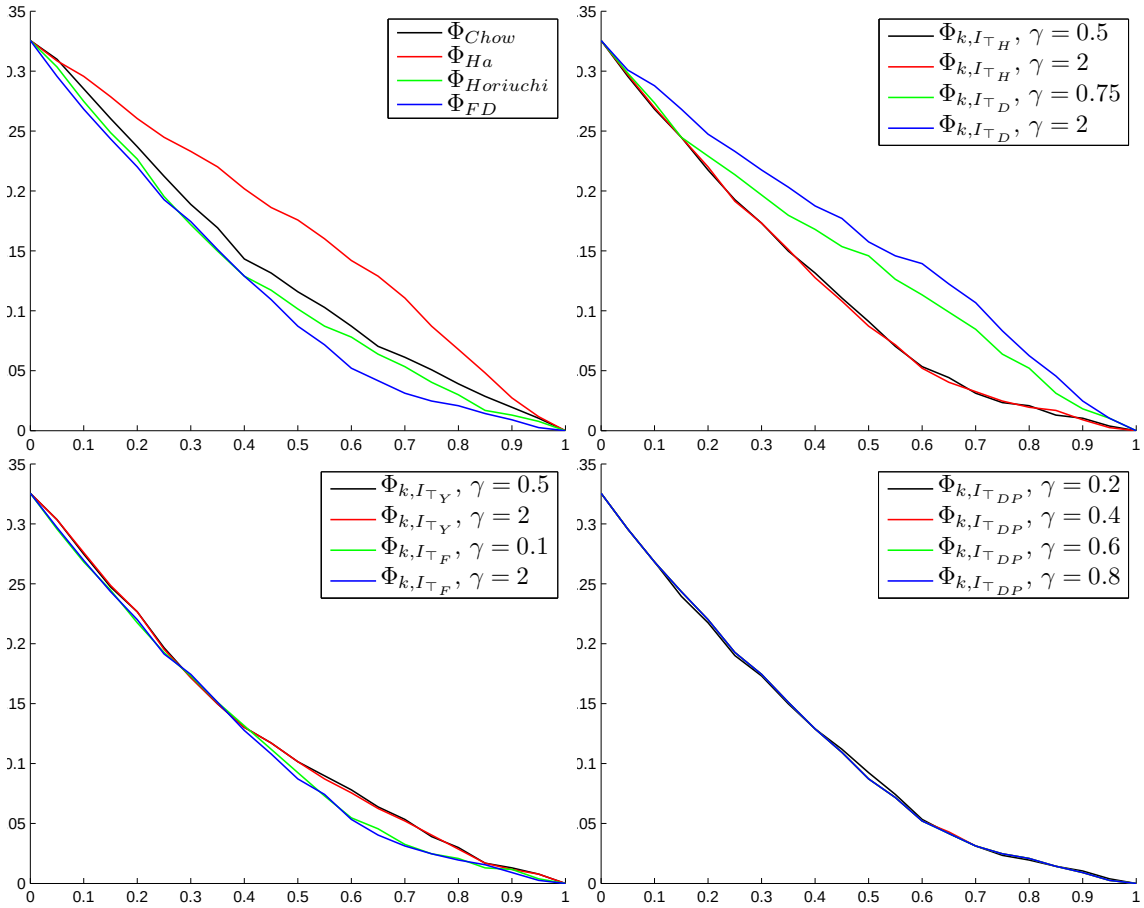

FIG. 4.12: Courbes *ER* sur les données *Pima*. Mesures usuelles (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

de données artificielles, les distributions sont connues et fixées, les degrés d'appartenance sont donc plus représentatifs des données, situation moins complexe que le cas de données réelles ne suivant pas de loi spécifique. Les implications se rapprochant de l'implication de Gödel ($\top_S$, correspondant à la mesure de Ha) donneront donc de bons résultats sur les données synthétiques, ce qui est le cas de l'implication de Dombi avec $\gamma = 0.75$, et de l'implication de Frank avec $\gamma = 0.1$.

Selon le degré de chevauchement des classes, on remarque que certaines mesures sont plus performantes que d'autres. En particulier, en cas de chevauchement léger (données *Vowel*, *Digits* et *Thyroid*), les implications de Frank et Dubois-Prade donnent de bons résultats, alors que dans le cas d'un chevauchement plus fort, les implications de Hamacher sont plus performantes. Une exception cependant, l'implication de Frank donne de bons résultats aussi bien sur des données présentant un léger chevauchement qu'un fort chevauchement. Sur les jeux de données considérés, les mesures obtenues à partir de l'implication de Yager ne donnent pas de bon résultats.

De manière générale, les mesures fondées sur des implications proches de celle de Gödel sont performantes pour des données synthétiques, alors que les mesures fondées sur des implications proches de celle de Goguen sont performantes pour des données réelles. La première n'utilise qu'une seule information (qu'un seul degré d'appartenance) et est donc

adaptée à la situation où la distribution est connue, puisque il n'y aura pas de dépendance entre classes. La seconde utilise deux informations, et prend donc comme hypothèse qu'il y a une interaction entre les classes, ce qui est le cas pour des données réelles non générées à partir de lois connues.

Remarque 4.2. Nous n'avons pas inclu l'opérateur de similarité par blocs $\Phi^{\mathcal{N}_\lambda}_{1,k+1}$ dans cette première étude, puisque c'est ici le rejet total qui est évalué. En d'autres terme, une ambiguïté d'ordre 2 suffit à rejeter l'observation. Dans le cas de l'ordre 2, ici $\Phi^{\mathcal{N}_\lambda}_{1,2}$, la mesure d'ambiguïté ne dépend ni de la norme triangulaire ni de la valeur du paramètre de résolution λ, c'est simplement le ratio $u_{(2)}/u_{(1)}$. Les résultats sont donc identiques à la mesure de Frélicot & Dubuisson.

Mesure Φ	$D1$	$D2$	DH	**Ionosphere**	**Forest**	**Vowel**	**Digits**	**Thyroid**	**Pima**	**Statlog**	**Glass**	**Iris**	**Cancer**
Chow	0.083	0.160	0.041	0.235	0.162	0.008	0.008	0.020	0.142	0.078	0.140	0.015	0.018
Mesure $\Phi_{k,I_\top}$													
Ha ($I_{\top_S}$)	0.053	0.107	0.009	0.149	0.156	0.006	0.011	0.012	0.175	0.070	0.088	0.005	0.053
Horiuchi ($I_{\top_L}$)	0.069	0.145	0.030	0.231	0.140	0.004	0.005	0.011	0.131	0.049	0.118	0.011	**0.017**
Frélicot ($I_{\top_A}$)	0.049	0.113	0.009	0.148	0.136	**0.002**	0.004	**0.007**	**0.123**	0.046	0.076	0.005	0.023
Hamacher, $\gamma = 0.5$	0.048	0.109	0.009	0.148	0.136	**0.002**	0.004	**0.007**	0.124	**0.044**	**0.075**	0.005	0.024
Hamacher, $\gamma = 2$	0.049	0.117	0.009	0.148	**0.135**	**0.002**	0.004	**0.007**	**0.123**	**0.044**	0.076	0.05	0.023
Dombi, $\gamma = 0.75$	**0.043**	**0.104**	**0.008**	0.149	0.138	0.003	0.006	0.011	0.151	0.052	0.079	**0.004**	0.041
Dombi, $\gamma = 2$	0.048	0.105	0.009	0.148	0.145	0.004	0.010	0.012	0.167	0.060	0.085	**0.004**	0.050
Yager, $\gamma = 0.5$	0.069	0.148	0.032	0.231	0.141	0.003	0.005	0.011	0.132	0.048	0.118	0.012	**0.017**
Yager, $\gamma = 2$	0.068	0.138	0.025	0.231	0.140	0.003	0.005	0.010	0.131	0.048	0.116	0.011	**0.017**
Frank, $\gamma = 0.1$	0.048	0.108	**0.008**	0.149	0.136	**0.002**	**0.003**	0.008	0.124	**0.044**	**0.075**	0.005	0.024
Frank, $\gamma = 2$	0.050	0.115	0.009	**0.147**	**0.135**	**0.002**	**0.003**	**0.007**	0.124	**0.044**	0.076	0.005	0.023
Dubois-Prade, $\gamma = 0.2$	0.049	0.105	0.009	0.148	0.137	**0.002**	**0.003**	**0.007**	**0.123**	0.045	0.076	**0.004**	0.024
Dubois-Prade, $\gamma = 0.4$	0.050	0.108	0.009	0.148	0.136	**0.002**	**0.003**	**0.007**	0.124	0.045	0.076	0.005	0.024
Dubois-Prade, $\gamma = 0.6$	0.050	0.110	0.009	0.148	0.136	**0.002**	**0.003**	**0.007**	0.124	0.045	0.076	0.005	0.024
Dubois-Prade, $\gamma = 0.8$	0.050	0.112	0.009	0.148	0.136	**0.002**	**0.003**	**0.007**	**0.123**	0.045	0.076	0.005	0.023

TAB. 4.2: Résultats pour les mesures $\Phi_{k,I_\top}$: Aire sous la courbe ER (AER), à minimiser. Les meilleurs résultats sont indiqués en gras et rouge.

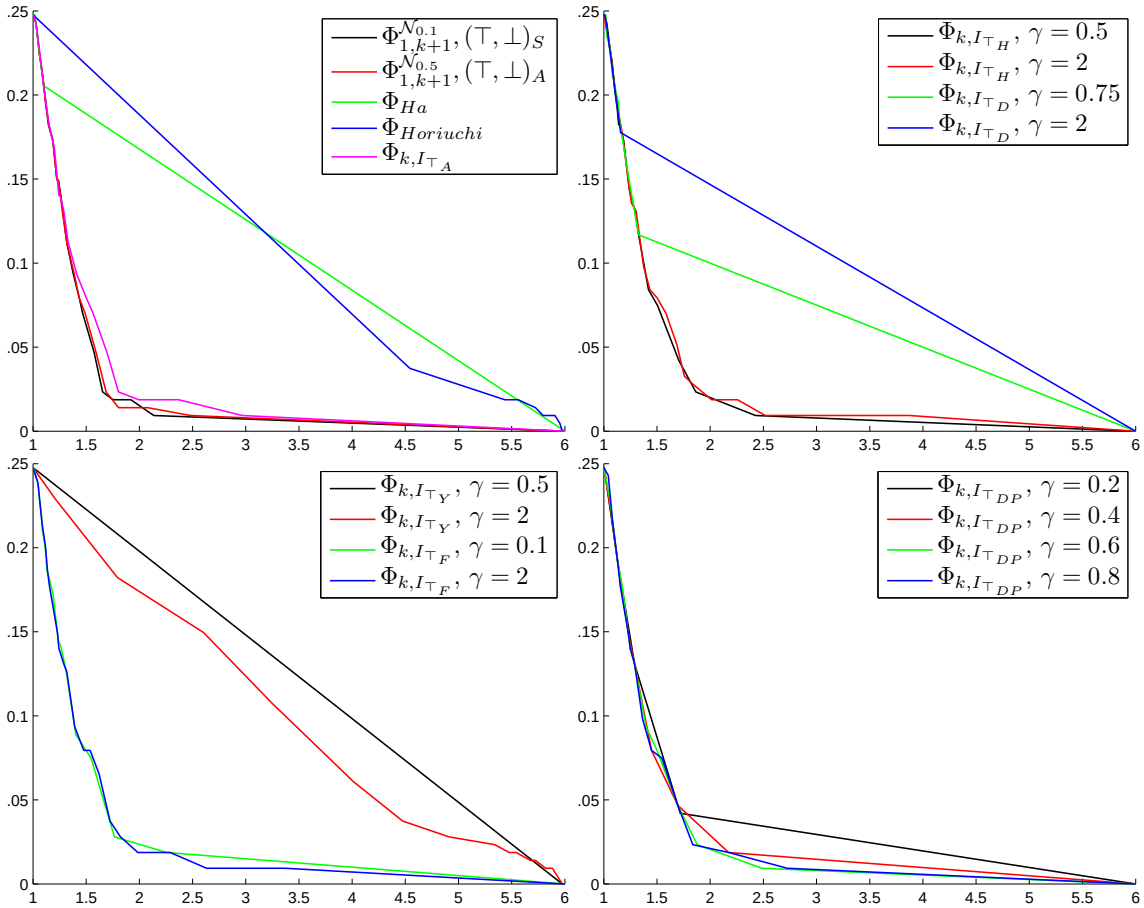

FIG. 4.13: Courbes $E\overline{n}$ sur les données *Glass*. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

Comme nous l'avons précisé dans la section précédente, l'évaluation de règles de sélection de classes est différente de celle d'une règle de rejet usuel. Comme l'objectif est de sélectionner plus d'une classe de manière à être plus sûr du résultat, tout en n'en choisissant pas trop, il faut trouver un compromis entre le taux d'erreur et le nombre de classes choisies. Nous présentons en TAB. 4.3 les résultats en termes d'aire sous la courbe $E\overline{n}$ de l'ensemble des règles de sélection de classes de la littérature ainsi que celles que nous proposons. Remarquons l'absence de la mesure de Chow, puisque elle ne définit pas un règle sélective. Nous donnons un exemple de courbe $E\overline{n}$ pour le jeu de données *Glass* en FIG. 4.13, l'ensemble des courbes $E\overline{n}$ pour les autres jeux de données étant disponible en ANNEXE D.

Ici encore, par mesure de concision, nous limiterons notre étude au cas où $\mathcal{K}$ est un noyau gaussien. Suite à la discussion dans le CHAPITRE 2 concernant l'influence du paramètre de résolution λ par rapport au choix du couple de t-normes, nous fixons $\lambda = 0.1$ lors de l'utilisation $(\top,\bot)_S$, et $\lambda = 0.5$ pour $(\top,\bot)_A$.
Pour la même raison que celle évoquée dans la Remarque 4.2, lorsque $c = 2$, l'évaluation de l'ambiguïté par blocs de $\Phi^{\mathcal{N}_\lambda}_{1,k+1}$ se réduit à la mesure de Frélicot & Dubuisson. L'intérêt principal de cet opérateur apparaît lorsque $c > 2$. Les résultats du TAB. 4.3 montrent que

lorsque $c > 2$ (c'est à dire les données $D2$, $Vowel$, $Digits$, $Thyroid$, $Statlog$ et $Glass$), $\Phi_{1,k+1}^{\mathcal{N}_\lambda}$ est le meilleur des opérateurs étudiés (hormis pour $D2$, où l'implication de Dombi, et $\gamma = 0.75$ donne un meilleur résultat).
Comme dans l'étude des courbes ER, les implications de Hamacher et Frank se révèlent globalement meilleures que les autres. Ici, dans le cas où les classes se chevauchent fortement, la règle sélective fondée sur l'implication de Hamacher sera préconisée ($Ionosphère$ et $Pima$), tandis que l'on choisira plutôt l'implication de Frank si le chevauchement est plus léger ($D1$ et DH).

En conclusion, nous avons proposé de nouvelles mesures d'ambiguïté pour des règles de rejet soit total, soit sélectif. Fondées sur des implications résiduelles floues, ces mesures généralisent des mesures précédemment proposées dans différents articles selon la norme triangulaire utilisée pour engendrer l'implication. Nous avons comparé ces mesures dans le cadre du rejet total avec celles existantes par le biais d'un protocole usuel lors de l'évaluation de l'option de rejet : l'aire sous la courbe ER. À partir des expériences sur des données synthétiques et réelles, il apparaît que l'utilisation de normes triangulaires supérieures à la t-norme Lukasiewicz et inférieures à la t-norme Standard pour calculer les implications résiduelles permet d'obtenir de meilleurs résultats. De manière générale, notre proposition permet d'obtenir des performances que n'atteignent pas les mesures existantes. En particulier, la mesure associée à l'implication résiduelle de Frank avec $\gamma = 2$ donne lieu à de très bons résultats.
Pour la comparaison des règles de rejet sélectif, nous avons introduit la notion d'aire sous la courbe Erreur-Nombre Moyen de Classes ($E\overline{n}$), par analogie aux aire sous les courbes ROC et ER. Dans le cas de rejet sélectif, les meilleurs résultats sont obtenus avec l'utilisation de l'opérateur d'ambiguïté par blocs. Si le problème présente seulement deux classes, la notion de bloc n'est plus d'actualité, et les mesures fondées sur les implications permettent d'obtenir de meilleurs résultats.
Selon les désirs de l'utilisateur, nous avons ainsi donné des éléments de choix d'une règle de rejet parmi les nombreuses possibles.
Parmi les perspectives possibles, citons l'évaluation du rejet en distance de l'ensemble de ces règles, puisque aucun des jeux de données étudiés ne présente de points à rejeter en distance.

Dans la section suivante, nous allons voir comment ces mesures d'ambiguïté, évaluant finalement le chevauchement et la séparation des classes, peuvent être utilisées en classification non supervisée, et particulièrement pour la détermination du nombre de classes d'un groupe de points.

Mesure Φ	$D1$	$D2$	DH	**Ionosphere**	**Forest**	**Vowel**	**Digits**	**Thyroid**	**Pima**	**Statlog**	**Glass**	**Iris**	**Cancer**
$\Phi^{\mathcal{N}_{0.1}}_{1,k+1},(\top,\bot)_S$	0.046	0.207	0.008	0.163	**0.128**	**0.002**	**0.008**	**0.006**	0.115	**0.086**	**0.106**	0.009	0.023
$\Phi^{\mathcal{N}_{0.5}}_{1,k+1},(\top,\bot)_A$	0.046	0.214	0.008	0.163	**0.128**	**0.002**	**0.008**	0.007	0.115	0.090	0.109	0.010	0.023
Mesure $\Phi_{k,I_\top}$													
Ha ($I_{\top_S}$)	0.070	0.223	0.011	0.186	0.149	0.146	0.171	0.056	0.164	0.436	0.526	0.034	0.045
Horiuchi ($I_{\top_L}$)	0.073	0.408	0.029	0.196	0.151	0.146	0.167	0.034	0.141	0.436	0.537	0.022	0.026
Algébrique ($I_{\top_A}$)	0.046	0.250	0.008	0.163	**0.128**	0.003	0.010	0.007	0.115	0.116	0.125	0.009	0.023
Hamacher, $\gamma = 0.5$	**0.045**	0.230	0.008	0.166	**0.128**	0.003	0.010	**0.006**	0.116	0.115	0.118	0.008	0.024
Hamacher, $\gamma = 2$	0.046	0.274	0.008	**0.156**	0.129	0.003	0.010	0.007	**0.114**	0.116	0.128	**0.004**	0.022
Dombi, $\gamma = 0.75$	0.049	**0.205**	0.009	0.184	0.131	0.086	0.150	0.025	0.144	0.358	0.332	0.009	0.043
Dombi, $\gamma = 2$	0.066	0.218	0.011	0.186	0.140	0.146	0.171	0.042	0.157	0.435	0.464	0.028	0.045
Yager, $\gamma = 0.5$	0.074	0.422	0.029	0.185	0.158	0.146	0.173	0.056	0.162	0.436	0.616	0.033	0.043
Yager, $\gamma = 2$	0.066	0.379	0.024	0.222	0.133	0.039	0.173	0.018	0.124	0.366	0.508	0.021	**0.016**
Frank, $\gamma = 0.1$	**0.045**	0.222	0.008	0.171	0.129	0.003	0.011	0.010	0.116	0.116	0.136	0.008	0.026
Frank, $\gamma = 2$	0.046	0.263	**0.007**	0.160	0.129	0.003	0.009	0.006	**0.114**	0.115	0.126	0.009	0.023
Dubois-Prade, $\gamma = 0.2$	0.048	0.212	0.009	0.176	0.139	0.005	0.022	0.013	0.117	0.132	0.181	0.009	0.030
Dubois-Prade, $\gamma = 0.4$	0.047	0.224	0.009	0.172	0.131	0.004	0.012	0.010	0.115	0.118	0.138	0.008	0.026
Dubois-Prade, $\gamma = 0.6$	0.046	0.236	0.008	0.165	0.129	0.003	0.010	**0.006**	0.115	0.117	0.120	0.008	0.025
Dubois-Prade, $\gamma = 0.8$	0.046	0.245	0.008	0.166	**0.128**	0.003	0.009	0.007	**0.114**	0.116	0.122	0.009	0.024

TAB. 4.3: Résultats pour les règles sélectives $\Phi^{\mathcal{N}_\lambda}_{1,k+1}$ et $\Phi_{k,I_\top}$: Aire sous la courbe $E\overline{n}$ ($AE\overline{n}$), à minimiser. Les meilleurs résultats sont indiqués en gras et rouge.

Chapitre 5

Validation de partitions floues

Résumé : *Dans ce chapitre, une seconde application est proposée : la validation de partitions en classification non supervisée. Lorsque l'on dispose de données non étiquetées, on doit découvrir une certaine structure dans les données, et en particulier déterminer le nombre de groupes distincts. La validation de partitions consiste à inspecter différentes structures proposées par des algorithmes, et de déterminer laquelle est la meilleure selon un critère fixé.*

5.1 Introduction et principe de sélection

Alors que nous travaillions en classification supervisée lors de l'introduction d'options de rejet dans la section précédente, nous nous intéressons maintenant à la classification non supervisée (ou *clustering*). Parmi les méthodes permettant de trouver des groupes de points appartenant au même cluster sur la base d'une similarité entre observations, on trouve les approches strictes (*c-means*), et les approches floues (par exemple *fuzzy c-means*), où chaque observation est associée à chaque cluster par le biais de degrés d'appartenance. Un défaut récurrent de ce genre de méthodes est leur non robustesse aux points isolés, aux points de bruit, ainsi qu'aux clusters qui se chevauchent. Dans cette section, nous travaillerons avec la méthode FCM introduite dans [Bezdek, 1981] et décrite en section 3.3.2.3. La principale limitation des méthodes de clustering est que l'utilisateur doit fixer le nombre c de clusters, nombre qui ne lui est pas forcément connu. Une c-partition floue U obtenue grâce à l'algorithme FCM doit donc être validée, puisque sa qualité dépendra beaucoup de ce paramètre c.

Trouver une valeur optimale de c est un problème difficile généralement appelé *Validation de Partition* (*Cluster Validity*) dans la littérature. Le principe général est de produire plusieurs partitions floues avec différentes valeurs de c, et de comparer les partitions obtenues grâce à un indice permettant de sélectionner la meilleure d'entre elles. De nombreux indices de validité de partitions ont été proposés pour cela dans les trente dernières années, et cela continue encore aujourd'hui, voir des états de l'art [Pal and Bezdek, 1995; Wang and Zhang, 2007].
Ces indices peuvent être classés selon le type d'information qu'ils utilisent, en particulier

nous distinguons les différents types donnés dans [Bezdek et al., 1999b], voir Tab. 5.1. Comme la plupart des indices proposés aujourd'hui, nous nous intéressons aux indices s'ap-

Type d'indice	Variables utilisées	Restriction sur U
Direct	U	Partition U stricte
Direct paramétrique	U et V	Partition U stricte
Direct paramétrique et données	U, V et X	Partition U stricte
Indirect	U	Partition U stricte, floue ou possibiliste
Indirect paramétrique	U et V	Partition U stricte, floue ou possibiliste
Indirect paramétrique et données	U, V et X	Partition U stricte, floue ou possibiliste

Tab. 5.1: Une classification des indices de validité de partitions.

pliquant sur des matrices de partition floues, c'est à dire que nous considérons seulement les indices indirect dans nos propositions.

On peut également classer les indices selon les propriétés des clusters considérées : certains indices évaluent la compacité de chaque cluster, d'autres la séparation entre clusters, certains les associent. Notons que ces catégories ne sont pas mutuellement exclusives, et que la plupart des indices proposés présentent plusieurs de ces propriétés mais pas nécessairement toutes, ce qui implique des avantages et des inconvénients pour chacun d'entre eux. Les premiers indices n'utilisent que les degrés d'appartenance de la matrice U, ils sont donc directs ou indirects. Ils présentent l'avantage d'être faciles à calculer, bien adaptés aux situations où les clusters se chevauchent, mais souffrent d'une tendance monotone par rapport à c. Un autre problème fréquemment cité est que ce type d'indices n'a pas de lien avec la structure géométrique des données. On peut cependant remarquer que la matrice U est construite à partir de l'information géométrique de V et X. Les indices plus récents utilisent des mesures de compacité et de séparation utilisant de manière conjointe la matrice de partition U et les informations spatiales contenues dans V. Ces indices sont moins monotones vis à vis de c, mais sont plus complexes à calculer, et pas forcément plus performants dans le cas de clusters qui se chevauchent. De plus, la façon dont les mesures de compacité et de séparation sont calculées ne permet pas de distinguer de nombreuses situations pourtant bien distinctes en réalité, voir [Kim et al., 2004] pour des exemples.

Les problèmes des indices existants, quelle que soit la catégorie à laquelle ils appartiennent, sont d'autant plus importants que les algorithmes de clustering utilisés ne sont pas capables de gérer les points de bruit et les points isolés. Ainsi, si l'on construit un pont entre des clusters, les points de bruit rendront le chevauchement des deux clusters plus important, réduisant ainsi le nombre de clusters trouvé, voir Fig. 5.1–(*gauche*). À l'inverse, les points isolés peuvent introduire des clusters composés d'un unique point, puisque ce cluster sera très compact et bien séparé des autres points. Cela augmente de manière erronée le nombre de clusters trouvé, voir Fig. 5.1–(*droite*).

La validation d'une partition (U,V) de X consiste à déterminer si celle-ci reflète ou non la structure des données. Comme l'utilisateur n'a aucune connaissance a priori sur

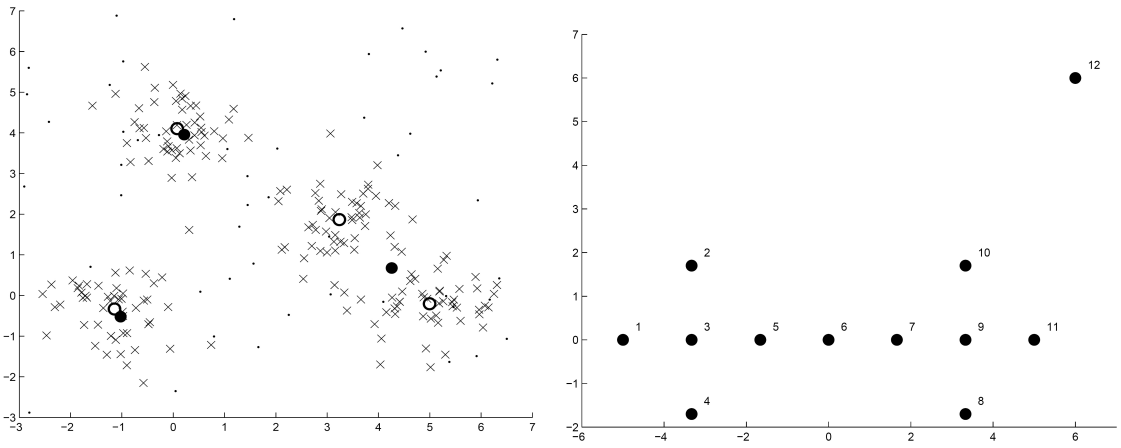

FIG. 5.1: Points de bruits additionnels (·) et centres d'une partition où $c = 3$ (•) au lieu de $c = 4$ (○), (*gauche*) - Point aberrant additionnel (12) augmentant le nombre de clusters de 2 à 3, (*droite*).

les données en classification non supervisée, les algorithmes de clustering tels que FCM utilisent un paramètre c (parmi d'autres) pour la spécification du nombre de clusters.

À partir d'un indice de validité de partition IVP, la procédure de sélection du nombre optimal $c^\star$ de clusters dans un ensemble de valeur possibles $[c_{min},c_{max}]$ est la suivante :

(1) choix des valeurs c_{min} et c_{max}
(2) pour $c = c_{min}$ à c_{max}
- calcul de (U,V) par un algorithme de clustering, par exemple FCM
- calcul de $IVP(c)$ à partir de (X,U,V)
(3) sélection de $c^\star$ tel que $IVP(c^\star)$ est optimal, et prendre la partition correspondante (U,V)

Le choix des valeurs c_{min} et c_{max} est toujours sujet à caution. Pour la valeur de c_{min}, on suppose que l'étape visant à déterminer si les données présentent une structure s'est conclue de manière positive, et que au moins deux groupes sont donc présents, on prend ainsi $c_{min} = 2$. Pour la valeur de c_{max}, plusieurs bornes supérieures ont été proposées, mais toutes de manière empirique. Nous prendrons l'une d'entre elles, utilisée le plus fréquemment; $c_{max} = \sqrt{n}$. Cette borne correspond en fait à la volonté de former des groupes constitués en moyenne de $\sqrt{n}$ observations.

5.2 Indices usuels

Dans cette section, après avoir rappelé quelques points importants sur l'algorithme FCM, nous considérons les indices de validité de partition s'y rapportant.
L'algorithme FCM produit une c-partition floue, c'est à dire une matrice $U = [\mathbf{u}_1,...,\mathbf{u}_n]$ de taille $(c \times n)$, où $\mathbf{u}_k$ est le vecteur de degrés d'appartenance de $\mathbf{x}_k$ aux c clusters. Pour rappel, cette matrice de partition floue U est obtenue par la minimisation de la fonctionnelle

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m ||\mathbf{x}_k - \mathbf{v}_i||^2 \tag{5.1}$$

On notera les restrictions suivantes sur cette matrice U, u_{ik} dénotant le degré d'appartenance de $\mathbf{x}_k$ au cluster i :

- $\sum_{i=1}^{c} u_{ik} = 1$
- $0 < \sum_{k=1}^{n} u_{ik} < n$

Une seconde matrice $V = [\mathbf{v}_1,...,\mathbf{v}_c]$ (de taille $(c \times p)$) de centres de clusters est également produite par FCM. Enfin, il y a un exposant flou $m \in]1, +\infty[$ rendant les partitions produites plus ou moins floues. Ce dernier paramètre a également une influence sur les indices de validité de partition. Comme la métrique utilisée dans FCM est la distance euclidienne, les clusters obtenus sont hyper-sphériques et contiennent un nombre de points similaire. Cette description des clusters n'est pas vraiment adaptée à de nombreuses situations : ponts, points isolés, points de bruit. La validation de partition est donc un problème encore plus difficile lorsque l'on utilise FCM au lieu d'autres algorithmes dédiés à ces problèmes spécifiques, par exemple PCM.

Beaucoup d'indices ont été proposés ces dernières années (voir [Rezaee et al., 1998; Wang and Zhang, 2007; Zarandi et al., 2007]), et les propositions ne cessent d'augmenter, si bien qu'il n'est pas possible de tous les décrire. Nous présentons ici plusieurs d'entre eux, parmi lesquels on trouve les plus fréquemment cités dans la littérature.

5.2.1 Indices indirects

Comme précisé dans l'introduction, les premiers indices n'utilisent que les degrés d'appartenance de la matrice U. Les deux plus importants sont le *Coefficient de Partition* [Bezdek, 1981] à valeurs dans $[\frac{1}{c},1]$:

$$PC(c) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^2 \tag{5.2}$$

et l'*Entropie de Classification* [Bezdek, 1974] à valeur dans $[0, \log(c)]$:

$$PE(c) = -\frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log(u_{ik}) \tag{5.3}$$

Plus PC est proche de 1, plus le clustering est strict, et inversement plus PC est faible, plus toutes les valeurs u_{ik} pour $i = 1, \cdots, c$ sont proches de $1/c$. Comme de grandes valeurs des u_{ik} suggèrent des clusters compacts et séparés, on cherche à maximiser PC afin de trouver $c^\star$.

Le second indice PE a une valeur faible si la partition examinée est stricte, se rapproche de $\log(c)$ lorsque la partition U se rapproche de la partition la plus floue possible $\overline{U} = [1/c]$. Ici, PE est donc à minimiser lors de la recherche de $c^\star$. Remarquons que cet indice est l'extension directe de l'entropie au sens de De Luca et Termini [De Luca and Termini, 1972] abordée au Chapitre 2. Un défaut de ces deux indices est qu'ils sont monotones par rapport à c, ainsi que leurs bornes. Un autre problème est leur sensibilité au paramètre m, puisque si celui ci tend vers 1, PC et PE ont les mêmes valeurs quel que soit c, et inversement, lorsque $m \to \infty$, la valeur $c^\star = 2$ est systématiquement choisie, voir [Pal and

Bezdek, 1995]. Afin de contrer cette tendance, Roubens propose une version normalisée de l'indice PC, [Roubens, 1978] :

$$NPC(c) = \frac{cPC - 1}{c - 1} \tag{5.4}$$

de manière à ce que NPC soit à valeurs dans $[0, 1]$. Dans Dunn [1977], Dunn modifie l'Entropie de Partition de la façon suivante :

$$NPE(c) = \frac{nPE}{n - c} \tag{5.5}$$

Nous utiliserons ces versions normalisées NPC et NPE pour la partie expérimentale.

5.2.2 Indices indirects paramétriques

La plupart des indices récents reposent sur l'utilisation de mesures de compacité définies à l'aide de U ou (U,X), et d'une mesure de séparation généralement fondée sur les centres V.

Xie and Beni [Xie and Beni, 1991] proposent un indice de validité pour $m = 2$ défini par :

$$XB(c) = \frac{J_m(U,V)\,/n}{\min_{i,j=1,c;j\neq i}\ ||\mathbf{v}_i - \mathbf{v}_j||^2} \tag{5.6}$$

où $J_m(U,V)$ est utilisée comme mesure de compacité et $\min_{i,j=1,c;j\neq i}\ ||\mathbf{v}_i - \mathbf{v}_j||^2$ est une mesure de séparation. Le nombre optimal de clusters $c^\star$ est trouvé par maximisation de l'équation (5.6). En suivant une idée similaire, Fukuyama et Sugeno [Fukuyama and Sugeno, 1989] proposent l'indice utilisant la fonctionnelle (3.41) comme mesure de compacité, et une mesure de séparation afin de la pénaliser :

$$FS(c) = J_m(U,V) - \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m\,||\mathbf{v}_i - \overline{\mathbf{v}}||^2 \tag{5.7}$$

où $\overline{\mathbf{v}}$ est la moyenne des centres $\mathbf{v}$. Le premier terme mesure la compacité et le second la séparation. La minimisation de (5.7) donnera le nombre optimal de clusters de X. Gath et Geva proposent l'Hypervolume Flou (FHV) dans [Gath and Geva, 1989] :

$$FHV(c) = \sum_{i=1}^{c}\sqrt{\det(C_i)} \tag{5.8}$$

où C_i est la matrice de covariance floue du cluster i définie par :

$$C_i = \frac{\sum_{k=1}^{n} u_{ik}^m(\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T}{\sum_{k=1}^{n} u_{ik}^m} \tag{5.9}$$

Une faible valeur de FHV indique que les clusters sont compacts, on minimise donc (5.8). Un autre indice, assez similaire à XB, a été introduit dans [Bensaid et al., 1996]. Il est défini comme le ratio d'une mesure de compacité et d'une mesure de séparation, mais utilise une norme différente lors du calcul de la distance entre points : une matrice A est introduite pour définir la déviation floue du point $\mathbf{x}$. L'indice SC est alors défini par

$$SC(c) = \sum_{i=1}^{c}\frac{\sum_{k=1}^{n} u_{ik}^m||\mathbf{x}_k - \mathbf{v}_i||_A^2}{n_i\sum_{j=1}^{c}||\mathbf{v}_i - \mathbf{v}_j||_A^2} \tag{5.10}$$

où n_i est la cardinalité floue du cluster i définie par $n_i = \sum_{k=1}^{n} u_{ik}$.
Afin de réduire la tendance monotone de (5.6) lorsque c tend vers n, Kwon étend l'indice XB dans [Kwon, 1998]. Un terme de pénalité défini par un degré de séparation entre les clusters est ajouté au numérateur de XB :

$$K(c) = \frac{J_m(U,V) + \frac{1}{c}\sum_{i=1}^{c} ||\mathbf{v}_i - \overline{\mathbf{v}}||^2}{\min_{i,j=1,c;j\neq i} \; ||\mathbf{v}_i - \mathbf{v}_j||^2} \tag{5.11}$$

où $\overline{\mathbf{v}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$. Comme XB, K a une tendance à sous-estimer le nombre de clusters. Cet indice a une valeur faible lorsque les clusters sont compacts et séparés, et est donc à minimiser. Wu et Yang [Wu and Yang, 2005] proposent un indice de validité défini par :

$$WY(c) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^2/u_M - \sum_{i=1}^{c} \exp\left(-\min_{j\neq i}\left(||\mathbf{v}_i - \mathbf{v}_j||^2/\beta_T\right)\right) \tag{5.12}$$

où $u_M = \min_{1\leq i\leq c}\left(\sum_{k=1}^{n} u_{ik}^2\right)$ et $\beta_T = \sum_{i=1}^{c} ||\mathbf{v}_i - \overline{\mathbf{v}}||^2/c$. Une grande valeur de WY signifie que les c clusters sont compacts et séparés. La maximisation de (5.12) donne le nombre optimal de clusters. Par cette formulation, un point isolé n'aura pas assez d'importance pour former un cluster, impliquant que WY donne de bons résultats en présence de points isolés.
Pakhira et al. proposent un nouvel indice pour des algorithmes de clustering stricts et flous dans [Pakhira et al., 2004]. L'indice PBM pour un clustering flou est défini par :

$$PBM(c) = \left(\frac{1}{c} \times \frac{E_1}{J_m(U,V)} \times D_c\right)^2 \tag{5.13}$$

où $E_1 = \sum_{k=1}^{n} ||\mathbf{x}_k - \overline{\mathbf{v}}||$, qui est constant pour des données X, et $D_c = \max_{i,j=1}^{c} ||\mathbf{v}_j - \mathbf{v}_i||$ est la séparation inter-clusters maximale. Le premier terme diminue lorsque c augmente. Le terme $\frac{E_1}{J_m(U,V)}$ est la somme des distances intra-clusters pondérées lorsque l'ensemble X est pris comme un seul cluster, divisée par la fonctionnelle (3.41). Le nombre optimal de clusters est donc obtenu par maximisation de (5.13).

En résumé, lorsque l'on cherche à valider une partition, on peut se concentrer sur plusieurs problèmes différents. Les premiers indices (PC, PE) cherchent une partition minimisant le caractère flou des degrés d'appartenance la composant, par exemple l'entropie ou la dispersion. Les indices plus récents se fondent sur une analyse de la compacité ou de la séparation des clusters produites par la partition, et à ce titre utilisent les informations supplémentaires fournies par l'algorithme : la matrice de centres V. Certains indices vont jusqu'à ne pas tenir compte de la matrice U autrement que par l'utilisation de la fonctionnelle (3.41). D'un certain point de vue, la partition évaluée est donc V et non pas U. Bien que performants dans certaines situations pour lesquelles un terme de pénalité particulier est introduit, ces indices ne possèdent pas une capacité de généralisation importante : une situation non prévue aboutira à un échec.

5.3 Un nouvel indice indirect

Si l'on suppose que les valeurs u_{ik} d'une matrice de partition floue représentent le degré de satisfaction au $i^{\text{ème}}$ groupe de $\mathbf{x}_k$, c'est à dire sa similarité aux prototypes de chaque

cluster, nous nous intéressons alors valeurs du vecteur $\mathbf{u}_k = [u_{1k}, \cdots, u_{ck}]$. À partir des informations contenues dans ce vecteur, le clustering consiste à sélectionner le groupe le plus approprié pour le point $\mathbf{x}_k$. L'opérateur max, ou $\perp_M$, est souvent utilisé dans cette situation, mais les valeurs inférieures interagissent également dans la décision à prendre. Ainsi l'opérateur max ne montre pas un comportement de compensation des valeurs agrégées, tandis que d'autres t-conormes triangulaires, particulièrement les archimédiennes [1], ont cette propriété intéressante, voir [Klement and Mesiar, 2005]. Cette propriété est en effet très intéressante, en particulier lorsque une forme satisfait plusieurs groupes : une partition exclusive n'est pas performante. Un point fondamental est donc la détermination du degré global d'appartenance exclusive à un groupe (ou cluster).

Dans Mascarilla et al. [2008], les auteurs définissent l'opérateur OU flou d'ordre l et l'utilisent dans le cadre de la classification supervisée avec options de rejet (voir section précédente). Cet opérateur évalue les degrés de satisfaction à un ordre donné par combinaison de normes triangulaires. Nous rappelons sa définition ici par commodité de lecture, mais une définition plus complète se trouve au Chapitre 1 :

$$\underset{i=1,\,c}{\overset{l}{\perp}} u_i = \underset{A\in\mathcal{P}_{l-1}}{\top}\left(\underset{j\in C\setminus A}{\perp} u_j\right) \tag{5.14}$$

Afin de rester cohérent avec les notations que l'on a utilisé dans ce chapitre, nous noterons cet opérateur $\Phi_{l,\top}$, où l est l'ordre considéré, et $\top$ correspond au couple $(\top, \perp)$ de normes triangulaires. Nous allons utiliser la capacité de cette mesure à évaluer les degrés d'appartenance pour des ordres différents afin de trouver le nombre optimal de cluster de données.

Un indice de validité fiable pour l'algorithme FCM doit considérer d'une part la compacité, mais aussi la séparation des données, observables à travers de la partition floue. Si seule la compacité est considérée, la meilleure partition sera obtenue lorsque chaque point est vu comme un cluster à part entière. D'un autre côté, si seule la séparation est prise en compte, la meilleure partition correspondra à un unique cluster contenant l'ensemble des données. Il est donc clair qu'un indice doit comporter d'une manière ou d'une autre ces deux types de mesures. Comme nous l'avons précisé, les mesures usuelles de compacité et de séparation ne reflètent pas la structure réelle des données. De nombreux indices utilisent la fonctionnelle (3.41) afin de quantifier la compacité des partitions floues. Pourtant, comme il est montré dans [Pal and Bezdek, 1995; Kwon, 1998], ces indices ont une tendance à décroitre de manière monotone lorsque le nombre de clusters se rapproche du nombre de points de X, n, en d'autres termes, $\lim_{c\to n} ||\mathbf{x}_k - \mathbf{v}_i||^2 = 0$. De plus cette fonctionnelle est déjà minimisée par l'algorithme FCM, et sa réutilisation ne permettra pas de corriger, ou de rattraper, certaines erreurs.
Les mesures de séparation sont généralement calculées par le biais d'une distance entre centres : $||\mathbf{v}_j - \mathbf{v}_i||^2$ pour $j \neq i$. L'utilisation exclusive de l'information des centres n'est pas suffisante pour l'interprétation de la structure géométrique des données, et donc la séparation des clusters, voir [Kim et al., 2004] pour des exemples sur les limitations de ces approches. Il est nécessaire d'inclure des informations supplémentaires pour appréhender

1. Dans ce cas, $\perp(a,a) > a$

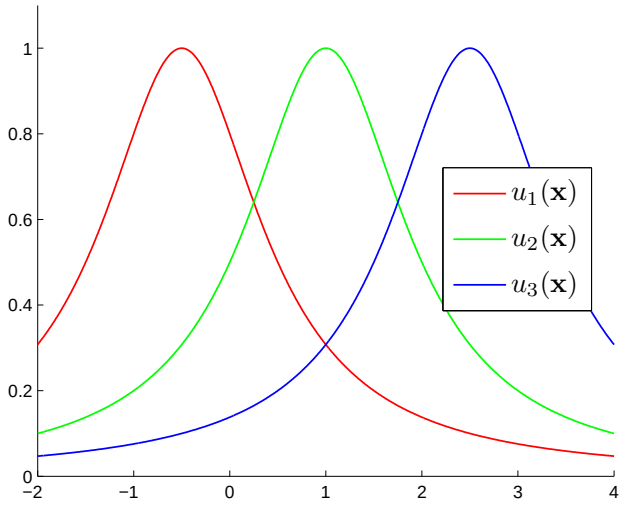

FIG. 5.2: Degrés d'appartenance à trois clusters différents.

de manière correcte la structure géométrique des données, mais cela se fait au prix d'une complexité importante aussi bien au niveau du temps de calcul que de la mise en place algorithmique de l'indice.

Nous proposons donc, pour chaque point $\mathbf{x}_k$, de définir deux mesures ne présentant pas ces inconvénients : une mesure de chevauchement flou évaluant le degré de chevauchement entre un nombre l spécifié de clusters flous, et une mesure de séparation floue indiquant le degré de chevauchement du cluster le plus probable, c'est à dire celui correspondant au degré d'appartenance maximum, par rapport aux $c-1$ autres clusters (grâce à la contrainte de somme à 1 sur l'ensemble des classes). Une faible valeur de séparation floue indiquera une séparation faible du cluster le plus probable par rapport aux autres.

5.3.1 Mesure de chevauchement

La capacité de gérer les clusters se chevauchant est aujourd'hui considérée comme un critère majeur dans la définition d'un indice de validité [Bouguessa et al., 2006]. En dépit de son importance, la majorité des mesures de chevauchement sont fondées sur des représentations intuitives. Une mesure de chevauchement entre l clusters flous pour chaque point $\mathbf{x}_k$ de X décrit par ses degrés d'appartenance peut être obtenue grâce à (5.14), comme c'est illustré en FIG. 5.2 – 5.3 – 5.4. En FIG. 5.2, les degrés d'appartenance à trois clusters sont indiqués par trois courbes de couleurs différentes. Dans les deux figures suivantes, les chevauchements entre $l = 2$ clusters (FIG. 5.3) et $l = 3$ clusters (FIG. 5.4) sont donnés pour différents couples de normes triangulaires. On remarque que la mesure de chevauchement d'ordre l est nulle lorsque $l - 1$ clusters se chevauchent, et augmente au fur et à mesure que les clusters se chevauchent de plus en plus. L'utilisation de différents couples permet de mettre en avant certaines caractéristiques du chevauchement : on distingue en particulier les mesures pour lesquelles on observe un pic ou un creux en $\mathbf{x} = 1$, c'est à dire au point où $u_2(\mathbf{x})$ est maximum, et où le chevauchement de $u_1(\mathbf{x})$ et $u_3(\mathbf{x})$ est maximum.

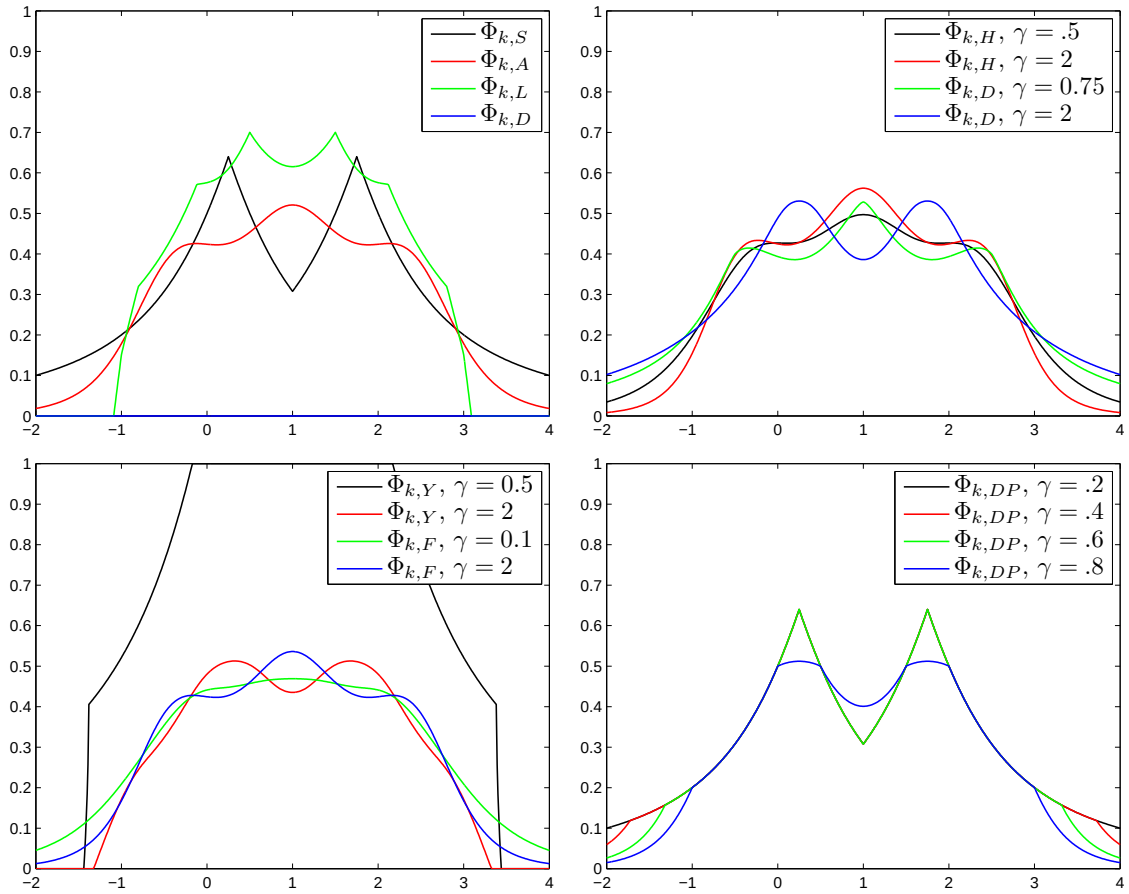

FIG. 5.3: Ou Flou d'ordre $l = 2$ pour les trois fonctions d'appartenance, et pour différents couples $(\top, \perp)$.

En calculant successivement $\Phi_{l,\top}$ pour différentes valeurs de l, nous obtenons une série de degrés de chevauchement d'ordre l pour $\mathbf{x}_k$. Afin de déterminer le degré global de chevauchement pour un point donné, il faut ensuite trouver lequel de ces ordres implique le plus grand chevauchement. L'(Les) ordre(s) le(s) plus satisfait(s) est(sont) obtenu(s) par une disjonction floue (voir CHAPITRE 1) des mesures de chevauchement d'ordre l pour $l = 2, \cdots, c$, et nous définissons la mesure de chevauchement comme :

$$O_{\perp}(\mathbf{x}_k, c) = \underset{l=2,c}{\overset{1}{\perp}} \Big(\Phi_{l,\top}(\mathbf{u}(\mathbf{x}_k)) \Big) \tag{5.15}$$

où $\perp$ est la t-conorme triangulaire duale de $\top$. Notons que plusieurs indices utilisent des mesures de chevauchement (par exemple [Kim et al., 2004; Zarandi et al., 2007]), mais ces mesures sont seulement calculées sur tous les couples possibles de clusters, et non sur tous les sous-ensembles comme c'est le cas implicitement ici. De telles approches correspondent en fait aux mesures de chevauchement d'ordre 2, et il faudrait donc $c \times (c-1)$ opérations supplémentaires, ce qui demanderait un temps de calcul non négligeable.

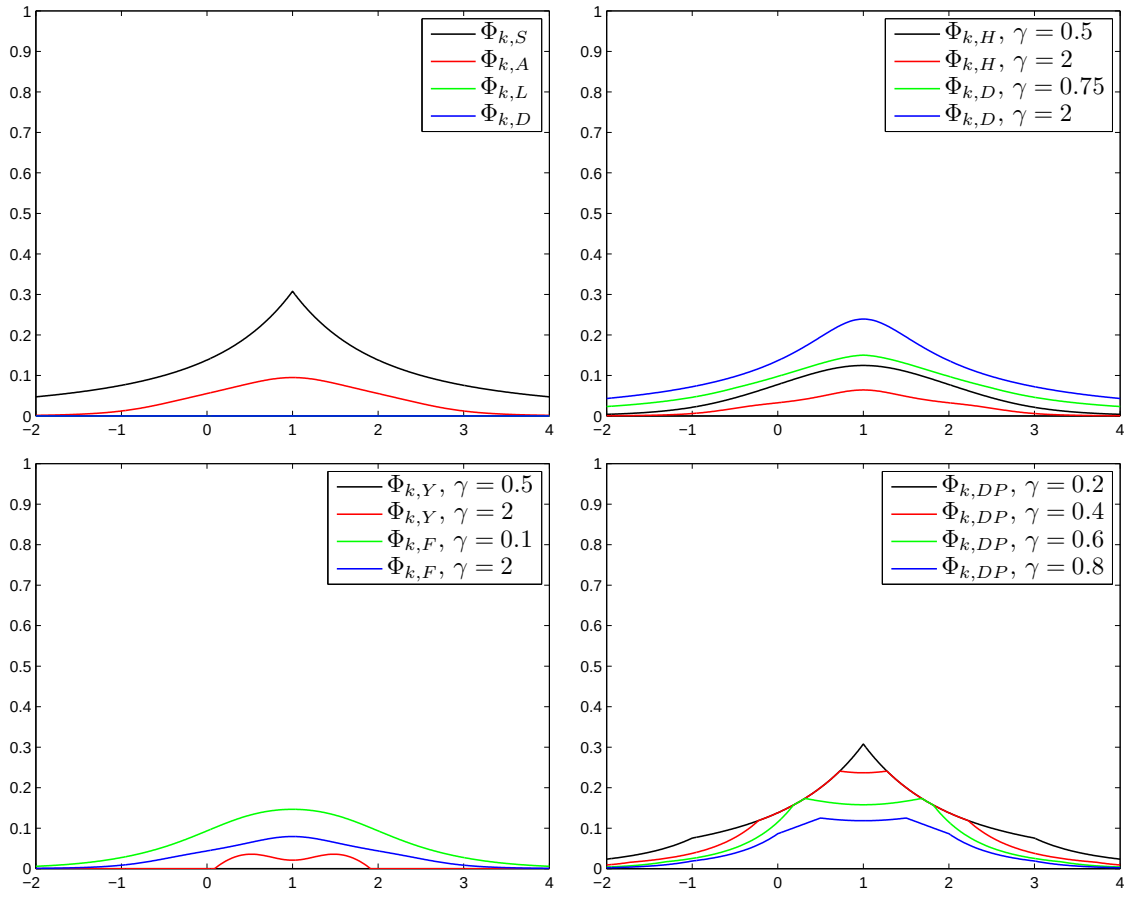

FIG. 5.4: Ou Flou d'ordre $l = 3$ pour les trois fonctions d'appartenance, et pour différents couples $(\top, \perp)$.

5.3.2 Mesure de séparation

Dans [Bezdek and Pal, 1998], Bezdek et Pal montrent que la séparation entre les clusters joue un rôle plus important que le diamètre des clusters en validation de partition. Nous proposons d'introduire une telle mesure en quantifiant la séparation floue de chaque point $\mathbf{x}_k$ par l'utilisation de $\perp \mathbf{u}_k$, qui est en fait la mesure de chevauchement d'un cluster flou, c'est à dire sa séparation des autres clusters, puisque $\sum_{i=1}^{c} u_{ik} = 1$. Cette agrégation correspond à la disjonction floue des degrés d'appartenance pour le point $\mathbf{x}_k$, ce qui sélectionne le cluster le plus probable. Les mesures de chevauchement sont donc combinées à l'aide d'un opérateur disjonctif, par exemple $\perp$.

Nous définissons la séparation floue de $\mathbf{x}_k$ par rapport aux c clusters comme :

$$S_{\perp}(\mathbf{x}_k, c) = \overset{1}{\perp}\left(\underbrace{\Phi_{1,\top}(\mathbf{u}(\mathbf{x}_k)), \cdots, \Phi_{1,\top}(\mathbf{u}(\mathbf{x}_k))}_{c-1 \text{ fois}}\right) \tag{5.16}$$

Nous donnons en FIG. 5.5 les valeurs de $\Phi_{1,\top}$ appliquée aux fonctions d'appartenance de la FIG. 5.2 pour différents couples de normes triangulaires. Cette séparation est maximum au pic des degrés d'appartenance de $u_1(\mathbf{x})$, $u_2(\mathbf{x})$ et $u_3(\mathbf{x})$, et décroît lorsque l'on s'en éloigne.

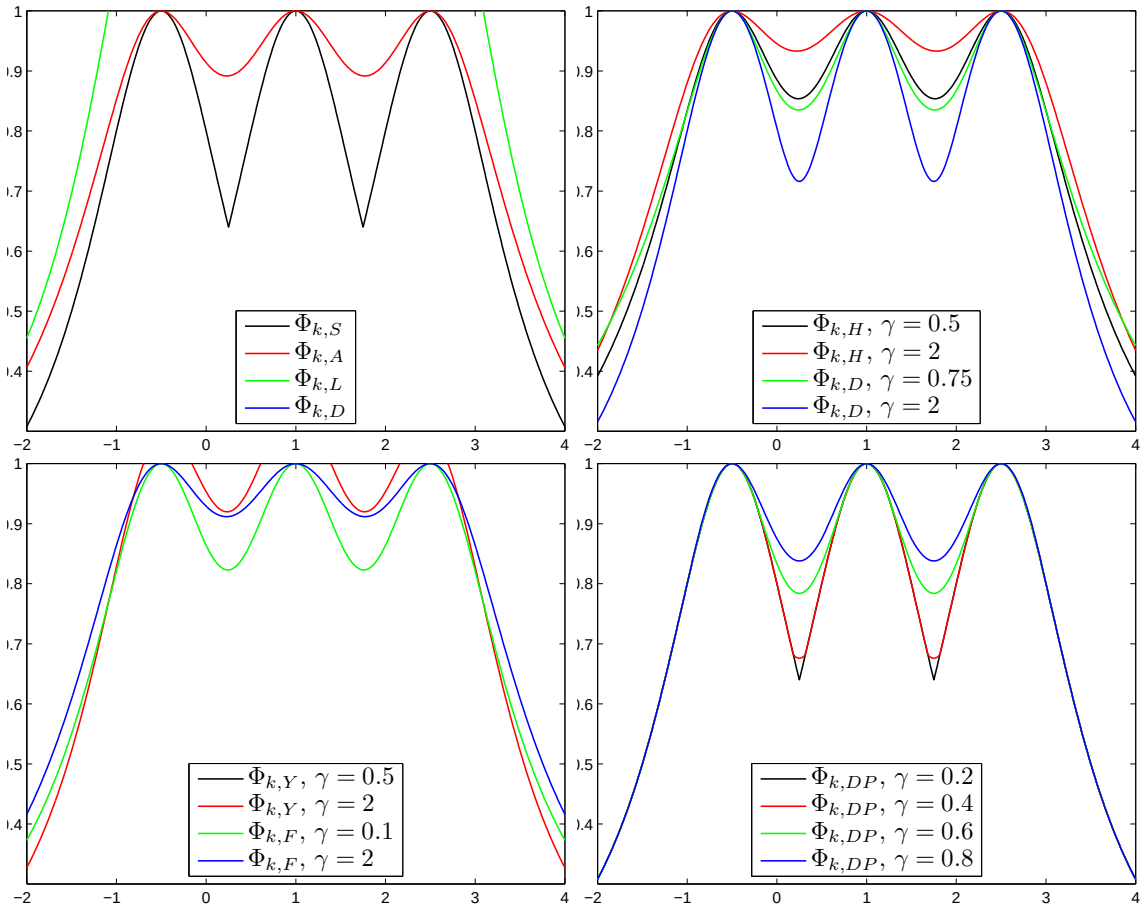

FIG. 5.5: Ou Flou d'ordre $k = 1$ pour les trois fonctions d'appartenance, et pour différents couples $(\top, \perp)$.

5.3.3 Une famille d'indices et ses propriétés

Une faible valeur de la mesure de chevauchement $O_\perp(\mathbf{x}_k,c)$ et une grande valeur de la mesure de séparation $S_\perp(\mathbf{x}_k,c)$ indiquent que le plus grand des degrés d'appartenance de $\mathbf{x}_k$ correspond à un cluster bien séparé ne présentant pas de chevauchement.

Définition 5.1 ([Le Capitaine and Frélicot, 2009c])**.** Nous proposons donc l'indice d'ordre k de chevauchement et séparation (*Overlap and Separation Index* - *OSI*) à valeurs dans [0,1] comme la moyenne des ratios des deux mesures $O_\perp(\mathbf{x}_k,c)$ et $S_\perp(\mathbf{x}_k,c)$ sur X :

$$OSI_\perp(c) = \frac{1}{n}\sum_{k=1}^{n}\frac{O_\perp(\mathbf{x}_k,c)}{S_\perp(\mathbf{x}_k,c)} \tag{5.17}$$

La minimisation de (5.17) sur $[c_{min},c_{max}]$ donne le nombre optimal de clusters pour les données X.

Remarque 5.1. Nous utilisons, dans la définition de (5.17), l'opérateur d'agrégation moyenne. Ce choix correspond à la quasi totalité des indices de validité de partition proposés dans la littérature. Une exception cependant, dans [Wu et al., 2009], les auteurs proposent une modification des indices les plus connus (*PC*, *PE*, *XB*) en utilisant l'opérateur médian à la place de la moyenne. Ceci peut tout à fait être fait pour l'indice *OSI*. Une étude plus approfondie pourra d'ailleurs être menée pour déterminer si d'autres opérateurs d'agrégation

seraient plus appropriés pour cette application précise.

Nous avions déjà proposé d'utiliser la moyenne du rapport d'une mesure de chevauchement et d'une mesure de séparation fondées sur le OU flou d'ordre l, [Le Capitaine and Frélicot, 2008a]. Dans la proposition antérieure, la mesure de séparation étant définie par

$$S_{\perp}(\mathbf{x}_k, c) = \Phi_{1,\top}(\mathbf{u}(\mathbf{x}_k)) \tag{5.18}$$

pouvait être inférieure à la mesure de chevauchement selon le couple ($\top$,$\perp$) utilisé, impliquant que la valeur de l'indice n'était pas bornée. La première conséquence est qu'il n'est pas garanti que l'indice ne soit pas monotone vis-à-vis de c et de n, à l'instar des autres indices utilisant uniquement U comme PC ou PE. La deuxième conséquence est que puisque le rapport est inférieur à 1 pour la plupart des points, l'influence d'une grande valeur de ce rapport affecte de manière très significative la valeur moyenne, de sorte que l'indice pourrait ne pas refléter correctement la structure des données. Le nouvel indice (5.17) ne présente pas ces défauts.

Dans la suite de cette section, nous allons montrer quelques propriétés satisfaites par l'indice proposé. Pour l'ensemble de ces propriétés, c est supposé être supérieur ou égal à 2.

Proposition 5.1 ([Le Capitaine and Frélicot, 2009c])**.** *Pour tout les couples* ($\top$,$\perp$) *de normes triangulaires, on a* $0 \leq OSI_{\perp}(c) \leq 1$.

Démonstration. Il est prouvé dans [Mascarilla et al., 2008] que $\overset{1}{\perp} \geq \overset{2}{\perp} \cdots \geq \overset{c}{\perp}$, pour tout ($\top$,$\perp$). Soit $\mathbf{o}$ le vecteur de dimension $(c-1)$ construit par $\Phi_{l,\top}(\mathbf{u})$, pour k variant de 2 à c, et $\mathbf{s}$ le vecteur de même dimension construit par les $(c-1)$ valeurs de $\Phi_{1,\top}(\mathbf{u})$. On a alors $o_i \leq s_i$ pour tout $i \in \{1, \cdots, c-1\}$. Par monotonie de (5.14), on a finalement $O_{\perp}(\mathbf{x}_k,c) \leq S_{\perp}(\mathbf{x}_k,c)$ pour tout $k \in \{1, \cdots, n\}$, d'où $O_{\perp}(\mathbf{x}_k,c)/S_{\perp}(\mathbf{x}_k,c) \leq 1$. Comme $O_{\perp}(\mathbf{x}_k,c)$ et $S_{\perp}(\mathbf{x}_k,c)$ appartiennent à $[0,1]$, on a également $O_{\perp}(\mathbf{x}_k,c)/S_{\perp}(\mathbf{x}_k,c) \geq 0$, ce qui conclut la preuve. □

Proposition 5.2 ([Le Capitaine and Frélicot, 2009c])**.** *Si U est une matrice de partition stricte, alors $OSI_{\perp}(c) = 0$, pour tout couple* ($\top$,$\perp$).

Démonstration. Puisque U est une matrice de partition stricte, alors pour tout k, $u_{ik} \in \{0,1\}$ et $\sum_{i=1}^{c} u_{ik} = 1$. Ceci a pour conséquence qu'une valeur vaut 1, et que toutes les autres valent 0. Prenons par exemple $u_{(1)k} = 1$ et $u_{(2)k} = \cdots = u_{(c)k} = 0$. Comme 0 est l'élément absorbant de $\top$ [2], on montre facilement que $O_{\perp}(\mathbf{x}_k,c) = 0$ pour tout ($\top$,$\perp$) :

$$\begin{aligned} \overset{1}{\underset{l=2,c}{\perp}} \left(\overset{l}{\underset{i=1,c}{\perp}} u_{ik} \right) &= \left(\overset{2}{\underset{i=1,c}{\perp}} (1,0,\cdots,0) \right) \overset{1}{\perp} \cdots \overset{1}{\perp} \left(\overset{c}{\underset{i=1,c}{\perp}} (1,0,\cdots,0) \right) \\ &= \overset{1}{\perp} (\underbrace{0,\cdots,0}_{c-1 \text{ fois}}) = 0. \end{aligned} \tag{5.19}$$

2. $a \top 0 = 0$, $\forall\ \top$

Enfin, comme 1 est l'élément absorbant de $\perp$[3], alors $\overset{1}{\perp}(1,0,\cdots,0) = 1$. Nous avons donc $OSI_{\perp}(c) = \frac{1}{n}\sum_{k=1}^{n} 0/1 = 0$, terminant la démonstration. □

Proposition 5.3 ([Le Capitaine and Frélicot, 2009c]). *Si $U = \overline{U}$ est une partition complètement floue, alors $OSI_{\perp}(c) = 1$ si on prend $(\top, \perp)_M$.*

Démonstration. Puisque U est une matrice de partition complètement floue, alors pour tout k, $u_{ik} = \frac{1}{c}$. À partir de (5.15), on peut écrire $O_{\perp}(\mathbf{x}_k,c)$ comme:

$$\underset{l=2,c}{\overset{1}{\perp}}\left(\underset{i=1,c}{\overset{l}{\perp}} u_{ik}\right) = \left(\underset{i=1,c}{\overset{2}{\perp}}\left(\frac{1}{c},\cdots,\frac{1}{c}\right)\right) \overset{1}{\perp} \cdots \overset{1}{\perp} \left(\underset{i=1,c}{\overset{c}{\perp}}\left(\frac{1}{c},\cdots,\frac{1}{c}\right)\right) \tag{5.20}$$

Si l'on utilise les normes triangulaires standard ($\top = \min, \perp = \max$), on peut simplifier (5.20) de la façon suivante :

$$\begin{aligned}
\underset{l=2,c}{\overset{1}{\perp}}\left(\underset{i=1,c}{\overset{l}{\perp}} u_{ik}\right) &= \overset{1}{\perp}\left(\underbrace{\frac{1}{c},\cdots,\frac{1}{c}}_{c-1 \text{ fois}}\right) \\
&= \max\left(\underbrace{\frac{1}{c},\cdots,\frac{1}{c}}_{c-1 \text{ fois}}\right) \\
&= \frac{1}{c}
\end{aligned} \tag{5.21}$$

Il est facile de vérifier que c'est la valeur de $S_{\perp}(\mathbf{u}_k,c)$ pour tout $\mathbf{u}_k$. Ainsi, nous obtenons $OSI_{\perp}(c) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{c}/\frac{1}{c} = 1$. □

Proposition 5.4 ([Le Capitaine and Frélicot, 2009c]). *Si l'on utilise les normes triangulaires standard min et max, alors $OSI_{\perp}(c) = \frac{1}{n}\sum_{k=1}^{n} \frac{u_{(2)k}}{u_{(1)k}}$.*

Démonstration. On considère le OU flou à l'ordre l. Puisque $\overset{l}{\perp}(\mathbf{u}_k) = u_{(l)k}$ si l'on utilise les normes triangulaires standard min et max, alors on a

$$OSI_{\perp}(c) = \frac{1}{n}\sum_{k=1}^{n} \frac{\underset{l=2,c}{\overset{1}{\perp}}\left(u_{(l)k}\right)}{\overset{1}{\perp} u_{(1)k}} \tag{5.22}$$

Puisque $\overset{1}{\perp}(\mathbf{u}_k)$ est la t-conorme triangulaire $\perp_M$, c'est à dire l'opérateur maximum, on obtient

$$OSI_{\perp}(c) = \frac{1}{n}\sum_{k=1}^{n} \frac{u_{(2)k}}{u_{(1)k}} \tag{5.23}$$

□

3. $a \perp 1 = 1, \forall\, \top$

	Point	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
$c=2$	u_{1k}	.946	.944	.996	.943	.945	.613	.174	.129	.018	.022	.039	.222
	u_{2k}	.054	.056	.004	.057	.055	.387	.826	.871	.982	.978	.961	.778
$c=3$	u_{1k}	.933	.913	.998	.921	.876	.466	.100	.055	.001	.058	.042	.000
	u_{2k}	.047	.063	.001	.061	.100	.477	.874	.905	.998	.824	.866	.001
	u_{3k}	.019	.024	.001	.018	.024	.057	.026	.039	.001	.119	.092	.999

TAB. 5.2: Degrés d'appartenance pour $c = 2{,}3$ clusters, et les données exemples *Diamond+*.

	Point	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
$(\top,\perp)_M$	$O(\mathbf{u}_k,c)$	.054	.056	.004	.057	.055	.387	.174	.129	.018	.022	.039	.222
	$S(\mathbf{u}_k,c)$	.946	.944	.996	.943	.945	.613	.826	.871	.982	.978	.961	.778
	$\frac{O(\mathbf{u}_k,c)}{S(\mathbf{u}_k,c)}$	.057	.059	.004	.060	.058	.631	.210	.148	.018	.022	.040	.285
$(\top,\perp)_P$	$O(\mathbf{u}_k,c)$	.051	.052	.003	.052	.051	.237	.143	.112	.017	.021	.037	.172
	$S(\mathbf{u}_k,c)$	.949	.947	.996	.947	.948	.762	.856	.887	.982	.979	.962	.827
	$\frac{O(\mathbf{u}_k,c)}{S(\mathbf{u}_k,c)}$	.053	.055	.003	.055	.054	.310	.167	.126	.017	.021	.038	.208
$(\top,\perp)_{H_{10}}$	$O(\mathbf{u}_k,c)$	.034	.035	0	.035	.035	.071	.058	.055	.012	.017	.027	.063
	$S(\mathbf{u}_k,c)$	.965	.964	.996	.964	.965	.924	.937	.944	.985	.982	.972	.932
	$\frac{O(\mathbf{u}_k,c)}{S(\mathbf{u}_k,c)}$	.035	.036	0	.036	.036	.076	.061	.058	.012	.017	.027	.067
$(\top,\perp)_{D_2}$	$O(\mathbf{u}_k,c)$	.054	.055	.003	.056	.054	.369	.123	.118	.017	.021	.038	.121
	$S(\mathbf{u}_k,c)$	.946	.944	.996	.943	.945	.630	.826	.871	.982	.978	.961	.778
	$\frac{O(\mathbf{u}_k,c)}{S(\mathbf{u}_k,c)}$	.057	.058	.003	.059	.057	.585	.149	.136	.017	.021	.040	.156

TAB. 5.3: Mesures de chevauchement, de séparation et $OSI_\perp$ pour les données *Diamond+* et $c = 2$ clusters.

Nous allons illustrer la capacité de l'indice proposé à trouver le bon nombre de clusters, et donc la bonne partition sur un exemple utilisant des données synthétiques inspirées de [Masson and Denoeux, 2008]. Ces données, baptisées *Diamond+*, contiennent onze points initialement introduits par Windham [Windham, 1985], plus un point isolé, voir FIG. 5.1-(*droite*). Une partition correcte devrait contenir $c^\star = 2$ clusters : les deux diamants qui se touchent. Les indices considérant uniquement la compacité et la séparation sélectionneront $c = 3$ clusters : les deux diamants et le point isolé.

En TAB. 5.2, nous donnons les degrés d'appartenance à chacun des clusters des douze points obtenus grâce à l'algorithme FCM, pour $c = 2$ et $c = 3$ clusters. Les TAB. 5.3–5.4 donnent les valeurs détaillées des mesures de chevauchement (5.15), des mesures de séparation (5.16), ainsi que les rapports formant l'indice OSI, pour les douze points et $c = 2$, $c = 3$, respectivement. Tout d'abord les normes triangulaires min et max sont utilisées par mesure de simplicité. Les valeurs trouvées sont $OSI_S(2) = 0.132$ et $OSI_S(3) = 0.142$, ce qui montre OSI trouve le bon nombre de clusters $c = 2$, c'est à dire une partition correcte. Comme on pouvait le prévoir, pour $c = 2$, $\mathbf{x}_6$ est le point pour lequel le chevauchement est le plus important (0.387), mais présente une séparation plus faible que $\mathbf{x}_{12}$ (0.613 contre 0.778). Ce point $\mathbf{x}_{12}$ ne prend pas assez d'importance pour pouvoir modifier la partition sélectionnée par l'indice proposé; un autre point, proche de $\mathbf{x}_{12}$, serait nécessaire pour augmenter le nombre de clusters à $c = 3$.

Lorsque le nombre de clusters est fixé à $c = 3$, le point $\mathbf{x}_{12}$ forme un cluster bien séparé ne présentant pas de chevauchement, mais $\mathbf{x}_6$ reste le point où le chevauchement est le plus grand. Le chevauchement de $\mathbf{x}_6$ induit par le choix de $c = 3$ rend cette partition moins bonne que la précédente, en d'autres termes, nous avons $OSI_M(2) \leq OSI_M(3)$.

	Point	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
$(\top,\perp)_M$	$O(\mathbf{u}_k,c)$	.047	.063	.001	.061	.100	.466	.100	.055	.001	.119	.092	.001
	$S(\mathbf{u}_k,c)$	.933	.913	.998	.921	.876	.477	.874	.905	.998	.824	.866	.999
	$\frac{O(\mathbf{u}_k,c)}{S(\mathbf{u}_k,c)}$	.050	.069	.001	.066	.114	.976	.114	.061	.001	.144	.106	.001
$(\top,\perp)_P$	$O(\mathbf{u}_k,c)$	.058	.072	.000	.067	.090	.192	.091	.078	.002	.119	.102	.000
	$S(\mathbf{u}_k,c)$	.996	.993	1	.994	.989	.929	.989	.992	1	.980	.986	1
	$\frac{O(\mathbf{u}_k,c)}{S(\mathbf{u}_k,c)}$	.058	.072	.000	.067	.091	.206	.092	.079	.002	.121	.103	.000
$(\top,\perp)_{H_{10}}$	$O(\mathbf{u}_k,c)$	.032	.037	0	.035	.042	.080	.048	.045	.010	.061	.052	.047
	$S(\mathbf{u}_k,c)$	.999	.999	1	.999	.999	.999	.999	.999	1	.999	.999	1
	$\frac{O(\mathbf{u}_k,c)}{S(\mathbf{u}_k,c)}$	.032	.037	0	.035	.042	.080	.048	.045	.010	.061	.052	.047
$(\top,\perp)_{D_2}$	$O(\mathbf{u}_k,c)$	.053	.068	.000	.064	.095	.358	.131	.103	.011	.137	.113	.040
	$S(\mathbf{u}_k,c)$	.952	.937	.999	.942	.916	.637	.915	.931	.998	.877	.901	.999
	$\frac{O(\mathbf{u}_k,c)}{S(\mathbf{u}_k,c)}$	.055	.073	.000	.068	.103	.562	.143	.110	.011	.156	.126	.040

TAB. 5.4: Mesures de chevauchement, de séparation et $OSI_\perp$ pour les données *Diamond+* et $c = 3$ clusters.

Nous pourrions procéder à la même analyse pour les deux autres couples $(\top,\perp)_{H_{10}}$ et $(\top,\perp)_{D_2}$. Dans cet exemple, l'utilisation du couple de normes triangulaires produit $(\top,\perp)_P$ n'aboutit pas au bon nombre de clusters. Comme nous le verrons dans la suite, le couple algébrique présente un fort comportement de compensation. Pour des données où le nombre d'observations est grand, l'influence des points isolés est faible, mais pour des données en nombre faible, ces points prennent trop d'importance par rapport aux autres. C'est le cas pour les données *Diamond+*.

5.4 Résultats expérimentaux

Dans cette section, nous évaluons les performances de l'indice proposé OSI en proposant une large comparaison de celui-ci avec les neuf autres indices présentés auparavant. Dans la première partie des résultats, l'exposant flou m est fixé à 2. Les résultats suivants se concentreront sur l'influence de ce paramètre m sur un jeu de données synthétiques. Le critère de terminaison de l'algorithme FCM est celui par défaut, 10^{-3}, et une distance euclidienne est utilisée. Le nombre optimal de clusters est recherché dans l'intervalle $[c_{min} = 2,c_{max}]$, où $c_{max} = 10$ pour les données réelles, et $c_{\max} = \min(10,\lfloor\sqrt{n}\rfloor)$ pour les données artificielles, où $\lfloor\cdot\rfloor$ dénote la partie entière, comme il est d'usage dans la littérature.

5.4.1 Données

Nous utilisons onze jeux de données présentant des structures différentes (bonne séparation, clusters qui se chevauchent, présence de points isolés, points de bruit) rendant le problème de validation de partition plus ou moins aisé. La plupart de ces jeux de données sont issus d'articles ayant un lien avec la validation de partitions. Les six premiers sont en dimension 2, permettant ainsi une vérification visuelle plus facile, tandis que les cinq autres décrivent des problèmes réels de classification :

1. *X30* [Bezdek and Pal, 1998], composé de $n = 30$ points divisés en trois clusters bien séparés, mais de tailles différentes, voir FIG. 5.6-(*gauche*).
2. *Bensaid* [Bensaid et al., 1996], caractérisé par trois clusters de tailles très différentes. Les effectifs de chacun des clusters sont respectivement 6, 3 et 40, voir FIG. 5.6-(*droite*).

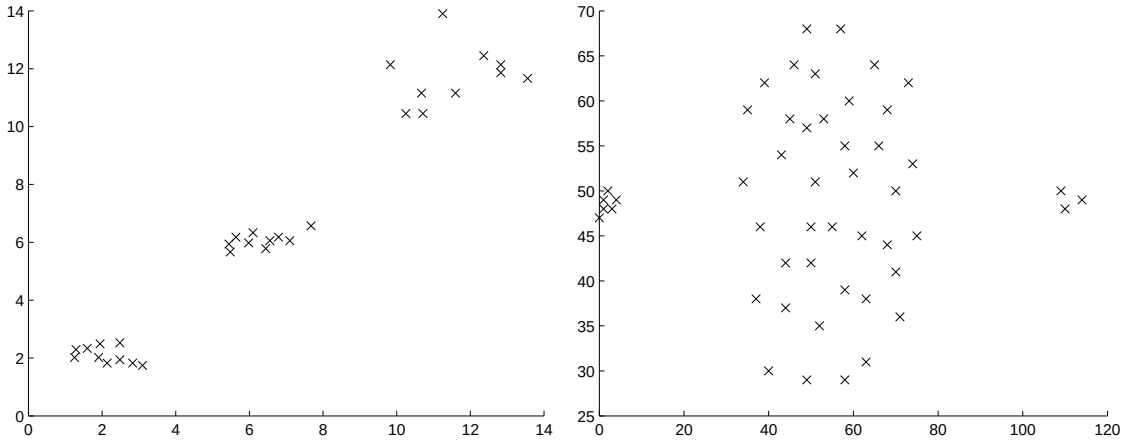

FIG. 5.6: Les données *X30* (*gauche*) et *Bensaid* (*droite*).

3. *Bridge*, contenant quatre clusters, tous reliés par un pont au centre, voir FIG. 5.7-(*gauche*).
4. *4over*, contenant $n = 200$ points distribués par un mélange de $c = 4$ distributions normales bi-variées de 50 points chacunes. Deux d'entre elles se chevauchent de manière assez légère, voir FIG. 5.7-(*droite*).
5. *4noise* est *4over* dans lequel nous avons ajouté 100 points distribués de manière uniforme. Ces points sont ajoutés de manière à simuler un environnement bruité, et par conséquent les deux clusters voisins se chevauchent de façon plus importante, voir FIG. 5.1-(*gauche*).
6. *Diamond+*, déjà décrite dans la section précédente, et visible en FIG. 5.1-(*droite*).
7. *Iris*, composé de 50 fleurs de 3 espèces d'iris : setosa, virginica et versicolor. Chaque fleur est décrite par 4 attributs : longueur et largeur des pétales et sépales. Deux des classes présentent un chevauchement, et bien que le nombre réel de cluster soit de 3, de nombreux auteurs défendent l'idée que d'un point de vue géométrique, le nombre correct est 2.
8. *Wine*, consistant en 13 attributs chimiques décrivant $n = 178$ vins italiens, divisés en trois clusters. Ces clusters sont bien séparés, et les indices de la littérature trouvent généralement le nombre correct de clusters.
9. Les données *Starfield* originales [Xie and Beni, 1991], contenant la position et la luminosité de $n = 51$ étoiles proches de Solaris. Le nombre attendu de clusters est de 8 ou 9, selon les articles.
10. Wisconsin Breast *Cancer*, composé de $n = 699$ cellules malignes/bénignes décrites par 9 attributs obtenus à partir d'images digitalisées.
11. *Pima*, composé de deux clusters se chevauchant, décrivant l'absence ou la présence de diabète, via 8 attributs, dans une population de $n = 768$ indiens.

5.4.2 Étude comparative

Le TAB. 5.5 résume les résultats obtenus avec les indices testés sur les jeux de données synthétiques et réelles. La colonne $c^\star$ donne le nombre correct de clusters pour chaque

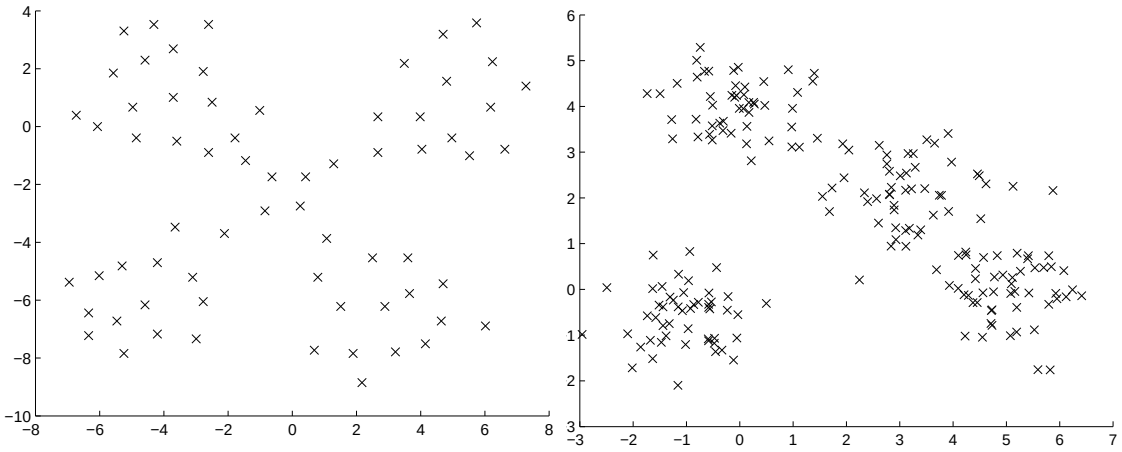

FIG. 5.7: Les données *Bridge* (*gauche*) et *4over*(*droite*).

ensemble de points, et les autres colonnes le nombre de clusters obtenus par l'indice correspondant. Par commodité de lecture, nous indiquerons en rouge et gras les nombres de clusters corrects trouvés par l'indice.
La structure des données *X30* est facile à déterminer (voir FIG. 5.6-(*gauche*)), et la plupart des indices parviennent à trouver le nombre de cluster exact $c = 3$, exceptés FS, SC et K. FS et SC divisent le cluster le plus dispersé, tandis que K sous-estime abusivement le nombre de clusters. Pour le jeu de données *Bensaid*, seuls les indices NPC, XB, K et $OSI_{\perp}$ trouvent le bon nombre de clusters, les autres indices ayant tendance à surestimer le nombre de ces clusters, par exemple 6 ou 7, en divisant le cluster central en plusieurs sous-clusters possédant environ le même nombre de points (voir FIG. 5.6-(*droite*)). Le même problème apparait pour les points connectant les quatre clusters du jeu de données *Bridge*, où seuls les indices NPC, PBM, K et $OSI_{\perp}$ trouvent $c = 4$. Pour le jeu *4over*, les indices NPE, XB, SC et K ne parviennent pas à trouver le bon nombre de clusters, car ils fusionnent les deux clusters les plus proches en un seul (voir FIG. 5.7-(*droite*)). Pour le jeu de données suivant, *4noise*, les seuls indices à trouver le bon nombre de clusters sont FHV, PBM et $OSI_{\perp=S,H_{10},D_2}$. Alors que l'indice WY est construit pour faire face aux points de bruit, il n'est pas forcément performant dans toutes les conditions. Pourtant, il montre de bonnes performances face aux points aberrants (points de bruit moins nombreux), comme c'est montré pour le jeu de données *Diamond+*, où les deux seuls autres indices à trouver le bon nombre de clusters sont NPE et notre indice $OSI_{\perp}$ [4].

Comme il est admis que le nombre correct de clusters pour les données *Iris* est deux ou trois, il n'est pas surprenant que l'ensemble des indices (sauf FS et SC) trouvent un nombre correct. Par contre, il est plus surprenant de voir que les indices NPE, XB, FS et SC ne parviennent pas à déterminer la structure des données *Wine*. Le faible nombre d'observations par rapport au nombre de clusters des données *Starfield* rend difficile la recherche du bon nombre de clusters. L'indice proposé $OSI_{\perp}$, et FHV, SC sont les seuls indices donnant un nombre acceptable de clusters (8 ou 9). La plupart des indices trouvent

4. La bonne performance de l'indice NPE sur ce jeu de données par rapport aux autres indices est à relativiser : cet indice présente une forte tendance à choisir $c = 2$ pour la plupart des données (tendance monotone).

un nombre correct de clusters pour le jeu de données *Cancer*. Les deux clusters de *Pima* présentent un chevauchement, impliquant une erreur de la part des indices FS, FHV, PBM et SC. Aucun des indices étudiés ne parvient à trouver le bon nombre de clusters $c^\star$ pour l'ensemble des jeux de données. Certains d'entre eux sont plus robustes aux point isolés, par exemple WY, au chevauchement entre clusters, par exemple PBM, mais échouent dans d'autres situations présentant par exemple du chevauchement et du bruit. L'indice $OSI_\perp$ proposé trouve le bon nombre $c^\star$ pour tous les jeux de données, excepté lorsqu'on utilise le couple de normes triangulaires produit avec les données *4noise* et *Diamond+*. Ceci s'explique par le fait que ce couple présente un fort comportement de compensation, particulièrement important lorsque des points de bruit sont présents (*4noise*), et lorsqu'il n'y a pas beaucoup d'observations (n faible) et présence de points isolés, ces points prenant un poids prépondérant sur les autres (*Diamond+*).

La fonctionnelle de l'algorithme FCM dépend d'un exposant flou m (voir (3.41)). Ce paramètre supplémentaire a une incidence non négligeable sur les partitions produites par l'algorithme :

- lorsque $m \to 1$, alors U tend vers une partition stricte, $u_{ik} \in \{0, 1\}$,
- lorsque $m \to \infty$, alors U tend vers une partition complètement floue $\overline{U}$, $u_{ik} = 1/c$.

Puisque les partitions sont modifiées, un indice de validité de partition doit également être analysé par rapport à sa robustesse vis à vis des variations de m. Dans [Bezdek and Pal, 1998], Bezdek et Pal montrent que l'algorithme FCM donne de meilleurs résultats lorsque on choisit une valeur de m dans l'intervalle [1.5,2.5]. Dans cette seconde expérience, nous comparons les performances des différents indices sur *4over* et *4noise* lorsque les valeurs de m sont choisies dans l'intervalle [1.5,2.5]. Ici, le paramètre c varie de $c_{min} = 2$ à $c_{max} = 8$. Afin de ne pas surcharger les résultats, seuls les résultats obtenus avec $OSI_{\perp_S}$ sont donnés. Les valeurs sélectionnées par l'ensemble des indices sont données dans le Tab. 5.6. Dans ce tableau, les valeurs correctes sont indiquées en gras et rouge. Comme on peut le voir, $OSI_\perp$ est moins sensible aux changements de m que les autres indices. On note tout de même que les indices WY et FHV montrent également une certaine robustesse par rapport à m, même si on peut trouver des valeurs de m pour lesquels ils se trompent.

En conclusion, nous avons proposé dans cette section une nouvelle famille d'indices de validité de partitions. Cette famille est simplement définie par la moyenne de deux mesures de chevauchement et de séparation n'utilisant pas les informations spatiales, mais uniquement les degrés d'appartenance de la matrice U (mesure indirecte, voir Tab. 5.1). Chacune des deux mesures est définie, pour chaque point $\mathbf{x}_k$ de X, par l'intermédiaire de combinaisons de normes triangulaires appliquées aux degrés d'appartenance de $\mathbf{x}_k$ aux c clusters. Comme il existe de nombreux couples de normes triangulaires, les opérateurs proposés forment donc une famille de combinaisons. Cette combinaison permet de prendre en compte l'importance relative des degrés en dépit de la contrainte due à l'approche floue de FCM et exigeant que la somme des degrés d'appartenance sur les clusters soit égale à 1. Une faible valeur de la mesure de chevauchement pour un point donné signifie qu'il n'appartient pas à une partie chevauchante du cluster le plus probable, tandis qu'une grande valeur de la mesure de séparation implique que ce point est proche du prototype de cluster le plus probable.

Une large comparaison avec les indices les plus fréquents de la littérature, ainsi que des indices récents, sur plusieurs données artificielles et réelles de structure très variée montre la supériorité de la famille d'indices proposée. Nous soulignons ses bonnes performances lorsque les données présentent un chevauchement entre clusters, des ponts entre clusters, des points isolés et des points de bruit, puisque les indices reposant sur l'utilisation exclusive des degrés d'appartenance ne sont généralement pas efficaces dans ces situations.

Le choix du couple de normes triangulaires n'est pas une tâche facile, et reste un problème d'un point de vue aussi bien théorique que pratique. Cela nécessiterait une étude approfondie à propos des propriétés de chacune d'elles, et comment ces propriétés influencent le comportement de l'opérateur par rapport aux valeurs des degrés d'appartenance pour une situation de clustering précise. Par exemple, nous avons observé que les t-normes présentant un fort comportement de compensation ne sont pas adaptées aux cas où le bruit présent dans les données est important.
Il serait également intéressant d'étudier l'influence de l'opérateur d'agrégation utilisé pour obtenir la valeur indice à partir des mesures individuelles sur les points, une première étude, [Wu et al., 2009], sur les opérateurs médian pour des indices existants montre des résultats intéressants.

Données	c_{max}	NPC	NPE	XB	FS	FHV	WY	PBM	SC	K	$OSI_{\perp}$				$c^{\star}$
											$\perp_M$	$\perp_{H_{10}}$	$\perp_P$	$\perp_{D_2}$	
X30	5	**3**	**3**	**3**	4	**3**	**3**	**3**	5	**2**	**3**	**3**	**3**	**3**	3
Bensaid	7	**3**	2	**3**	7	6	6	6	7	**3**	**3**	**3**	**3**	**3**	3
Bridge	8	**4**	2	5	6	5	5	**4**	8	**4**	**4**	**4**	**4**	**4**	4
4over	10	**4**	3	3	**4**	**4**	**4**	**4**	3	3	**4**	**4**	**4**	**4**	4
4noise	10	3	2	2	5	**4**	3	**4**	8	3	**4**	**4**	3	**4**	4
Diamond+	4	3	**2**	3	3	4	**2**	3	4	3	**2**	**2**	3	**2**	2
Iris	10	**3**	**2**	**2**	5	**3**	**2**	**3**	**3**	**2**	**2**	**2**	**2**	**2**	2 ou 3
Wine	10	**3**	2	2	10	**3**	**3**	**3**	6	2	**3**	**3**	**3**	**3**	3
Starfield	10	2	2	6	7	**9**	3	4	**8**	3	**8**	**9**	**9**	**9**	8 ou 9
Cancer	10	**2**	**2**	**2**	3	**2**	**2**	**2**	4	**2**	**2**	**2**	**2**	**2**	2
Pima	10	**2**	**2**	**2**	3	7	**2**	3	6	**2**	**2**	**2**	**2**	**2**	2

TAB. 5.5: Valeurs $c^{\star}$ trouvées par les différents indices sur des données synthétiques, puis réelles.

Données	m	NPC	NPE	XB	FS	FHV	WY	PBM	SC	K	$OSI_\perp$			
											$\perp_M$	$\perp_{H_{10}}$	$\perp_P$	$\perp_{D_2}$
4over	1.5	**4**	**4**	3	5	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
	1.7	**4**	**4**	3	5	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
	1.9	**4**	3	3	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
	2.1	**4**	2	3	**4**	**4**	**4**	**4**	**4**	3	**4**	**4**	**4**	**4**
	2.3	**4**	2	3	**4**	**4**	**4**	**4**	**4**	3	**4**	**4**	5	**4**
	2.5	3	2	3	**4**	**4**	**4**	**4**	6	3	**4**	**4**	5	**4**
4noise	1.5	**4**	3	3	6	3	**4**	5	**4**	6	**4**	5	3	**4**
	1.7	**4**	2	3	6	**4**	**4**	5	6	3	**4**	5	3	**4**
	1.9	3	2	3	5	**4**	**4**	**4**	6	3	**4**	**4**	3	**4**
	2.1	**4**	2	2	5	**4**	3	**4**	6	3	**4**	**4**	3	**4**
	2.3	3	2	3	5	**4**	**4**	**4**	6	3	**4**	**4**	3	**4**
	2.5	3	2	3	**4**	**4**	**4**	2	6	3	**4**	**4**	**4**	**4**

TAB. 5.6: Valeurs $c^\star$ trouvées par les différents indices sur les jeux de données 4*over* et 4*noise*, pour des valeurs de m variant de 1.5 à 2.5.

Conclusion générale

Conclusion

Une sortie, c'est une entrée que l'on prend dans l'autre sens.

— L'Herbe rouge
Boris Vian (1950)

Dans cette conclusion, nous résumons puis discutons les principaux résultats obtenus lors de ce travail, et donnons quelques perspectives de travail. Dans les deux parties principales de ce manuscrit, nous avons donné dans un premier temps un état de l'art aussi complet que possible sur les champs d'étude concernés : opérateurs d'agrégation (Partie 1, Chapitre 1), puis reconnaissance de formes (Partie 2, Chapitre 1 et Chapitre 3). Une fois ces présentations faites, nous avons proposé de nouveaux outils pour chaque champ (Partie 1, Chapitre 2) et (Partie 2, Chapitre 4 et Chapitre 5)

Apports scientifiques

Les mesures de similarité constituent le cœur du Chapitre 2. Nous avons présenté dans un premier temps les mesures de caractérisation de vecteurs décrivant un objet, c'est à dire les mesures entropiques et d'incertitude, de spécificité. Ensuite, nous avons décrit les mesures de comparaison entre deux vecteurs, c'est à dire une comparaison entre deux objets. Ces mesures comprennent les mesures d'inclusion, de similarité, de distance et bien d'autres encore. Nous avons ensuite proposé de nouvelles mesures de similarité. Les trois premières propositions sont des mesures de caractérisation, tandis que la quatrième introduit des mesures de comparaison :

1. à partir du OU flou d'ordre k, le ET flou d'ordre k, utilisant une combinaison de t-normes et t-conormes afin d'évaluer à quel point les k plus grandes (respectivement plus faibles) valeurs sont similaires. Ces mesures sont des opérateurs d'union et d'intersection floue de k sous-ensembles, et les unions et intersections floues fondées sur des normes triangulaires en sont donc des cas particuliers. On peut également voir ces opérateurs comme une généralisation des statistiques d'ordre k, voir section 1.5.
2. l'opérateur de similarité par blocs de valeurs, où au lieu de considérer les k plus grandes ou plus faibles valeurs, les blocs constitués de k valeurs, quelle que soient leur positions, sont évalués. De plus, l'autre nouveauté par rapport à la première proposition consiste à ajouter un noyau symétrique permettant une prise en compte progressive des valeurs du bloc selon leur position dans le bloc, voir section 2.2.3.

3. l'opérateur de similarité fondé sur une approche logique : l'implication logique réciproque de valeurs est vue comme une mesure de similarité. De nombreuses implications peuvent être utilisées, et nous avons présenté dans un premier temps les différents types d'implications floues, pour nous concentrer ensuite sur les implications résiduelles floues, voir section 2.2.4.
4. des mesures de comparaison, (voir section 2.3.4) :
   1 une mesure d'inclusion $\mathcal{I}$ de deux ensembles flous fondée sur l'agrégation de valeurs d'implications individuelles pour chaque x_i de X.
   2 une mesure de similarité $\mathcal{S}$ de deux ensembles flous, également fondée sur les valeurs d'implications des composantes des ensembles.
   3 une mesure de distance $\mathcal{D}$ définie comme le complément de la mesure de similarité $\mathcal{S}$.

   Selon la mesure, des restrictions sont imposées sur les implications ou les opérateurs d'agrégation utilisés : il faut un opérateur d'agrégation conjonctif ou strictement monotone de compensation pour définir une mesure d'inclusion, alors que le choix est libre pour $\mathcal{S}$ et $\mathcal{D}$. Dans tous les cas, nous avons vu qu'il fallait utiliser une implication résiduelle.

Pour l'ensemble de nos propositions, les multiples possibilités de choix de normes triangulaires (en réalité infinies) rendent celles-ci très générales, et constituent de ce fait des familles de mesures. Ce grand nombre de possibilités a aussi des inconvénients : en pratique comment choisir un opérateur plutôt qu'un autre dans une application donnée ? Pour cette raison, nous avons analysé, pour chacune des quatre propositions, leur comportement en fonction du choix du couple de normes triangulaires, mais également en fonction du choix du paramètre γ si le couple est paramétrique. Nous espérons que l'utilisateur y trouvera un guide pour faire son choix.

Dans le Chapitre 4 et le Chapitre 5, deux applications impliquant des études théoriques ont été abordées. La première de ces applications est l'option de rejet. Dans cette partie, nous avons proposé d'utiliser les outils introduits au Chapitre 2. On distingue ici deux types de règles, correspondant à la partition opérée sur l'espace de représentation. Dans le cas rejet total, nous avons proposé de se servir la famille de mesures de similarité introduite en section 2.2.4 en tant que mesure d'ambiguïté. Nous avons montré comment ces mesures correspondent à de nombreuses propositions de la littérature selon la norme triangulaire utilisée. Dans le second cas, où l'on choisit les classes parmi les c possibles, nous avons proposé d'utiliser l'opérateur de similarité par blocs proposé en section 2.2.3, ici encore en tant que mesure d'ambiguïté. Nous avons introduit une règle générique de sélection de classes incluant le rejet en distance. Afin d'évaluer le rejet total, nous avons utilisé l'aire sous la courbe Erreur-Rejet. Dans le cas du rejet sélectif, nous avons introduit une nouvelle mesure d'évaluation : l'aire sous la courbe Erreur-Nombre Moyen de Classes. Pour les deux cas considérés, nos propositions permettent d'obtenir de meilleures performances sur des données aussi bien artificielles que réelles.
La seconde application proposée est la validation de partitions, consistant principalement à trouver le nombre optimal de groupes dans un problème de classification non supervisée.

Nous avons proposé d'utiliser l'opérateur introduit en section 1.5 afin de définir des mesures de chevauchement et de séparation pour chaque forme observée. La comparaison de ce nouvel indice par rapport aux autres sur des jeux de données artificiels et réels montre l'intérêt de l'approche : la mesure est plus performante et plus stable aux variations du paramètre d'exposant flou m de l'algorithme FCM.

L'ensemble des publications relatives aux travaux de thèse est listé en pages 153-154. Une sélection d'articles non développés ou représentatifs des travaux de thèse est disponible en Annexe E.

Futures pistes de travail

Les opérateurs proposés ont une multitude d'utilisation possibles, et nous ne les avons appliqués que sur une partie d'entre elles. Parmi les domaines qui nous intéressent, prenons le traitement d'image. Dans [Le Capitaine et al., 2007b], nous soulignons les liens existants entre l'opérateur de similarité par blocs et la morphologie mathématique floue. En particulier, il serait intéressant d'analyser les résultats des opérations morphologiques usuelles (dilatation, érosion, gradient) si l'on utilise nos outils. Nous avons également, dans [Le Capitaine and Frélicot, 2009], proposé une méthode automatique de segmentation d'images couleurs en cherchant le nombre optimal de clusters, ou plus généralement de régions homogènes, grâce à la méthode proposée au Chapitre 5. Toujours dans le domaine de l'image, nous voudrions étudier les performances pour l'appariement d'images des mesures de similarité $\mathcal{S}$ proposée au Chapitre 2. Enfin, puisque les opérateurs flou d'ordre k sont des extensions des statistiques d'ordre, dont l'opérateur médian est un cas particulier, nous souhaitons évaluer à quel point l'utilisation de nos opérateurs pour le filtrage ayant pour but de réduire le bruit dans une image est performante.

Nous projetons également de continuer nos recherches dans le domaine de la reconnaissance de formes. Les mesures d'inclusion $\mathcal{I}$ sont d'un intérêt tout à fait important dans ce cadre. Dans [Fan et al., 1999], les auteurs proposent l'utilisation de mesures d'inclusion pour la définition d'indices de validité de partitions. Nos mesures permettraient sans doute une plus grande souplesse grâce à leur paramètre γ lié aux normes triangulaires. Dans le même ordre d'idée, dans [Kuncheva et al., 2001], les auteurs utilisent aussi des mesures d'inclusion afin de fusionner les sorties de multiples classifieurs. Comme les mesures décrites dans cette contribution ne sont pas robustes (utilisation de min ou max pour l'opérateur d'agrégation), l'utilisation de nos mesures permettraient d'obtenir de meilleurs résultats en terme de taux d'erreur. La capacité des opérateurs définis à évaluer si k valeurs sont importantes peut être utilisée dans plusieurs domaines. Ici, ces valeurs sont comprises entre 0 et 1. Une analyse en composantes principales normée fournit des vecteur propres dont les valeurs propres associées sont dans ce même intervalle, nos opérateurs peuvent donc aider à déterminer le nombre de valeurs propres à conserver. De même en *spectral clustering*, où l'on calcule généralement les valeurs et vecteur propres du Laplacien d'une matrice de similarité entre pixels. Le choix du nombre de groupes dans ce contexte étant difficile, il est envisageable d'utiliser nos opérateurs pour faciliter ce choix [Bach and Jordan, 2006].

Dans le contexte de la sélection de variables discriminantes en classification supervisée, on retrouve une nouvelle fois le problème du choix du nombres d'attributs utilisés pour la description des objets, [Semani et al., 2004]. Enfin, nous n'avons évidemment pas terminé notre travail sur l'option de rejet. Par exemple, considérons l'opérateur de similarité par blocs. Dans la définition actuelle, nous utilisons des noyaux symétriques. Un évolution notable consisterait à les remplacer par une véritable mesure floue, prenant en compte les interactions entre sous-ensembles de classes.

Références de l'auteur

C. Frélicot & H. Le Capitaine. Chapter Class-selective Rejection Rules based on the Aggregation of Pattern Soft Labels, in Pattern Recognition, Intech, 2009.

C. Frélicot & H. Le Capitaine. Détermination du nombre de classes d'une partition floue par mesures de séparation et de chevauchement fondées sur des opérateurs d'agrégation adaptés. XVIth Joint Meeting of the French Society of Classification, SFC09. Grenoble, France, 2009.

H. Le Capitaine & C. Frélicot. Segmentation d'images couleur par des mesures de chevauchement et de séparation fondées sur l'agrégation de partition floue. To appear In Rencontres Francophones sur la Logique Floue et ses Applications LFA09. Annecy, France, 2009.

H. Le Capitaine & C. Frélicot. Classification supervisée avec options de rejet : vers une approche généralisée par implications floues. To appear In Rencontres Francophones sur la Logique Floue et ses Applications LFA09. Annecy, France, 2009.

H. Le Capitaine & C. Frélicot. A fuzzy modeling approach to cluster validity. In Proc. IEEE Int. Conf. on Fuzzy Systems, FUZZIEEE09. Jeju Island, Korea, 2009.

H. Le Capitaine & C. Frélicot. Classification with reject options in a logical framework: a fuzzy residual implication approach. In Proc. Int. Fuzzy Systems Assoc. World Congress, IFSA -EUSFLAT 09. Lisboa, Portugal, 2009.

H. Le Capitaine & C. Frélicot. Towards a unified logical framework of fuzzy implications to compare fuzzy sets. In Proc. Int. Fuzzy Systems Assoc. World Congress, IFSA -EUSFLAT 09. Lisboa, Portugal, 2009.

H. Le Capitaine & C. Frélicot. A family of cluster validity indexes based on a l-order fuzzy OR operator. Joint IAPR Int. Workshops on Structural and Syntactic Pattern Recognition and Statistical Techniques in Pattern Recognition, SSPR&SPR 2008, LNCS 5342, pp. 622-631, Orlando, Florida, 2008.

H. Le Capitaine & C. Frélicot. A Class-Selective Rejection Scheme based on Blockwise Similarity of Typicality Degrees. In Proc. 19th Int. Conf. on Pattern Recognition, ICPR 08. Tampa, Florida, 2008.

H. Le Capitaine & C. Frélicot. Une approche par inférence floue pour le classement avec rejet et la validation de partitions. In Rencontres Francophones sur la Logique Floue et ses Applications LFA08, Cépaduès, pp. 126-133, Lens, France, 2008.

H. Le Capitaine & C. Frélicot. A new fuzzy 3-rules pattern classifier with reject options based on aggregation of membership degrees. In Proc. 12th Int. Conf. on Information Processing and Management of Uncertainty, IPMU08. Spain, 2008.

H. Le Capitaine, T. Batard, C. Frélicot & M. Berthier. Blockwise similarity in [0,1] via triangular norms and Sugeno integrals - Application to cluster validity. In Proc. IEEE Int. Conf. on Fuzzy Systems, FUZZIEEE07. London, United Kingdom, 2007.

H. Le Capitaine, T. Batard, C. Frélicot & M. Berthier. Mesure de similarité par blocs via les normes triangulaires et l'intégrale de Sugeno - Application à la détection de contours. In Rencontres Francophones sur la Logique Floue et ses Applications LFA07, Cépaduès. Nîmes, France, 2007.

Articles soumis au moment de l'envoi :

H. Le Capitaine & C. Frélicot. A Cluster Validity Index combining an Overlap Measure and a Separation Measure based on Fuzzy Aggregation Operators. Soumis à IEEE Transactions on Fuzzy Systems.

Articles en fin de rédaction :

H. Le Capitaine & C. Frélicot. A class-selective rejection rule based on a similarity measure of ordered typicality degrees. Pour soumission à Pattern Recognition.

Bibliographie

J. Aczél. Sur les opérations définies pour nombres réels. *Bulletin de la Société Mathématiques de France*, 76:59–64, 1949. [cité en p. 19]

J. Aczél and C. Alsina. Characterizations of some classes of quasilinear functions with applications to triangular norms and to synthesizing judgements. *Methods Oper. Res.*, 48:3–22, 1984. [cité en p. 179]

M. Akay, editor. *Time-frequency and Wavelets in Biomedical Signal Processing*. John Wiley & Sons Inc., 1998. [cité en p. 77]

C. Alsina and E. Trillas. On the symmetric difference of fuzzy sets. *Fuzzy Sets and Systems*, 153(2):181–194, 2005. [cité en p. 57]

C. Alsina, M. Frank, and B. Schweizer. *Associative functions: triangular norms and copulas*. World Scientific, 2006. [cité en p. 16, 172]

F. Bach and M. Jordan. Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research*, 7:1963–2001, 2006. [cité en p. 151]

V. Balopoulos, A. Hatzimichailidis, and B. Papadopoulos. Distance and similarity measures for fuzzy operators. *Information Sciences*, 177:2336–2348, 2007. [cité en p. 69]

W. Bandler and L. Kohout. Fuzzy power sets and fuzzy implication operators. *Fuzzy Sets and Systems*, 4:13–30, 1980. [cité en p. 56, 58, 59, 66]

P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008. [cité en p. 100]

G. Beliakov. Learning weights in the generalized owa operators. *Fuzzy Optimization and Decision Making*, 4(2):119–130, 2005. [cité en p. 15]

G. Beliakov, A. Pradera, and T. Calvo. *Aggregation Functions: A guide for practitionners*. Studies in Fuzziness and Soft Computing. Springer-Verlag, 2008. [cité en p. 10]

R. Bellman. *Adaptive Control Processes*. Princeton University Press, 1961. [cité en p. 92]

A. Bensaid, L. Hall, J. Bezdek, L. Clarke, M. Silbiger, J. Arrington, and R. Murtagh. Validity-guided (re)clustering with applications to image segmentation. *IEEE Transactions on Fuzzy Systems*, 4(2):112–123, 1996. [cité en p. 129, 139]

P. Benvenuti and D. Vivona. General theory of the fuzzy integral. *Mathware & Soft Computing*, 3(1-2):199–209, 1996. [cité en p. 25]

J. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3:58–72, 1974. [cité en p. 128]

J. Bezdek. *Pattern Recognition with fuzzy objective function algorithm*. Plenum Press, 1981. [cité en p. 89, 125, 128]

J. Bezdek and N. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics*, 23(3):301–315, 1998. [cité en p. 134, 139, 142]

J. Bezdek, D. Dubois, and H. Prade, editors. *Fuzzy Sets in Approximate Reasoning and Information Systems*. Springer-Verlag, 1999a. [cité en p. 177]

J. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer Academic, 1999b. [cité en p. 24, 91, 126]

C. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [cité en p. 80]

C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995. [cité en p. 80]

C. Blake and C. Merz. Uci repository of machine learning databases, 1998. [cité en p. 116]

I. Bloch. Distances in fuzzy sets for image processing derived from fuzzy mathematical morphology. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 1307–1312, Granada, 1996. [cité en p. 54]

I. Bloch. On fuzzy distances and their use in image processing. *Pattern Recognition*, 32 (11):1873–1895, 1999. [cité en p. 57]

K. Bosteels and E. Kerre. A triparametric family of cardinality-based fuzzy similarity measures. *Fuzzy Sets and Systems*, 158(22):2466–2479, 2007. [cité en p. 58]

F. Botana. Deriving fuzzy subsethood measures from violations of the implication between elements. In *11th Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, volume LNCS 1415, pages 234–243. Springer, 1998. [cité en p. 68]

B. Bouchon-Meunier, M. Rifqi, and S. Bothorel. Towards general measures of comparison of objects. *Fuzzy Sets and Systems*, 84(2):143–153, 1996. [cité en p. 55, 57, 67]

M. Bouguessa, S. Wang, and H. Sun. An objective approach to cluster validation. *Pattern Recognition Letters*, 27(13):1419–1430, 2006. [cité en p. 132]

N. Bruijn, C. van Ebbenhorst Tengbergen, and D. Kruyswijk. On the set of divisors of a number. *Nieuw Arch. Wisk.*, 23(2):191–193, 1951. [cité en p. 28]

R. Brunelli. *Template Matching Techniques in Computer Vision: Theory and Practice*. John Wiley & Sons Inc., 2009. [cité en p. 80]

H. Bunke and A. Sanfeliu, editors. *Syntactic And Structural Pattern Recognition - Theory And Applications*. World Scientific, 1990. [cité en p. 79]

P. Burillo, N. Frago, and R. Fuentes. Inclusion grade and fuzzy implication operators. *Fuzzy Sets and Systems*, 114(3):143–153, 2000. [cité en p. 68, 69]

H. Bustince, V. Mohedano, E. Barrenechea, and M. Pagola. Definition and construction of fuzzy di-subsethood measures. *Information Sciences*, 176(21):3190–3231, 2006. [cité en p. 69]

T. Calvo and R. Mesiar. Generalized medians. *Fuzzy Sets and Systems*, 124(1):59–64, 2001. [cité en p. 16]

T. Calvo and R. Mesiar. Weigthed triangular norms-based aggregation operators. *Fuzzy Sets and Systems*, 137(1):3–10, 2003. [cité en p. 174]

T. Calvo, B. De Baets, and J. Fodor. The functional equations of frank and alsina for uninorms and nullnorms. *Fuzzy Sets and Systems*, 120:385–394, 2001. [cité en p. 174]

T. Calvo, G. Mayor, and R. Mesiar, editors. *Aggregation Operators: New Trends and Applications*. Physica-Verlag, 2002. [cité en p. 10, 176]

A. Cauchy. *Cours d'analyse de l'Ecole Royale Polytechnique, Vol. I. Analyse algébrique,*. Imprimerie Royale, Debure, 1821. [cité en p. 13]

V. Chandola, A. Banerjee, and V. Kumar. Outlier detection: A survey. Technical Report, 2007. [cité en p. 93]

S. Chen, M. Yeh, and P. Hsio. A comparison of similarity measures of fuzzy values. *Fuzzy Sets and Systems*, 72(1):79–89, 1995. [cité en p. 65]

G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1954. [cité en p. 22, 23]

C. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, 6(4):247–254, 1957. [cité en p. 100, 104]

C. Chow. On optimum error and reject tradeoff. *IEEE Transactions on Information Theory*, 16:41–46, 1970. [cité en p. 100, 102, 108, 114]

R. Clemen and R. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2):187–203, 1999. [cité en p. 9]

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. [cité en p. 77]

V. Cross and T. Sudkamp. *Similarity and Compatibility in Fuzzy Set Theory*. Physica-Verlag; Heidelberg, 2002. [cité en p. 35, 36]

R. Davé. Generalized fuzzy c-shells clustering and detection of circular and elliptical boundaries. *Pattern Recognition*, 25(7):713–721, 1992. [cité en p. 91]

B. De Baets, H. De Meyer, and H. Naessens. On rational cardinality-based inclusion measures. *Fuzzy Sets and Systems*, 128(2):169 – 183, 2002. [cité en p. 58, 59]

A. De Luca and S. Termini. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information Control*, 20:301–312, 1972. [cité en p. 36, 57, 67, 68, 128]

C. De Stefano, C. Sansone, and M. Vento. To reject or not to reject: that is the question-an answer in caseof neural classifiers. *IEEE Transactions on Systems, Man and Cybernetics, part C*, 30(1):84–94, 2000. [cité en p. 103, 105]

G. Della Riccia, D. Dubois, H.-J. Lenz, and R. Kruse, editors. *Preferences and Similarities*. Springer Wien New-York, 2008. [cité en p. 36]

M. Detyniecki. *Mathematical Aggregation Operators and their application to video querying*. PhD thesis, University of Paris VI, 2000. [cité en p. 10]

P. Devijver and J. Kittler. *Pattern Recognition : A statistical approach*. Prentice-Hall, 1982. [cité en p. 76]

J. Dombi. A general class of fuzzy operators, the de morgan class of fuzzy operators and fuzziness measures induced by fuzzy operators. *Fuzzy Sets and Systems*, 8:149–163, 1982a. [cité en p. 21, 181]

J. Dombi. Basic concepts for a theory of evaluation: The aggregative operator. *European Journal of Operational Research*, 10(3):282–293, 1982b. [cité en p. 174]

J. Dombi and L. Porkolab. On measures of fuzziness. *Annales Univ. Sci. Budapest, Sect. Comp.*, 12:69–78, 1991. [cité en p. 37]

D. Dubois and H. Prade, editors. *Fundamentals of Fuzzy Sets*. Springer-Verlag, 2000. [cité en p. 177]

D. Dubois and H. Prade. On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, 142(11):143–161, 2004. [cité en p. 10]

D. Dubois and H. Prade. *Fuzzy Sets and Systems - Theory and Applications*. Academic Press, New York, 1980. [cité en p. 21, 64, 177, 182]

D. Dubois and H. Prade. A unifying view of comparison indices in a fuzzy set-theoretic framework. In R. Yager, editor, *Fuzzy Set and Possibility Theory. Recent Developments*, pages 3–13. Pergamon Press, 1982a. [cité en p. 35, 58]

D. Dubois and H. Prade. A class of fuzzy measures based on triangular norms a general framework for the combination of uncertain information. *International Journal of General Systems*, 8(1):43–61, 1982b. [cité en p. 22]

D. Dubois and H. Prade. A review of fuzzy set aggregation connectives. *Information Sciences*, 36:85–121, 1985. [cité en p. 10, 13]

D. Dubois and H. Prade. The principle of minimum specificity as a basis for evidential reasoning. In B. Bouchon-Meunier and R. Yager, editors, *Uncertainty in Knowledge-Based Systems*, pages 75–84. Springer Berlin, 1986. [cité en p. 37]

D. Dubois and H. Prade. *Possibility Theory : An Approach to Computerized Processing of Uncertainty*. Plenum Press, 1988. [cité en p. 9, 178]

D. Dubois, H. Prade, and C. Testemale. Weighted fuzzy pattern matching. *Fuzzy Sets and Systems*, 28(3):313–331, 1988. [cité en p. 86]

D. Dubois, H. Fargier, and H. Prade. Beyond min aggregation in multicriteria decision : (ordered) weighted min, discri-min, leximin. In R. Yager and J. Kacprzyk, editors, *The ordered weighted averaging operators. Theory and applications*, pages 181–192. Kluwer Academic Publishers, 1997. [cité en p. 16]

B. Dubuisson and M. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26:155–165, 1993. [cité en p. 93, 98, 101, 104, 105]

R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley Interscience, 2001. [cité en p. 74]

J. Dunn. Indices of partition fuzziness and the detection of clusters in large data sets. In *Fuzzy Automata and Decision processes*. Elsevier, NY, 1977. [cité en p. 129]

W. Edwards. Probability-preferences in gambling. *American Journal of Psychology*, 66: 349–364, 1953. [cité en p. 23]

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. of the American Statistical Association*, 78:316–331, 1983. [cité en p. 79]

J. Fan, W. Xie, and J. Pei. Subsethood measure: new definitions. *Fuzzy Sets and Systems*, 106(2):201–209, 1999. [cité en p. 59, 68, 69, 151]

T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. [cité en p. 115]

J. Fodor and J. Marichal. On nonstrict means. *Aequationes Mathematicae*, 54(1-2):308–327, 1997. [cité en p. 13]

J. Fodor, R. Yager, and A. Rybalov. Structure of uninorms. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 5(4):411–427, 1997. [cité en p. 173]

P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Multiclassification: reject criteria for the bayesian combiner. *Pattern Recognition*, 32(8):1435–1447, 1999. [cité en p. 103, 105]

P. Fonck, J. Fodor, and M. Roubens. An application of aggregation procedures to the definition of measures of similarity between fuzzy sets. *Fuzzy Sets and Systems*, 97(1): 67–74, 1998. [cité en p. 68]

L. Fono, H. Gwet, and B. Bouchon-Meunier. Fuzzy implication operators for difference operations for fuzzy sets and cardinality-based measures of comparison. *European Journal of Operational Research*, 183(1):314–326, 2007. [cité en p. 51, 69]

M. Frank. On the simultaneous associativity of $f(x,y)$ and $x + y - f(x,y)$. *Aequationes Mathematicae*, 19:194–226, 1979. [cité en p. 21, 184]

C. Frélicot. Multiprototype-based fuzzy classification and reject options. In *5th IEEE International Conference on Fuzzy Systems*, volume 3, pages 2026–2031, 1996. [cité en p. 102]

C. Frélicot and B. Dubuisson. A multi-step predictor of membership function as an ambiguity reject solver in pattern recognition. In *Proceedings of the 4th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU*, pages 709–715, 1992. [cité en p. 103, 110]

C. Frélicot and H. Le Capitaine. Class-selective rejection rules based on the aggregation of pattern soft labels. In *Pattern Recognition*. Intech, 2009. [cité en p. 38, 104]

C. Frélicot, L. Mascarilla, and A. Fruchard. An ambiguity measure for pattern recognition problems using triangular-norms combination. *WSEAS Transactions on Systems*, 8(3): 2710–2715, 2004. [cité en p. 25]

C. Frélicot and L. Mascarilla. Reject strategies driven combination of pattern classifiers. *Pattern Analysis and Applications*, 5(2):234–243, 2002. [cité en p. 103]

Y. Fukuyama and M. Sugeno. A new method for choosing the number of clusters for the fuzzy c-means method. In *Proc. 5th Fuzzy Systems Symposium*, pages 247–250, 1989. [cité en p. 129]

R. Fuller. On obtaining owa operator weights: a short survey of recent developments. In *5th IEEE Int. Conf. on Computational Cybernetics*, pages 241–244, 2007. [cité en p. 15]

G. Fumera and F. Roli. Support vector machines with embedded reject option. In *Fisrt International Workshop on Pattern Recognition with Support Vector Machines*, 2002. [cité en p. 100]

G. Fumera, F. Roli, and G. Giacinto. Reject option with multiple thresholds. *Pattern Recognition*, 33(12):2099–2101, 2000. [cité en p. 101, 102, 103, 104, 105]

L. Garmendia, R. Yager, E. Trillas, and A. Salvador. On t-norms based measures of specificity. *Fuzzy Sets and Systems*, 133(2):237–248, 2003. [cité en p. 38]

I. Gath and A. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):773–780, 1989. [cité en p. 91, 129]

J. Goguen. The logic of inexact concepts. *Synthese*, 19:325–373, 1969. [cité en p. 64]

M. Golfarelli, D. Maio, and D. Maltoni. On the error-reject trade-off in biometric verification systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):786–796, 1997. [cité en p. 114]

M. Grabisch. Symmetric and asymmetric fuzzy integrals: the ordinal case. In *6th Int. Conf. on Soft Computing*, Iizuka, Japan, 2000a. [cité en p. 10]

M. Grabisch. Fuzzy integral for classification and feature extraction. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals — Theory and Applications*, pages 348–374. Physica Verlag, 2000b. [cité en p. 9, 24, 86]

M. Grabisch. A new algorithm for identifying fuzzy measures and its application to pattern recognition. In *Int. Joint Conf. of the 4th IEEE Int. Conf. on Fuzzy Systems and the 2nd Int. Fuzzy Engineering Symposium*, pages 145–150, Yokohama, Japan, 1995. [cité en p. 86]

M. Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2):167–189, 1997. [cité en p. 22, 23]

M. Grabisch and C. Labreuche. Bi-capacities–i: definition, möbius transform and interaction. *Fuzzy Sets and Systems*, 151(2):211–236, 2005a. [cité en p. 23]

M. Grabisch and C. Labreuche. Bi-capacities–ii: the choquet integral. *Fuzzy Sets and Systems*, 151(2):237–259, 2005b. [cité en p. 23]

M. Grabisch and C. Labreuche. A decade of application of the choquet and sugeno integrals in multi-criteria decision aid. *4OR*, 6:1–44, 2008. [cité en p. 24]

M. Grabisch, G. Bienvenu, A. Ayoun, J. Grandin, A. Lemer, and M. Moruzzi. A formal comparison of probabilistic and possibilistic frameworks for classification. In *7th IFSA World Congress*, Prague, 1997. [cité en p. 86]

M. Grabisch, S. Orlovski, and R. Yager. Fuzzy aggregation of numerical preferences. In R. Slowinski, editor, *The Handbook of Fuzzy Sets Series, Vol. 4: Fuzzy Sets in Decision Analysis, Operations Research and Statistics*, pages 31–68. Kluwer Academic, 1998. [cité en p. 11]

M. Grabisch, T. Murofushi, and M. Sugeno, editors. *Fuzzy Measures and Integrals. Theory and Applications*. Physica Verlag, 2000. [cité en p. 10]

M. Grabisch, J. Marichal, R. Mesiar, and E. Pap. *Aggregation Functions*. Number 127 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2009. [cité en p. 10, 32]

Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu. Support vector machines with a reject option. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 537–544. NIPS, 2008. [cité en p. 100]

R. Gunderson, J. Bezdek, C. Coray, and J. Watson. Detection and characterization of cluster substructure: I. linear structure : fuzzy c-lines. *SIAM Journal of Applied Mathematics*, 40(2):339–357, 1981. [cité en p. 91]

S. Gupta. On some multiple decision (selection and ranking) rules. *Technometrics*, 7: 225–245, 1965. [cité en p. 106]

D. Gustafson and W. Kessel. Fuzzy clustering with fuzzy covariance matrix. In *Proc. IEEE Conference on Decision and Control*, pages 761–766, San Diego, California, 1979. [cité en p. 91]

T. Ha. The optimum class-selective rejection rule. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):608–615, 1997. [cité en p. 103, 106, 107, 109, 115]

H. Hamacher. On logical connectives of fuzzy statements and their affiliated truth functions. *Progress in cybernetics and system research*, 3:276–288, 1978. [cité en p. 21, 185]

L. Hansen, C. Liisberg, and P. Salamon. The error-reject tradeoff. *Open Systems and Information Dynamics*, 4:159–184, 1997. [cité en p. 114]

R. Hartley. The measurement of information. *Bell System Technical Journal*, 7(3):535–563, 1928. [cité en p. 37]

W. Highleyman. Linear decision functions, with application to pattern recognition. *Proceedings of the IRE*, 50(6):1501–1514, 1962. [cité en p. 116]

K. Hirota and W. Pedrycz. Matching fuzzy quantities. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(6):1580–1586, 1991. [cité en p. 49, 58, 66, 108]

T. Horiuchi. Class-selective rejection rule to minimize the maximum distance between selected classes. *Pattern Recognition*, 31:1579–1588, 1998. [cité en p. 103, 107, 110]

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons Inc., 2001. [cité en p. 77]

H. Ishibuchi and T. Nakashima. Fuzzy classification with reject options by fuzzy if-then rules. In *IEEE World Congress on Computational Intelligence*, 1998. [cité en p. 103, 105]

P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Société de Vaud des Sciences Naturelles*, 44:223, 1908. [cité en p. 57]

A. Jain, R. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000. [cité en p. 75, 76, 80, 92]

J. Karmeshu, editor. *Entropy measures, maximum entropy principle and emerging applications*. Springer-Verlag, 2003. [cité en p. 37]

A. Kehagias and M. Konstantinidou. L-fuzzy valued inclusion measure, l-fuzzy similarity and l-fuzzy distance. *Fuzzy Sets and Systems*, 136(3):313–332, 2003. [cité en p. 68, 69]

D. Kim, K. Lee, and D. Lee. On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition*, 37(10):2009–2025, 2004. [cité en p. 126, 131, 133]

E. Klement and R. Mesiar. *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*. Elsevier, 2005. [cité en p. 16, 131]

E. Klement, R. Mesiar, and E. Pap. *Triangular Norms*. Kluwer Academic, 2000. [cité en p. 16, 17]

G. Klir. *Uncertainty and Information: Foundations of Generalized Information Theory*. Wiley, 2006. [cité en p. 37]

G. Klir and M. Mariano. The uniqueness of possibilistic measures of uncertainty and information. *Fuzzy Sets and Systems*, 24(2):197–219, 1987. [cité en p. 37]

O. Külpe. *Outlines of psychology*. Arno Press, New-York, 1895. [cité en p. 35]

J. Knopfmacher. On measures of fuzziness. *Journal of Mathematical Analysis and Applications*, 49:529–534, 1975. [cité en p. 36]

A. Kolmogorov. Sur la notion de moyenne. *Rendiconti Accademia dei Lincei*, 12(6):388–391, 1930. [cité en p. 13]

B. Kosko. *Neural networks and fuzzy systems*. Prentice Hall, Englewood Cliffs, NJ, 1992. [cité en p. 66, 68]

R. Krishnapuram and J. Keller. A possibilsitic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110, 1993. [cité en p. 90, 99]

R. Krishnapuram and J. Keller. The possibilistic c-means algorithm: insights and recommendations. *IEEE Transactions on Fuzzy Systems*, 4(3):385–393, 1996. [cité en p. 90]

R. Krishnapuram, H. Frigui, and O. Nasraoui. Fuzzy and possibilistic shell clustering algorithms and their application and surface approximation, i–ii. *IEEE Transactions on Fuzzy Systems*, 3(1):29–60, 1995. [cité en p. 91]

S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951. [cité en p. 36]

L. Kuncheva. How good are fuzzy if-then classifiers? *IEEE Transactions on Systems, Man and Cybernetics*, 30(4):501–509, 2000. [cité en p. 85, 86]

L. Kuncheva. Using measures of similarity and inclusion for multiple classifier fusion by decision templates. *Fuzzy Sets and Systems*, 122(3):401–407, 2001. [cité en p. 70]

L. Kuncheva. *Combining Pattern Classifiers*. Wiley-IEEE, 2004. [cité en p. 24]

L. Kuncheva, J. Bezdek, and R. Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314, 2001. [cité en p. 70, 151]

S. Kwon. Cluster validity index for fuzzy clustering. *Electronic Letters*, 34(22):2176–2177, 1998. [cité en p. 130, 131]

T. Landgrebe and R. Duin. Efficient multiclass roc approximation by decompostion via confusion matrix perturbation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):810–822, 2008. [cité en p. 115]

T. Landgrebe, D. Tax, P. Paclik, and R. Duin. The interaction between classification and reject performance for distance-based reject options classifiers. *Pattern Recognition Letters*, 27:908–917, 2006. [cité en p. 102, 105, 115]

H. Le Capitaine and C. Frélicot. A family of cluster validity indexes based on a l-order fuzzy or operator. In *Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshops, SSPR&SPR*, Orlando, USA, 2008a. [cité en p. 136]

H. Le Capitaine and C. Frélicot. A class-selective rejection scheme based on blockwise similarity of typicality degrees. In *International Conference on Pattern Recognition, ICPR*, Tampa, USA, December 2008b. IEEE. [cité en p. 47, 107, 108]

H. Le Capitaine and C. Frélicot. A new fuzzy 3-rules pattern classifier with reject options based on aggregation of membership degrees. In *International Conference on Information Processing and Management of Uncertainty, IPMU*, Malaga, Spain, June 2008c. [cité en p. 86, 105]

H. Le Capitaine and C. Frélicot. Classification with reject options in a logical framework: a fuzzy residual implication approach. In *13th International Fuzzy Systems Association World Congress, IFSA*, Lisboa, Portugal, 2009a. [cité en p. 50, 105, 108, 109, 110, 112]

H. Le Capitaine and C. Frélicot. Towards a unified logical framework of fuzzy implications to compare fuzzy sets. In *13th International Fuzzy Systems Association World Congress, IFSA*, Lisboa, Portugal, 2009b. [cité en p. 58, 60, 61, 62, 64, 65]

H. Le Capitaine and C. Frélicot. A cluster validity index combining an overlap measure and a separation measure based on fuzzy aggregation operators. *Soumis à IEEE Transactions on Fuzzy Systems*,?:?–?, 2009c. [cité en p. 26, 135, 136, 137]

H. Le Capitaine and C. Frélicot. Segmentation d'images couleur par des mesures de chevauchement et de séparation fondées sur l'agrégation de partition floue. In *Rencontres Francophones sur la Logique Floue et ses Applications LFA09*, 2009. [cité en p. 151]

H. Le Capitaine, T. Batard, C. Frélicot, and M. Berthier. Blockwise similarity in [0,1] via triangular norms and sugeno integrals - application to cluster validity. In *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE*, London, England, July 2007a. [cité en p. 42, 47]

H. Le Capitaine, T. Batard, C. Frélicot, and M. Berthier. Mesure de similarité par blocs via les normes triangulaires et l'intégrale de sugeno - application à la détection de contours. In *Rencontres Francophones sur la Logique Floue et ses Applications, LFA*, Nimes, France, November 2007b. Cepadues. [cité en p. 47, 151]

X. Liu. Entropy, distance measure and similarity measure of fuzzy sets and their relations. *Fuzzy Sets and Systems*, 52(3):305–318, 1992. [cité en p. 61, 62]

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. [cité en p. 88]

X. Luo and N. Jennings. A spectrum of compromise aggregation operators for multi-attribute decision making. *Artificial Intelligence*, 171(2-3):161–184, 2007. [cité en p. 12]

K. Maes, S. Saminger, and B. De Baets. On the characterization of self-dual aggregation operators. In *EUSFLAT - LFA*, pages 159–164, 2005. [cité en p. 176]

J. Marichal. On an axiomatization of the quasi-arithmetic mean values without the symmetry axiom. *Aequationes Mathematicae*, 59(1-2):74–83, 2000. [cité en p. 13]

J. Marichal. *Aggregation Operators for Multicriteria Decision Aid*. PhD thesis, University of Liège, 1998. [cité en p. 10]

M. Mas, M. Monserrat, J. Torrens, and E. Trillas. A survey on fuzzy implication functions. *IEEE Transactions on Fuzzy Systems*, 15(6):1107–1121, 2007. [cité en p. 48, 49, 58]

L. Mascarilla and C. Frélicot. A class of reject-first possibilistic classifiers based on dual triples. In *Proceedings of the 9th International Fuzzy Systems Association World Congress*, pages 743–747, 2001. [cité en p. 38, 105]

L. Mascarilla, M. Berthier, and C. Frélicot. A k-order fuzzy or operator for pattern classification with k-order ambiguity rejection. *Fuzzy Sets and Systems*, 159(15):2011–2029, 2008. [cité en p. 16, 26, 27, 28, 104, 131, 136]

M. Masson and T. Denoeux. Ecm: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384–1397, 2008. [cité en p. 138]

G. Mayor and J. Torrens. On a family of t-norms. *Fuzzy Sets and Systems*, 41(2):161–166, 1991. [cité en p. 188]

K. Menger. Statistical metrics. *Proc. National Academy of Science USA*, 28(12):535–537, 1942. [cité en p. 16]

J. Merigó and A. Gil-Lafuente. The induced generalized owa operator. *Information Sciences*, 179(6):729–741, 2009. [cité en p. 15]

R. Mesiar and A. Mesiarova. Fuzzy integrals – what are they? *International Journal of Intelligent Systems*, 23(2):199–212, 2008. [cité en p. 25]

P. Miranda, M. Grabisch, and P. Gil. p-symmetric fuzzy measures. *Int. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 10(supplement):105–123, 2002. [cité en p. 23]

H. Mouchere and E. Anquetil. A unified strategy to deal with different natures of reject. In *18th International Conference on Pattern Recognition*, 2006. [cité en p. 103]

T. Murofushi and S. Soneda. Techniques for reading fuzzy measures (iii): interaction index. In *9th Fuzzy System Symposium*, pages 693–696, Saporo, Japon, 1993. [cité en p. 23]

T. Murofushi and M. Sugeno. Fuzzy t-conorm integrals with respect to fuzzy measures: generalization of sugeno integral and choquet integral. *Fuzzy Sets and Systems*, 42:57–71, 1991. [cité en p. 25]

R. Muzzolini, Y. Yang, and R. Pierson. Classifier design with incomplete knowledge. *Pattern Recognition*, 31(4):345–369, 1998. [cité en p. 101]

Y. Narukawa and V. Torra. Fuzzy measure and probability distributions: distorted probabilities. *IEEE Transactions on Fuzzy Systems*, 13(5):617–629, 2005. [cité en p. 23]

Y. Narukawa and V. Torra. Multidimensional generalized fuzzy integral. *Fuzzy Sets and Systems*, 160(6):802–815, 2009. [cité en p. 25]

R. Nelsen. *An introduction to copulas*. Springer, 2006. [cité en p. 10, 32, 184]

L. Oliveira, A. Britto, and R. Sabourin. Improving cascading classifiers with particle swarm optimization. In *ICDAR*, pages 570–574, 2005. [cité en p. 102]

M. Pakhira, S. Bandyopadhyay, and U. Maulik. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(3):487–501, 2004. [cité en p. 130]

N. Pal and J. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*, 3(3):370–379, 1995. [cité en p. 125, 128, 131]

B. Papadopoulos, G. Trasanides, and A. Hatzimichailidis. Optimization method for the selection of the appropriate fuzzy implication. *Journal of Optimization Theory and Applications*, 134(1):135–141, 2007. [cité en p. 69]

C. Pappis and N. Karacapilidis. A comparative assessment of measures of similarity of fuzzy values. *Fuzzy Sets and Systems*, 56(2):171–174, 1993. [cité en p. 65]

F. Pitrelli, J. Subrahmonia, and P. Perrone. Confidence modeling for handwriting recognition: algorithms and applications. *International Journal on Document Analysis and Recognition*, 8(1):35–46, 2006. ISSN 1433-2833. [cité en p. 102]

A. Pradera, E. Trillas, and T. Calvo. A general class of triangular norm-based aggregation operators: quasi-linear t-s operators. *International Journal of Approximate Reasonning*, 30(1):57–72, 2002. [cité en p. 175]

P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994. [cité en p. 76]

M. R. Rezaee, B. Lelieveldt, and J. Reiber. A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters*, 19(3-4):237–246, 1998. [cité en p. 128]

M. Rifqi, V. Berger, and B. Bouchon-Meunier. Discrimination power of measures of comparison. *Fuzzy Sets and Systems*, 110(2):189–196, 2000. [cité en p. 55]

B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996. [cité en p. 80]

M. Roubens. Pattern classification problems and fuzzy sets. *Fuzzy Sets and Systems*, 1(4): 239–253, 1978. [cité en p. 129]

M. Saar-Tsechansky and F. Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8:1625–1657, 2007. [cité en p. 92]

R. Sabbadin. *Une approche ordinale de la décision dans l'incertain : Axiomatisation, représentation logique et application à la décision séquentielle*. PhD thesis, Université Paul Sabatier, Toulouse, 1998. [cité en p. 24]

C. Sansone, F. Tortorella, and M. Vento. A classification reliability driven reject rule for multi-expert systems. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(6):1–19, 2001. [cité en p. 103]

S. Santini and R. Jain. Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):871–883, 1999. [cité en p. 57]

C. Santo-Pereira and A. Pires. On optimal reject rules and roc curves. *Pattern Recognition Letters*, 26(7):943–952, 2004. [cité en p. 114]

B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. [cité en p. 77]

V. Schroeder. Quasi-metric and metric spaces. *Conformal Geometry and Dynamics*, 10: 355–360, 2006. [cité en p. 56]

B. Schweizer and A. Sklar. *Probabilistic Metric Spaces*. North-Holland, Amsterdam, 1983. [cité en p. 10, 16, 189]

D. Semani. *Une méthode supervisée de selection et de discrimination avec rejet - Application au projet Aquathèque*. PhD thesis, Université de La Rochelle, 2004. [cité en p. 76]

D. Semani, C. Frélicot, and L. Mascarilla. Discrimination possibiliste avec options de rejet : une nouvelle approche. In *12èmes Rencontres Francophones sur la Logique Floue et ses Applications, LFA*, 2002. [cité en p. 105]

D. Semani, C. Frélicot, and P. Courtellemont. Combinaison d'étiquettes floues/possibilistes pour la sélection de variables. In *14 ème Congrès de Reconnaissance de Formes et Intelligence Artificielle, RFIA*, volume 2, pages 479–488, 2004. [cité en p. 152]

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976. [cité en p. 23]

C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423, 623–656, 1948. [cité en p. 36]

L. Shapley. A value for n-person games. In H. Kuhn and A. Tucker, editors, *Contribution to the Theory of Games, vol. II*, pages 303–317. Princeton University Press, 1953. [cité en p. 23]

H. Shi, P. Gader, and W. Chen. Fuzzy integral filters: properties and parallel implementation. *Real-Time Imaging*, 4(4):233–241, 1998. [cité en p. 24]

N. Shilkret. Maxitive measure and integration. *Indagatione Math.*, 33:109–116, 1971. [cité en p. 25]

W. Silvert. Symmetric summation: a class of operations on fuzzy sets. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(10):657–659, 1979. [cité en p. 13, 176]

D. Sinha and E. R. Dougherty. Fuzzification of set inclusion : theory and applications. *Fuzzy Sets and Systems*, 55(1):15–42, 1993. [cité en p. 64, 68]

P. Smyth. Probability density estimation and local basis function neural networks. In *Proceedings of the workshop on Computational learning theory and natural learning systems (vol. 2) : intersections between theory and experiment*, pages 233–248, Cambridge, MA, USA, 1994. MIT Press. [cité en p. 105]

M. Sugeno. *Theory of fuzzy integrals and its applications*. PhD thesis, Tokyo Institute of Technology, 1974. [cité en p. 21, 22, 24, 192]

H. Tahani and J. Keller. Information fusion in computer vision using the fuzzy integral. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(3):733–741, 1990. [cité en p. 24, 86]

T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics*, 15:116–132, 1985. [cité en p. 85]

D. Tax and R. Duin. Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29(10):1565–1570, 2008. [cité en p. 104]

A. Temko, D. Macho, and C. Nadeu. Fuzzy integral based information fusion for classification of highly confusable non-speech sounds. *Pattern Recognition*, 41(5):1831–1840, 2008. [cité en p. 24]

S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 2006. [cité en p. 74, 75]

Y. Tolias, S. Panas, and L. Tsoukalas. Generalized fuzzy indices for similarity matching. *Fuzzy Sets and Systems*, 120(2):255–270, 2001. [cité en p. 57]

V. Torra. The weighted owa operator. *International Journal of Intelligent Systems*, 12: 153–166, 1997. [cité en p. 15]

V. Torra and Y. Narukawa. The interpretation of fuzzy integrals and their application to fuzzy systems. *International Journal of Approximate Reasonning*, 41(1):43–58, 2006. [cité en p. 24]

V. Torra and Y. Narukawa. *Modeling Decisions: Information Fusion and Aggregation Operators*. Springer, 2007. [cité en p. 10]

F. Tortorella. A roc-based reject rule for dichotomizers. *Pattern Recognition Letters*, 26: 167–180, 2005. [cité en p. 115]

E. Trillas and C. Alsina. Logic: going farther from tarski? *Fuzzy Sets and Systems*, 53(1): 1–13, 1993. [cité en p. 22]

I. Turksen. Interval-valued fuzzy sets and "compensatory and". *Fuzzy Sets Systems*, 51(3): 295–307, 1992. [cité en p. 13]

I. Turksen, V. Kreinovich, and R. Yager. A new class of fuzzy implications. axioms of fuzzy implication revisited. *Fuzzy Sets and Systems*, 100(1-3):267–272, 1998. [cité en p. 49]

A. Tversky. Features of similarity. *Psychological review*, 84(4):327–352, 1977. [cité en p. 55, 57, 67]

W. Wang and Y. Zhang. On fuzzy cluster validity indices. *Fuzzy Sets and Systems*, 158 (19):2095–2117, 2007. [cité en p. 125, 128]

W.-J. Wang. New similarity measures on fuzzy sets and on elements. *Fuzzy Sets and Systems*, 85:305–309, 1997. [cité en p. 65, 67]

X. Wang, B. De Baets, and E. Kerre. A comparative study of similarity measures. *Fuzzy Sets and Systems*, 73(22):259–268, 1995. [cité en p. 56]

A. Webb. *Statistical Pattern Recognition*. John Wiley & Sons Inc., 2002. [cité en p. 80]

S. Weber. A general concept of fuzzy connectives, negations, and implications based on t-norms and t-conorms. *Fuzzy Sets and Systems*, 11:115–134, 1983. [cité en p. 192]

T. Whalen. Parameterized r-implications. *Fuzzy Sets and Systems*, 134(2):231–281, 2003. [cité en p. 49]

M. Windham. Numerical classification of proximity data with assignment measure. *Journal of Classification*, 2:157–172, 1985. [cité en p. 138]

K. Wu and M. Yang. A cluster validity index for fuzzy clustering. *Pattern Recognition Letters*, 26(9):1275–1291, 2005. [cité en p. 130]

K.-L. Wu, M.-S. Yang, and J.-N. Hsieh. Robust cluster validity indexes. *Pattern Recognition*, 42(11):2541–2550, 2009. [cité en p. 135, 143]

X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991. [cité en p. 129, 140]

Z. Xu. An overview of methods for determining owa weights. *International Journal of Intelligent Systems*, 20(8):843–865, 2005. [cité en p. 15]

Z. Xu. Dependent owa operators. In *3rd Intl. Conf. Modeling Decisions for Artificial Intelligence*, pages 172–178, 2006. [cité en p. 15]

Z. Xu and Q. Da. The uncertain owa operator. *International Journal of Intelligent Systems*, 17(6):569–575, 2002. [cité en p. 15]

R. Yager. Heavy owa operators. *Fuzzy Optimization and Decision Making*, 1(4):379–397, 2002. [cité en p. 15]

R. Yager. Noble reinforcement in disjunctive aggregation operators. *IEEE Transactions on Fuzzy Systems*, 11(6):754–767, 2003. [cité en p. 13]

R. Yager. Generalized owa aggregation operators. *Fuzzy Optimization and Decision Making*, 3(1):93–107, 2004. [cité en p. 15]

R. Yager. Aggregation of ordinal information. *Fuzzy Optimization and Decision Making*, 6 (3):199–219, 2007. [cité en p. 10]

R. Yager. Measures of specificity over continuous spaces under similarity relations. *Fuzzy Sets and Systems*, 159(17):2193–2210, 2008. [cité en p. 37, 38]

R. Yager. On the general class of fuzzy connectives. *Fuzzy Sets and Systems*, 4(3):235–242, 1980. [cité en p. 21, 193]

R. Yager. Measuring tranquility and anxiety in decision making: an application of fuzzy sets. *International Journal of General Systems*, 9(3):249–260, 1982. [cité en p. 37]

R. Yager. Ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 18:183–190, 1988. [cité en p. 14, 15]

R. Yager. Ordinal measures of specificity. *International Journal of General Systems*, 17(1): 57–72, 1990. [cité en p. 37]

R. Yager and D. Filev. Induced ordered weighted averaging operators. *IEEE Transactions on Systems, Man, and Cybernetics–Part B*, 29(2):141–150, 1999. [cité en p. 15]

R. Yager and J. Kacprzyk. *The Ordered Weighted Averaging Operation: Theory, Methodology and Applications*. Kluwer: Norwell, 1997. [cité en p. 14]

R. Yager and A. Rybalov. Uninorm aggregation operators. *Fuzzy Sets and Systems*, 80(1): 111–120, 1996. [cité en p. 173]

R. Yager and A. Rybalov. Full reinforcement operators in aggregation techniques. *IEEE Transactions on Systems, Man and Cybernetics*, 28(6):757–769, 1998. [cité en p. 13]

V. R. Young. Fuzzy subsethood. *Fuzzy Sets and Systems*, 77(3):371–384, 1996. [cité en p. 59, 64, 67]

L. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1): 3–28, 1978. [cité en p. 178]

L. Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, 90(2):111–127, 1997. [cité en p. 37]

M. H. F. Zarandi, E.Neshat, and I. B. Türksen. A new cluster validity index for fuzzy clustering based on similarity measure. In *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, 11th International Conference*, pages 127–135, 2007. [cité en p. 128, 133]

W. Zeng and H. Li. Inclusion measures, similarity measures, and the fuzziness of fuzzy sets and their relations. *International Journal of Intelligent Systems*, 21(6):639–653, 2006. [cité en p. 55, 56]

H.-Y. Zhang and W.-X. Zhang. Hybrid monotonic inclusion measure and its use in measuring similarity and distance between fuzzy sets. *Fuzzy Sets and Systems*, 160(1):107–118, 2009. [cité en p. 59, 69]

H.-J. Zimmermann and P. Zysno. Latent connectives in human decision making. *Fuzzy Sets and Systems*, 4:37–51, 1980. [cité en p. 13, 175]

H.-J. Zimmermann and P. Zysno. Quantifying vagueness in decision models. *European Journal of Operational Research*, 22(2):148–158, 1985. [cité en p. 99]

Troisième partie

Annexes

Annexe A

Opérateurs d'agrégation

A.1 Propriétés

A.1.1 Idempotence

En algèbre, l'idempotence est une propriété liée à une opération $*$ par laquelle un élément x est idempotent, c'est à dire $x * x = x$. Dans le cadre des opérateurs n-aires, cette notion s'étend à la propriété suivante.

Définition A.1. Un opérateur d'agrégation $\mathcal{A}$ possède un élément idempotent x si

$$\mathcal{A}(x, \cdots, x) = x \tag{A.1}$$

On dira qu'un opérateur d'agrégation est idempotent si la propriété (A.1) est respectée pour tout $x \in I$.

Cette propriété, aussi appelée unanimité, s'interprète de la manière suivante : si on agrège n fois la même valeur, alors on s'attend à obtenir cette même valeur initiale. On voit aisément que si les conditions aux bornes (1.2) et (1.3) sont respectées, alors 0 et 1 sont des éléments idempotents, que l'on appellera triviaux. Désirer l'idempotence d'un opérateur ainsi que sa monotonie revient à vouloir un comportement de compensation, et l'on se restreindra dans ce cas aux opérateurs respectant $\min \leq \mathcal{A} \leq \max$.

A.1.2 Continuité

Un opérateur d'agrégation est dit continu si la fonction d'agrégation est continue dans le sens usuel de ce terme, c'est à dire que pour tout $n \in \mathbb{N}$, l'opération n-aire $\mathcal{A}$ est continue. Cette propriété est souvent nécessaire dans de nombreuses applications, puisque elle contraint l'opérateur à ne pas se comporter de manière chaotique. En effet, la continuité assure qu'un changement, ou une erreur faible en entrée ne mènera pas à un grand changement, ou erreur, à la sortie.

A.1.3 Symétrie

La symétrie, ou la commutativité, d'opérateur binaires, c'est à dire la propriété $x*y = y*x$ peut facilement être étendue au cas d'opérations n-aires. Lorsque $n > 2$, on parle alors de

symétrie.

Définition A.2. On dit qu'un opérateur d'agrégation $\mathcal{A}$ est symétrique si

$$\mathcal{A}(x_1,\cdots,x_n) = \mathcal{A}(x_{\sigma(1)},\cdots,x_{\sigma(n)}) \tag{A.2}$$

pour tout permutation σ de l'ensemble $N = \{1,\cdots,n\}$.

Les opérations maximum, minimum, ou encore les moyennes sont symétriques, mais nous verrons en section 1.2 que les moyennes pondérées ne le sont pas. Définir un opérateur d'agrégation symétrique revient à associer à chacune des valeurs d'entrée la même importance, ce qui pousse certains auteurs à qualifier cette propriété d'*anonymat*.

A.1.4 Associativité

L'associativité d'une opération binaire $*$ définie sur un domaine I signifie que le couple $(I,*)$ est un demi-groupe.

Définition A.3. Un opérateur d'agrégation $\mathcal{A}$ est associatif si

$$\begin{gathered}\forall\, n,n' \in \mathbb{N},\forall\, x_1,\cdots,x_n \quad et \quad y_1,\cdots,y_{n'} \in I,\\ \mathcal{A}(x_1,\cdots,x_n,y_1,\cdots,y_{n'}) = \mathcal{A}(\mathcal{A}(x_1,\cdots,x_n),\mathcal{A}(y_1,\cdots,y_{n'})).\end{gathered} \tag{A.3}$$

Cette propriété est intéressante dans la mesure où l'on peut commencer à agréger les valeurs sans connaître l'ensemble des arguments d'entrée, ce qui peut être une contrainte d'un système en ligne par exemple. Pratiquement, on divise ainsi les entrées en tuples que l'on agrège, et l'ordre d'apparition de ces tuples lors de l'opération d'agrégation ne doit pas changer le résultat final. À partir d'un opérateur associatif défini pour deux opérandes, il est donc aisé de l'étendre à son équivalent n-aire. Un exemple d'opérateur associatif est le produit. Inversement, la moyenne arithmétique n'est pas associative. On pourra se référer à [Alsina et al., 2006] pour une étude détaillée des fonctions associatives.

A.1.5 Élément neutre et élément absorbant

Venant encore une fois d'une notion connue dans le cadre des opérations binaires, un élément e est appelé élément neutre de l'opération $*$ si pour tout x, on a $x*e = e*x = x$. À partir de cette définition, on pourra donc dire que l'action d'un élément neutre a le même effet que s'il n'était pas présent.

Définition A.4. Un opérateur d'agrégation $\mathcal{A}$ possède un élément neutre e si

$$\begin{gathered}\forall\, n \in \mathbb{N},\forall\, x_1,\cdots,x_n \in I, \exists i \text{ tel que } x_i = e, \text{ alors}\\ \mathcal{A}(x_1,\cdots,x_n) = \mathcal{A}(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n).\end{gathered} \tag{A.4}$$

Inversement, un élément a est appelé élément absorbant de l'opération $*$ si pour tout x, on a $x*a = a*x = a$. On transpose alors cette définition à l'opérateur d'agrégation n-aire de la façon suivante.

Définition A.5. Un opérateur d'agrégation $\mathcal{A}$ possède un élément absorbant a si

$$\begin{gathered}\forall n \in \mathbb{N},\forall\, x_1,\cdots,x_n \in I, \exists i \text{ tel que } x_i = a, \text{ alors}\\ \Downarrow\\ \mathcal{A}(x_1,\cdots,x_n) = a.\end{gathered} \tag{A.5}$$

C'est l'attitude contraire à l'élément neutre, dans la mesure où la simple présence de cet élément parmi les valeurs d'entrée suffit à provoquer une sortie respectant celui-ci. Cela implique donc ce comportement, ou vote, que l'on qualifiera de *veto*.

A.2 Combinaisons

A.2.1 Uninormes et nullnormes

Les t-normes et t-conormes fournissent un renforcement par le bas et par le haut, respectivement, mais pas de manière conjointe. Afin de remédier à ce problème, Yager et Rybalov [Yager and Rybalov, 1996] proposent un nouvel opérateur d'agrégation, l'uninorme $\mathcal{U}$. Cet opérateur est une généralisation des deux précédents, dans la mesure où l'élément neutre e peut être fixé dans l'intervalle unité de manière libre.

Définition A.6. Une uninorme est une opération binaire $\mathcal{U}$ commutative, associative et croissante possédant un élément neutre e appartenant à l'intervalle unité, c'est à dire que pour tout $x \in [0,1]$, on a $\mathcal{U}(x,e) = x$.

Le point intéressant des uninormes est qu'elles permettent à des valeurs séparées par l'élément neutre de se compenser. Le lien entre les uninormes et les normes triangulaires est évidemment important, et l'on peut même écrire une t-norme sous la forme

$$\top_{\mathcal{U}}(x,y) = \frac{\mathcal{U}(ex,ey)}{e} \tag{A.6}$$

et alternativement, l'opération

$$\perp_{\mathcal{U}}(x,y) = \frac{\mathcal{U}(e+(1-e)x,e+(1-e)y)-e}{1-e} \tag{A.7}$$

est une t-conorme [Fodor et al., 1997]. La structure des uninormes sur $[0,e]^2$ est donc fortement liée aux t-normes, tandis qu'elle est liée aux t-conormes sur $[e,1]^2$. Sur le reste du carré unité, $\mathcal{U}$ est bornée par le minimum et le maximum, c'est à dire que pour tout $(x,y) \in [0,1]^2 \backslash ([0,e]^2 \cup [e,1]^2)$, $\mathcal{U}$ sera un opérateur de compensation au sens de (1.8). Comme $\mathcal{U}$ est une opération associative, on a $\mathcal{U}(0,1) \in \{0,1\}$. On appelle uninorme conjonctive une uninorme $\mathcal{U}$ telle que $\mathcal{U}(0,1) = 0$, et une uninorme disjonctive $\mathcal{U}(0,1) = 1$. Les transformations définies par (A.8) et (A.9) forment deux classes générales d'opérateurs.

$$\mathcal{U}(x,y) = \begin{cases} e\top(\frac{x}{e},\frac{y}{e}) & \text{si } (x,y) \in [0,e[^2 \\ e+(1-e)\perp\left(\frac{x-e}{1-e},\frac{y-e}{1-e}\right) & \text{si } (x,y) \in [e,1]^2 \\ \min(x,y) & \text{sinon} \end{cases} \tag{A.8}$$

$$\mathcal{U}(x,y) = \begin{cases} e\top(\frac{x}{e},\frac{y}{e}) & \text{si } (x,y) \in [0,e]^2 \\ e+(1-e)\perp\left(\frac{x-e}{1-e},\frac{y-e}{1-e}\right) & \text{si } (x,y) \in]e,1]^2 \\ \max(x,y) & \text{sinon} \end{cases} \tag{A.9}$$

Dans (A.8), on note que $\mathcal{U}(0,1) = 0$ ce qui implique que $\mathcal{U}$ est une uninorme conjonctive, tandis que dans (A.9), on a $\mathcal{U}(0,1) = 1$ ce qui rend $\mathcal{U}$ disjonctive. La Fig. A.1 présente une visualisation de la structure d'une uninorme, où $\top^\star$ et $\perp^\star$ sont donnés par:

$$\top^\star(x,y) = e\top(\frac{x}{e},\frac{y}{e}) \tag{A.10}$$

$$\perp^\star(x,y) = e+(1-e)\perp(\frac{x-e}{1-e},\frac{y-e}{1-e}) \tag{A.11}$$

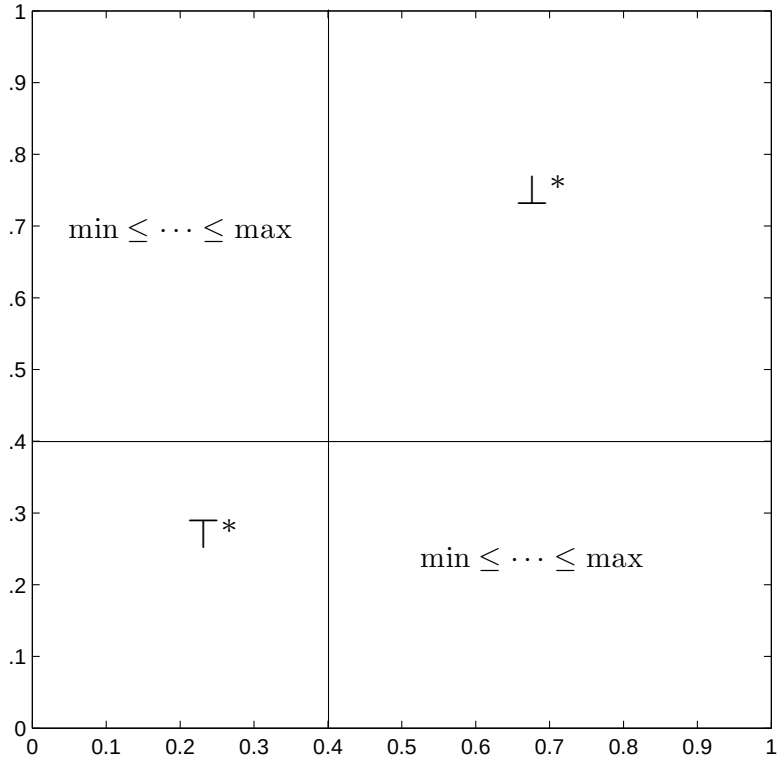

FIG. A.1: Structure d'une uninorme ayant pour élément neutre $e = 0.40$.

Les propriétés de $\mathcal{U}$, $\top$ et $\bot$ permettent l'extension de chaque uninorme à son opérateur n-aire [Calvo and Mesiar, 2003] :

$$\mathcal{U}(x_1,\cdots,x_n) = \mathcal{U}\Big(\top^\star(min(x_1,e),\cdots,min(x_n,e)),\\ \bot^\star(max(x_1,e),\cdots,max(x_n,e))\Big)$$

L'effet de compensation pour des valeurs séparées par l'élément neutre apparaît pour la classe des uninormes archimédiennes [Dombi, 1982b].

Définition A.7. Un opérateur d'agrégation $\mathcal{U}$ est une uninorme archimédienne continue si en tout point de $(x_1,\cdots,x_n) \in \bigcup_{n\in\mathbb{N}}[0,1]^n$, $\{0,1\} \subset \{x_1,\cdots,x_n\}$, si et seulement si il existe une bijection monotone $g : [0,1] \to [-\infty,\infty]$ avec $g(e) = 0$ telle que

$$\mathcal{U}(x_1,\cdots,x_n) = g^{-1}\Big(\sum_{i=1}^{n} g(x_i)\Big) \tag{A.12}$$

On dira alors que $\mathcal{U}$ une uninorme engendrée par un générateur additif g et élément neutre e.

Cette uninorme peut être reliée aux moyennes quasi-arithmétiques, mis à part le fait que ces dernières ne possèdent pas d'élément neutre.

Nous avons vu que les uninormes étaient fortement liées aux sommes ordinales sur deux régions différentes. La contrepartie de cet opérateur sur I est la nullnorme $\mathcal{V}$ [Calvo et al., 2001].

Définition A.8. Une nullnorme est une opération commutative, associative et croissante,

admettant un élément absorbant $a \in [0,1]$, et qui satisfait $\mathcal{V}(x,0) = x$ pour tout $x \leq a$, et $\mathcal{V}(x,1) = x$ pour tout $x \geq a$.

Comme pour les uninormes, on peut démontrer qu'une nullnorme peut s'écrire de la façon suivante

$$\mathcal{V}(x,y) = \begin{cases} a\bot(\frac{x}{a},\frac{y}{a}) & \text{si } (x,y) \in [0,a]^2 \\ a + (1-a)\top\left(\frac{x-a}{1-a},\frac{y-a}{1-a}\right) & \text{si } (x,y) \in [a,1]^2 \\ a & \text{sinon} \end{cases} \tag{A.13}$$

De la même manière que l'on peut obtenir une uninorme à partir de fonctions génératrices, ceci est également possible pour les nullnormes. On peut ainsi trouver une nullnorme sous certaines conditions :

Définition A.9. Un opérateur $\mathcal{V}$ est une nullnorme continue nilpotente avec élément absorbant $a \in]0,1[$ si et seulement si il existe une bijection croissante $q : [0,1] \to [0,1]$ telle que

$$\mathcal{V}(x_1,\cdots,x_n) = q^{-1}\Big(\text{med}(\sum_{i=1}^{n} q(x_i), \sum_{i=1}^{n} q(x_i) - (n-1), q(a))\Big) \tag{A.14}$$

où on dit que $\mathcal{V}$ est une nullnorme nilpotente si pour tout $x \in [0,1]$, il existe un k entier tel que

$$\mathcal{V}(\underbrace{x,\cdots,x}_{k \text{ fois}}) \in \{0,a,1\}. \tag{A.15}$$

A.2.2 Convexes linéaires et exponentielles

De nombreux opérateurs utilisent des combinaisons de t-normes et t-conormes. Comme on a pu le voir lorsque l'on a évoqué les uninormes, ces opérateurs ne fournissent pas de renforcement bilatéral. C'est pourquoi de nombreuses propositions s'appuyant sur des combinaisons convexes de normes triangulaires sont apparues. On distingue deux catégories de combinaisons : les combinaisons convexes linéaires et exponentielles, introduites par Zimmermann et Zysno [Zimmermann and Zysno, 1980]. Dans un premier temps, proposées pour des normes triangulaires particulières, elles ont été ensuite généralisées pour n'importe quelle norme triangulaire. En effet l'opérateur Γ est défini de la manière suivante

$$\Gamma(x_1,\cdots,x_n) = \Big(\prod_{i=1}^{n} x_i\Big)^{1-\gamma}\Big(1 - \prod_{i=1}^{n}(1-x_i)\Big)^{\gamma} \tag{A.16}$$

qui est en fait une sous-classe des opérateurs de combinaison convexe exponentielle, c'est à dire d'une moyenne géométrique pondérée de $\top$ et $\bot$, définis par

$$\mathcal{E}(x_1,\cdots,x_n) = \top(x_1,\cdots,x_n)^{1-\gamma}\bot(x_1,\cdots,x_n)^{\gamma} \tag{A.17}$$

pour laquelle on obtient (A.16) si l'on utilise le couple $(\top,\bot)_A$. L'autre approche, également proposée dans [Zimmermann and Zysno, 1980], puis étendue, repose sur le concept de moyenne arithmétique pondérée. Ainsi, la combinaison convexe linéaire est définie par

$$\mathcal{L}(x_1,\cdots,x_n) = (1-\gamma)\cdot\top(x_1,\cdots,x_n) + \gamma\cdot\bot(x_1,\cdots,x_n) \tag{A.18}$$

Plus récemment, Pradera et al. [Pradera et al., 2002] proposent une généralisation de ces deux propositions. Pour cela, ils s'appuient sur une méthode de composition, et appliquent

la définition générale d'une moyenne quasi-arithmétique pondérée donnée par (1.11). Ils obtiennent ainsi la définition suivante des opérateurs $\top - \bot$ quasi-linéaires :

$$\mathcal{QL}(x_1, \cdots, x_n) = f^{-1}\Big((1-\gamma) \cdot f(\top(x_1, \cdots, x_n)) + \gamma \cdot f(\bot(x_1, \cdots, x_n))\Big) \tag{A.19}$$

Évidemment, $\mathcal{E}$ et $\mathcal{L}$ sont des cas particuliers de $\mathcal{QL}$. Ces trois opérateurs présentent un comportement de compensation entre les valeurs agrégées, à distinguer, rappelons-le, d'un comportement de compromis.

A.2.3 Sommes symétriques

Les opérateurs auto-duals ont été étudiés dans [Calvo et al., 2002], où un tel opérateur est construit comme la moyenne entre un opérateur d'agrégation et son dual. Un cadre plus général est proposé plus récemment par Maes et al. [Maes et al., 2005]. Initialement, Silvert [Silvert, 1979] avait introduit des opérations permettant de fusionner deux ensembles flous de manière à ce que le complément de la combinaison soit la combinaison du complément, en d'autres termes,

$$1 - \mathcal{SS}(x_1, \cdots, x_n) = \mathcal{SS}(1-x_1, \cdots, 1-x_n) \tag{A.20}$$

où l'opérateur $\mathcal{SS}$ est appelé somme symétrique. Silvert montrera ainsi que les sommes symétriques sont des opérations continues, non décroissantes, commutatives que l'on peut écrire sous la forme

$$\mathcal{SS}(x,y) = \frac{f(x,y)}{f(x,y) + f(1-x, 1-y)} \tag{A.21}$$

où f est une fonction continue croissante respectant $f(0,0) = 0$. Dans l'esprit des combinaisons convexes décrites à la section précédente, cette fonction f peut être une t-norme ou une t-conorme. On pourra ainsi remarquer que si l'on prend $f(x,y) = x \cdot y$, on obtient l'opérateur de renforcement complet défini dans l'exemple 1.2.

Annexe B

Théorie des ensembles flous

Cette annexe introduit de façon assez brève des éléments et définitions de la théorie des ensembles flous. Les propriétés de la théorie classique des ensembles seront utilisés afin d'introduire leurs équivalents flous. Ce chapitre ne se voulant absolument pas exhaustif, le lecteur pourra se référer à [Dubois and Prade, 1980; Bezdek et al., 1999a; Dubois and Prade, 2000] pour plus d'informations. Par la suite, nous dénoterons

- $X = \{x_1, \cdots, x_n\}$ le (supposé fini) univers de discours,
- $\mathcal{C}(X)$ et $\mathcal{F}(X)$ les ensembles de tous les ensembles stricts et flous de X, respectivement,
- $f_A(x)$, $\forall x \in X$, la fonction d'appartenance d'un ensemble flou A sur X.

Dans un ensemble classique $A \subset X$, A dénote une collection d'éléments. Les éléments $x \in X$ peuvent ou non appartenir à l'ensemble A. Contrairement à l'ensemble classique, l'ensemble flou utilise une fonction d'appartenance pour chaque x, dénotant ainsi à quel point (degré) l'élément x appartient à A.

Définition B.1. Un ensemble flou A sur X est défini par la donnée d'une application

$$f_A : X \to I \tag{B.1}$$

$$f_A(x) \mapsto [0, 1] \tag{B.2}$$

où $f_A(x)$ s'interprète comme le degré d'appartenance de x à A.

Evidemment, cette définition est la simple extension de la définition stricte, où $f_A(x)$ peut prendre les valeurs 0 ou 1.

Définition B.2. Le **support** d'un ensemble flou A est l'ensemble des éléments pour lesquels le degré d'appartenance n'est pas nul

$$Supp(A) = \{x \in X : f_A(x) > 0\}$$

Définition B.3. Le **noyau** d'un ensemble flou A est l'ensemble des éléments pour lesquels le degré d'appartenance vaut 1

$$Noyau(A) = \{x \in X : f_A(x) = 1\}$$

Définition B.4. Un ensemble flou A est dit **normal** lorsque son noyau est non-vide, en d'autres termes, on peut trouver un x tel que $f_A(x) = 1$.

Définition B.5. L'α-coupe d'un ensemble A est défini par

$$A_\alpha = \{x \in X : f_A(x) \geq \alpha\}$$

Si l'on utilise l'inégalité stricte, alors l'α-coupe sera dite stricte.

On peut dès lors formuler les opérations ensemblistes floues usuelles sur deux ensembles flous A et B :

Egalité	$A = B$	$\forall x \in X, f_A(x) = f_B(x)$	(B.3)
Inclusion	$A \subseteq B$	$\forall x \in X, f_A(x) \leq f_B(x)$	(B.4)
Intersection	$A \cap B$	$\forall x \in X, f_{A\cap B}(x) = \min(f_A(x), f_B(x))$	(B.5)
Union	$A \cup B$	$\forall x \in X, f_{A\cap B}(x) = \max(f_A(x), f_B(x))$	(B.6)
Complémentaire	$\bar{A}$	$\forall x \in X, f_{\bar{A}}(x) = 1 - f_A(x)$	(B.7)

De manière plus générale, l'union et l'intersection d'ensembles flous peuvent être définies via les normes triangulaires. On peut donc modifier les équations (B.5-B.6) et obtenir

Intersection	$A \cap B$	$\forall x \in X, f_{A\cap B}(x) = \top(f_A(x), f_B(x))$	(B.8)
Union	$A \cup B$	$\forall x \in X, f_{A\cap B}(x) = \bot(f_A(x), f_B(x))$	(B.9)

Les deux lois fondamentales en théorie ensembliste classique sont le principe du tiers exclu ($A \cup \bar{A} = X$) et le principe de non-contradiction ($A \cap \bar{A} = \emptyset$). Ces deux principes ne sont pas respectés en théorie des ensembles flous.

Nous terminerons ce bref aperçu de la théorie des ensembles flous en présentant rapidement une dérivée de celle-ci : la théorie des possibilités [Zadeh, 1978; Dubois and Prade, 1988]. Si nous considérons un ensemble de référence X, la croyance en un événement défini sur X est complètement déterminée par la connaissance des degrés de nécessité et de possibilité.

Définition B.6. Une mesure de possibilité Π est une fonction définie sur l'ensemble des parties de X, $\mathcal{P}(X)$, à valeurs dans I telle que

$$\Pi(X) = 1, \qquad \Pi(\emptyset) = 0 \tag{B.10}$$

$$\forall A, B \in X, \Pi(A \cup B) = \max(\Pi(A), \Pi(B)) \tag{B.11}$$

Définition B.7. Une mesure de nécessité N est une fonction définie sur l'ensemble des parties de X, $\mathcal{P}(X)$, à valeurs dans I telle que

$$N(X) = 1, \qquad N(\emptyset) = 0 \tag{B.12}$$

$$\forall A, B \in X, N(A \cap B) = \min(N(A), N(B)) \tag{B.13}$$

Par analogie aux fonctions d'appartenance des ensembles flous, on peut utiliser une fonction attribuant un degré de possibilité à tout élément x de X.

Définition B.8. Une distribution de possibilité est une fonction $\pi : X \to I$ telle qu'il existe un élément $x \in X$ pour lequel $\pi(x) = 1$: $\sup_{x\in X} \pi(x) = 1$.

Annexe C

Familles de Normes Triangulaires

Nous présentons dans cette section des informations liées aux familles de normes triangulaires paramétriques que l'on retrouve régulièrement dans la littérature, et qui n'auront, par manque de place et par mesure de concision, pas été développées dans le corps principal de ce manuscrit. Pour chaque couple, nous introduirons leur définition, ainsi qu'une visualisation sur le carré unité. Cette approche graphique permettra ainsi une meilleure compréhension de l'importance, et des particularités qu'induisent l'introduction d'un paramètre dans ces couples. Pour l'ensemble des iso-surfaces représentées dans cette section, la couleur rouge correspond à une valeur importante, et la couleur bleue à une valeur faible.

C.1 Aczél-Alsina

Initialement introduite dans [Aczél and Alsina, 1984], cette famille est la seule pour laquelle quels que soient p,q dans $\lambda \in]0,1[\cup]0,\infty[$, le nombre $\frac{\log p}{\log q}$ est irrationnel, et pour tout $(a,b) \in [0,1]^2$, on a $\top(a^p,b^p) = \top(a,b)^p$, de même que $\top(a^q,b^q) = \top(a,b)^q$. La norme triangulaire d'Aczél-Alsina $\top_{AA_\lambda}$ est définie par

$$\top_{AA_\lambda}(a,b) = \begin{cases} \top_D(a,b) & \text{si } \lambda = 0 \\ \top_M(a,b) & \text{si } \lambda = \infty \\ \exp\Big(- \big((\log a)^\lambda + (\log b)^\lambda\big)^{1/\lambda}\Big) & \text{si } \lambda \in]0,\infty[\end{cases}$$

où $\lambda \in [0, +\infty]$, voir la première ligne de la Fig. C.1.
La conorme triangulaire d'Aczél-Alsina $\perp_{AA_\lambda}$ est définie par

$$\perp_{AA_\lambda}(a,b) = \begin{cases} \perp_D(a,b) & \text{si } \lambda = 0 \\ \perp_M(a,b) & \text{si } \lambda = \infty \\ 1 - \exp\Big(- \big(- (\log(1-a))^\lambda + (-\log(1-b))^\lambda\big)^{1/\lambda}\Big) & \text{si } \lambda \in]0,\infty[\end{cases}$$

où $\lambda \in [0, +\infty]$, voir la seconde ligne de la Fig. C.1.
Les fonctions génératrices de $\top_{AA_\lambda}$ sont définies par

$$\begin{array}{ll} t(a) = (-\log a)^\lambda & \text{si additive} \\ \theta(a) = \exp(-(-\log a)^\lambda) & \text{si multiplicative} \end{array}$$

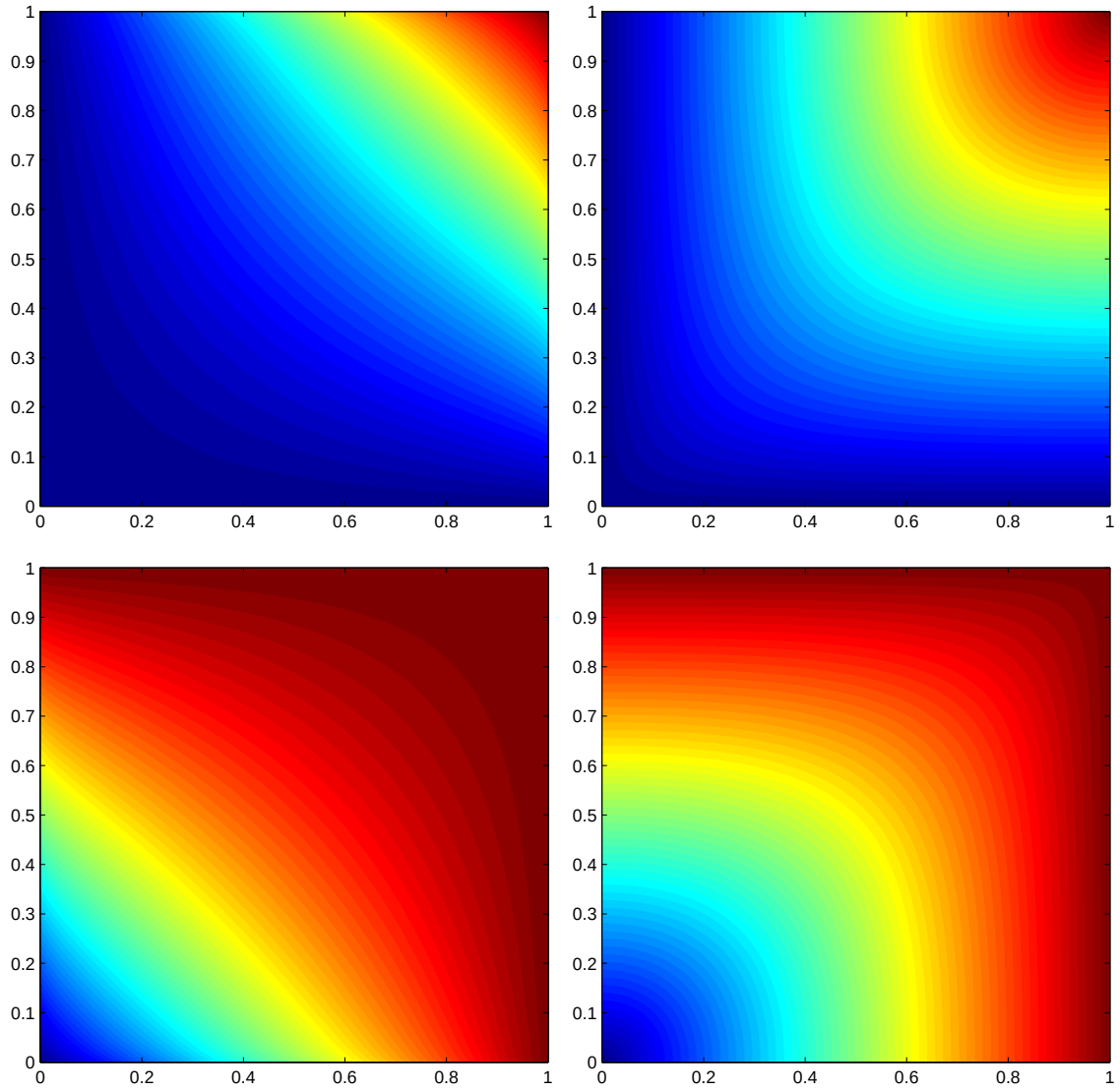

FIG. C.1: Première ligne : Iso-surface de $\top_{AA_\lambda}$ pour $\lambda = 0.75$ et $\lambda = 2$. Deuxième ligne : Iso-surface de $\perp_{AA_\lambda}$ pour $\lambda = 0.75$ et $\lambda = 2$.

À noter que l'on voit aisément que si l'on fixe $\lambda = 1$, on obtient la famille $(\top,\perp)_P$. Par ailleurs, toutes les normes triangulaires générées par cette famille sont continues, à l'exception du cas où $\lambda = 0$.

C.2 Dombi

Cette famille fut introduite dans [Dombi, 1982a] dans un article ayant pour objectif l'étude d'opérateurs conjonctifs et disjonctifs. La norme triangulaire de Dombi $\top_{D_\lambda}$ est définie par

$$\top_{D_\lambda}(a,b) = \begin{cases} \top_D(a,b) & \text{si } \lambda = 0 \\ \top_M(a,b) & \text{si } \lambda = \infty \\ \dfrac{1}{1+\left(\left(\frac{1-a}{a}\right)^\lambda+\left(\frac{1-b}{b}\right)^\lambda\right)^{1/\lambda}} & \text{si } \lambda \in]0,\infty[\end{cases}$$

où $\lambda \in [0,+\infty]$, voir la première ligne de la Fig. C.2.
La conorme triangulaire de Dombi $\perp_{D_\lambda}$ est définie par

$$\perp_{D_\lambda}(a,b) = \begin{cases} \perp_D(a,b) & \text{si } \lambda = 0 \\ \perp_M(a,b) & \text{si } \lambda = \infty \\ 1 - \dfrac{1}{1+\left(\left(\frac{a}{1-a}\right)^\lambda+\left(\frac{b}{1-b}\right)^\lambda\right)^{1/\lambda}} & \text{si } \lambda \in]0,\infty[\end{cases}$$

où $\lambda \in [0,+\infty]$, voir la seconde ligne de la Fig. C.2.
Les fonctions génératrices de $\top_{D_\lambda}$ sont définies par

$$\begin{array}{ll} t(a) = (\frac{1-a}{a})^\lambda & \text{si additive} \\ \theta(a) = \exp(-(\frac{1-a}{a})^\lambda) & \text{si multiplicative} \end{array}$$

A noter que l'on voit aisément que si l'on fixe $\lambda = 1$, on obtient la famille de Hamacher, où le paramètre λ est fixé à 0. Par ailleurs, toutes les normes triangulaires générées par cette famille sont continues, à l'exception du cas où $\lambda = 0$.

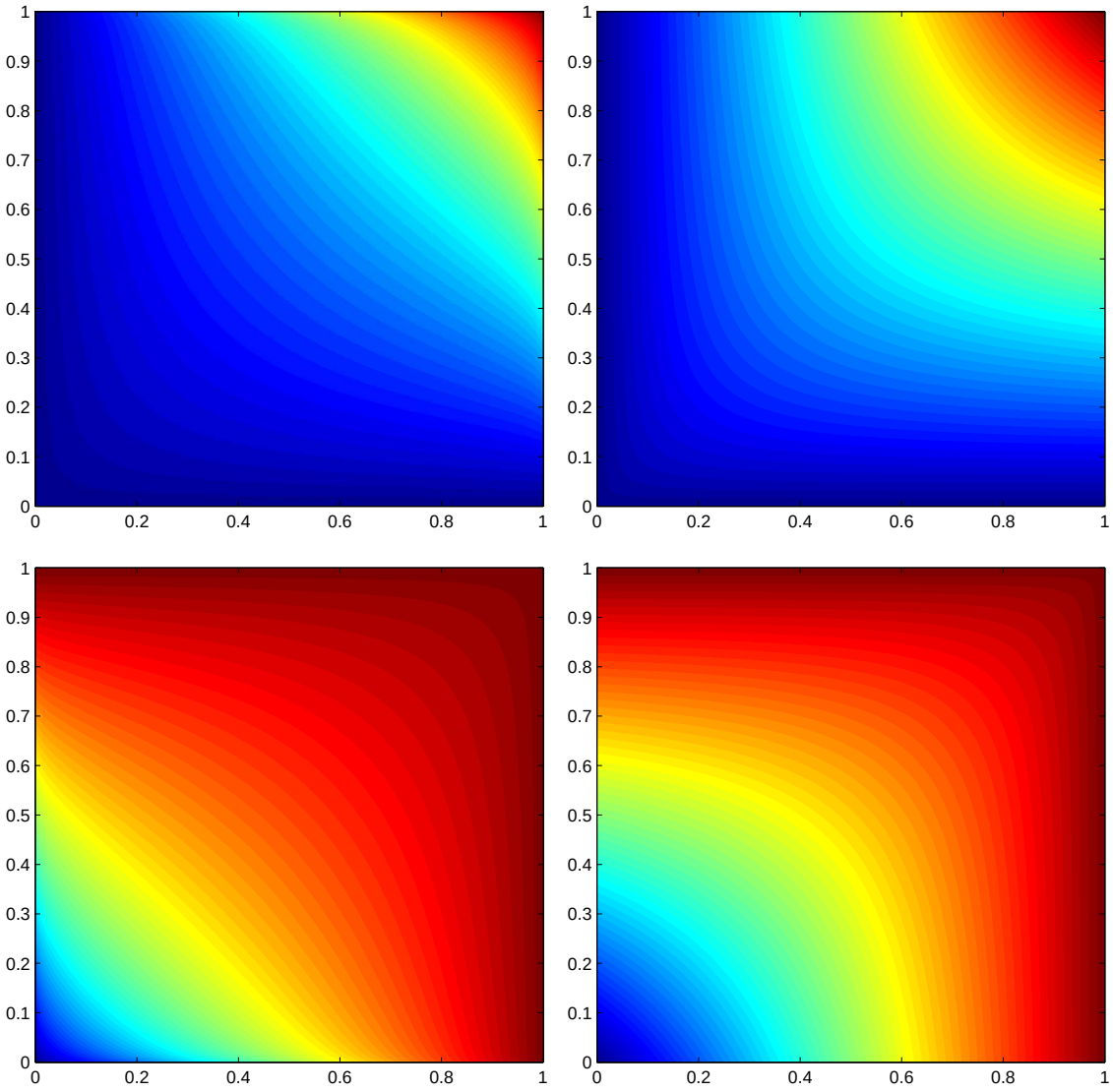

FIG. C.2: Première ligne : Iso-surface de $\top_{D_\lambda}$ pour $\lambda = 0.5$ et $\lambda = 1$. Deuxième ligne : Iso-surface de $\perp_{D_\lambda}$ pour $\lambda = 0.5$ et $\lambda = 1$.

C.3 Dubois-Prade

Dans [Dubois and Prade, 1980], les auteurs introduisent une nouvelle famille de normes triangulaires. Cette famille a ceci de particulier qu'elle n'a pas de fonctions génératrices additives ou multiplicatives, comme c'est habituellement le cas lors la construction de nouvelles familles. La norme triangulaire de Dubois-Prade $\top_{DP_\lambda}$ est définie par

$$\top_{DP_\lambda}(a,b) = \frac{ab}{\max(a,b,\lambda)}$$

où $\lambda \in [0,1]$, voir la première ligne de la Fig. C.3.
La conorme triangulaire de Dubois-Prade $\perp_{DP_\lambda}$ est définie par

$$\perp_{DP_\lambda}(a,b) = 1 - \frac{(1-a)(1-b)}{\max((1-a),(1-b),\lambda)}$$

où $\lambda \in [0, +\infty]$, voir la seconde ligne de la Fig. C.3.

On voit que pour $\lambda = 1$, on obtient le couple $(\top,\perp)_P$, et fixer $\lambda = 0$ permet d'obtenir le

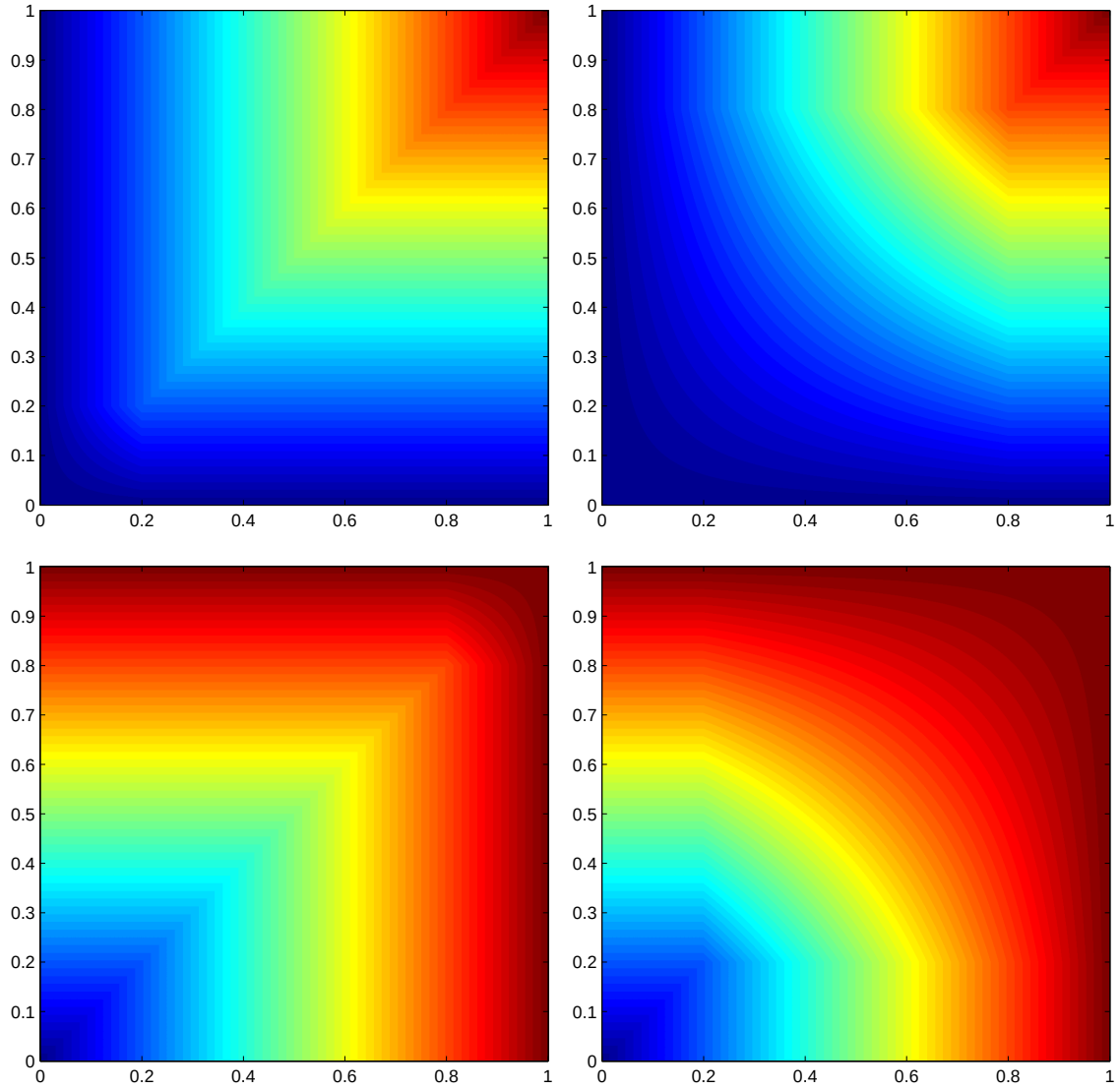

FIG. C.3: Première ligne : Iso-surface de $\top_{DP_\lambda}$ pour $\lambda = 0.2$ et $\lambda = 0.8$. Deuxième ligne : Iso-surface de $\perp_{DP_\lambda}$ pour $\lambda = 0.2$ et $\lambda = 0.8$.

couple $(\top,\perp)_M$.

C.4 Frank

Les recherches sur l'associativité des duals de copules poussent M. Frank [Frank, 1979] à introduire une nouvelle famille, qui sera utilisée par sommation ordinale. La norme triangulaire de Frank $\top_{F_\lambda}$ est définie par

$$\top_{F_\lambda}(a,b) = \begin{cases} \top_M(a,b) & \text{si } \lambda = 0 \\ \top_P(a,b) & \text{si } \lambda = 1 \\ \top_L(a,b) & \text{si } \lambda = \infty \\ \log_\lambda\left(1 + \frac{(\lambda^a-1)(\lambda^b-1)}{\lambda-1}\right) & \text{si } \lambda \in]0,1[\cup]1,\infty[\end{cases}$$

où $\lambda \in [0,+\infty]$, voir la première ligne de la Fig. C.4.
La conorme triangulaire de Frank $\perp_{F_\lambda}$ est définie par

$$\perp_{F_\lambda}(a,b) = \begin{cases} \perp_M(a,b) & \text{si } \lambda = 0 \\ \perp_P(a,b) & \text{si } \lambda = 1 \\ \perp_L(a,b) & \text{si } \lambda = \infty \\ 1 - \log_\lambda\left(1 + \frac{(\lambda^{1-a}-1)(\lambda^{1-b}-1)}{\lambda-1}\right) & \text{si } \lambda \in]0,1[\cup]1,\infty[\end{cases}$$

où $\lambda \in [0,+\infty]$, voir la seconde ligne de la Fig. C.4.
Les fonctions génératrices de $\top_{F_\lambda}$ sont définies par

$$t(a) = \begin{cases} -\log a & \text{si } \lambda = 1 \\ 1 - a & \text{si } \lambda = \infty \\ \log \frac{\lambda-1}{\lambda^a-1} & \text{si } \lambda \in]0,1[\cup]1,\infty[\end{cases} \quad \text{si additive}$$

$$\theta(a) = \begin{cases} a & \text{si } \lambda = 1 \\ \exp(a-1) & \text{si } \lambda = \infty \\ \frac{\lambda^a-1}{\lambda-1} & \text{si } \lambda \in]0,1[\cup]1,\infty[\end{cases} \quad \text{si multiplicative}$$

L'ensemble des normes triangulaires obtenues par ces définitions sont évidemment continues, et sont également des copules [Nelsen, 2006].

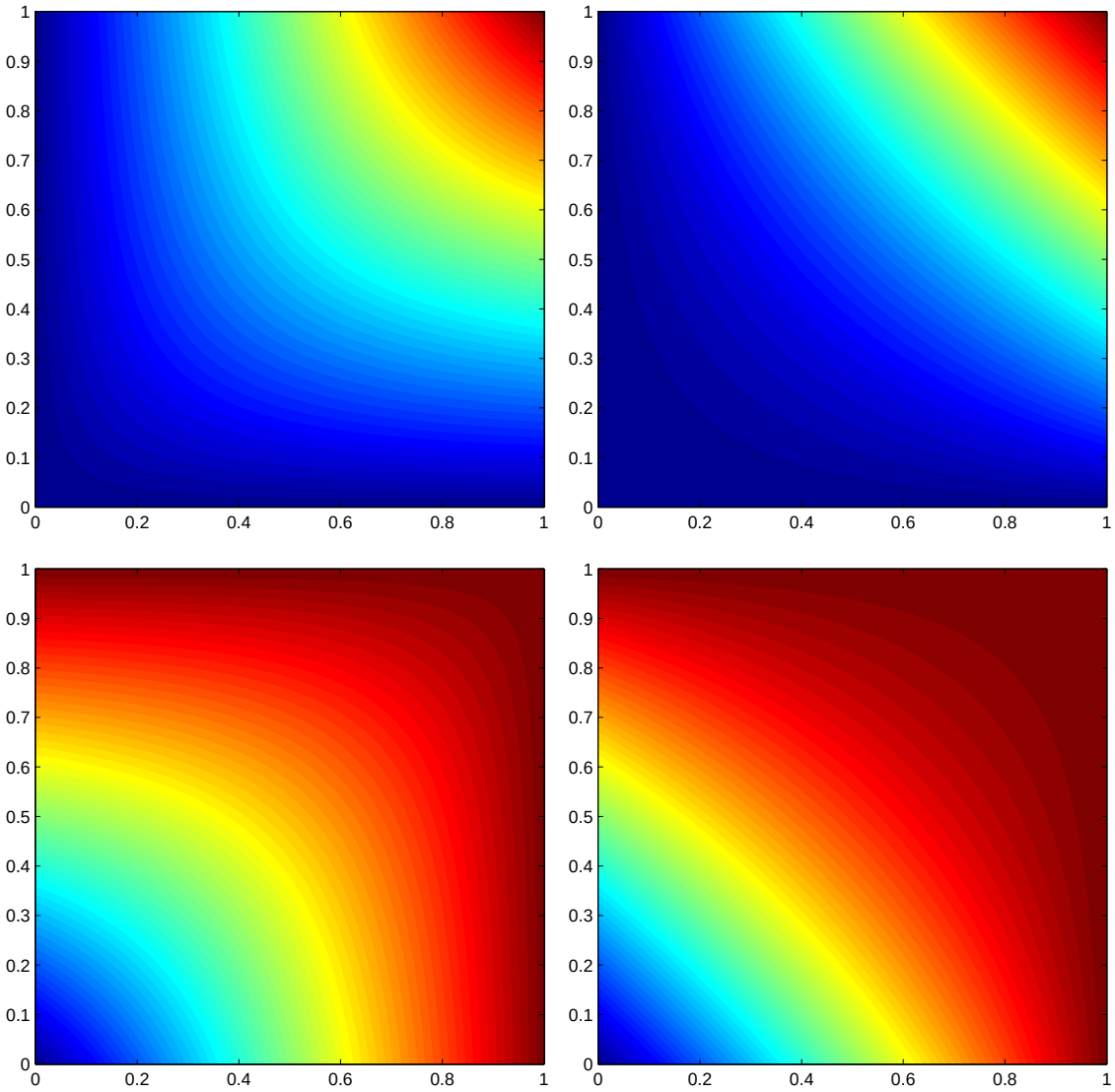

FIG. C.4: Première ligne : Iso-surface de $\top_{F_\lambda}$ pour $\lambda = 0.1$ et $\lambda = 10$. Deuxième ligne : Iso-surface de $\perp_{F_\lambda}$ pour $\lambda = 0.1$ et $\lambda = 10$.

C.5 Hamacher

Dans [Hamacher, 1978], Hamacher propose une approche axiomatique de connecteurs logiques conjonctifs et disjonctifs pouvant être exprimés par des fonctions rationnelles. Une norme triangulaire est en fait un quotient de deux polynômes si et seulement si elle appartient à la famille de Hamacher. La norme triangulaire de Hamacher $\top_{H_\lambda}$ est définie par

$$\top_{H_\lambda}(a,b) = \begin{cases} \top_D(a,b) & \text{si } \lambda = \infty \\ 0 & \text{si } \lambda = a = b = 0 \\ \frac{ab}{\lambda+(1-\lambda)(a+b-ab)} & \text{sinon} \end{cases}$$

où $\lambda \in [0, +\infty]$, voir la première ligne de la Fig. C.5.
La conorme triangulaire de Hamacher $\perp_{H_\lambda}$ est définie par

$$\perp_{H_\lambda}(a,b) = \begin{cases} \perp_D(a,b) & \text{si } \lambda = \infty \\ 1 & \text{si } \lambda = 0 \text{ et } a = b = 1 \\ \frac{a+b-ab-(1-\lambda)ab}{1-(1-\lambda)ab} & \text{sinon} \end{cases}$$

où $\lambda \in [0, +\infty]$, voir la seconde ligne de la Fig. C.5.
Les fonctions génératrices de $\top_{H_\lambda}$ sont définies par

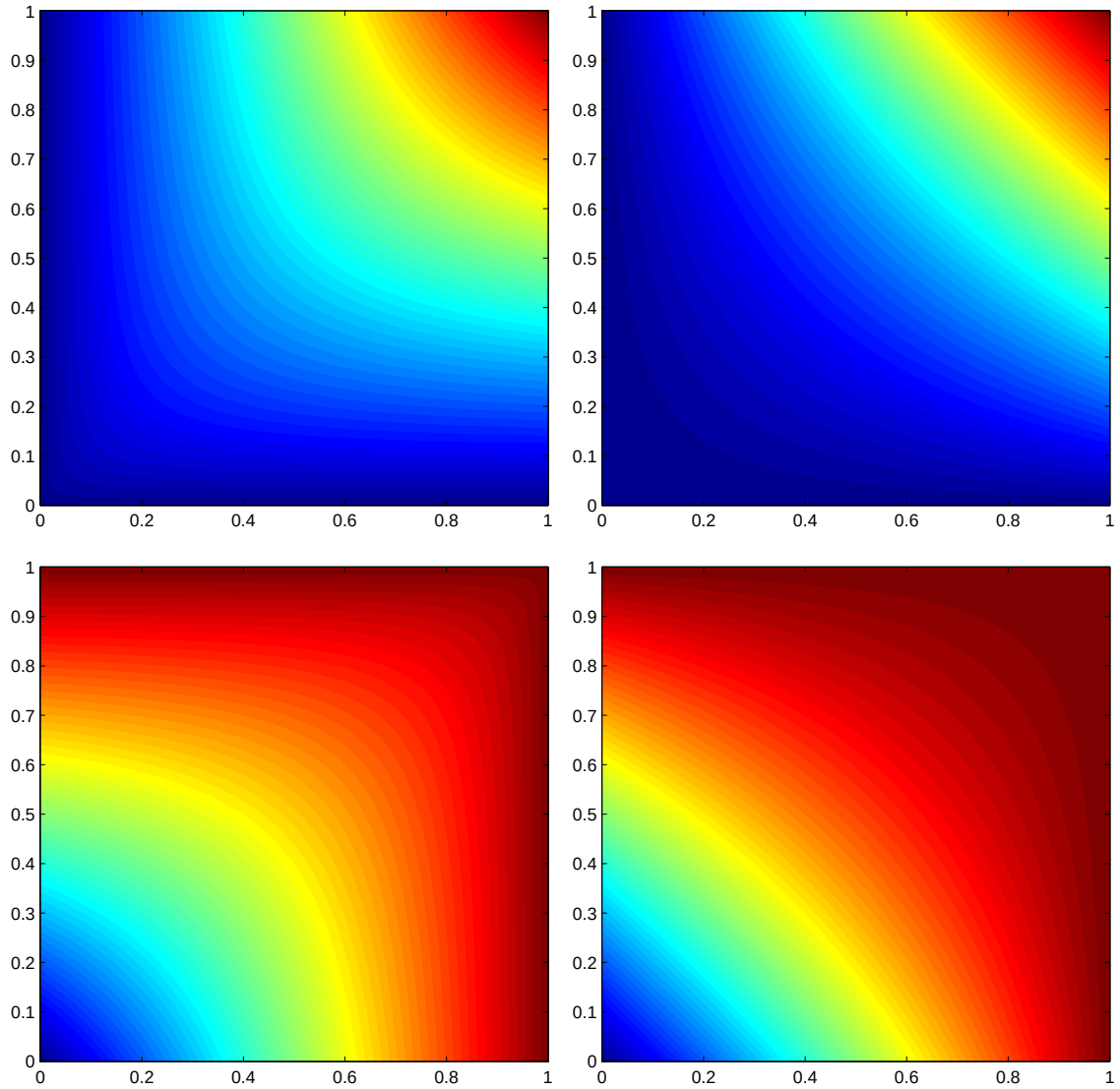

Fig. C.5: Première ligne : Iso-surface de $\top_{H_\lambda}$ pour $\lambda = 0$ et $\lambda = 2$. Deuxième ligne : Iso-surface de $\perp_{H_\lambda}$ pour $\lambda = 0$ et $\lambda = 2$.

$$t(a) = \begin{cases} \frac{1-a}{a} & \text{si } \lambda = 0 \\ \log\left(\frac{\lambda+(1-\lambda)a}{a}\right) & \text{si } \lambda \in]0,\infty[\end{cases} \quad \text{si additive}$$

$$\theta(a) = \begin{cases} \exp(\frac{a-1}{a}) & \text{si } \lambda = 0 \\ \frac{a}{\lambda+(1-\lambda)a} & \text{si } \lambda \in]0,\infty[\end{cases} \quad \text{si multiplicative}$$

On voit immédiatement que pour $\lambda = 1$, on obtient le couple $(\top,\perp)_P$. On appelle parfois $\top_{H_0}$ le *produit de Hamacher*, et $\perp_{H_2}$ la *somme de Einstein*. Si l'on a $\lambda \in [0,2]$, alors $(\top,\perp)_{H_\lambda}$ est une famille de copules.

C.6 Mayor-Torrens

Dans [Mayor and Torrens, 1991], Mayor et Torrens introduisent une nouvelle famille, dont la norme triangulaire est la seule continue et satisfaisant, pour tout $(a,b) \in [0,1]^2$, la relation $\top(a,b) = \max(\top(\max(a,b),\max(a,b)) - |a-b|,0)$. La norme triangulaire de Mayor-Torrens $\top_{MT_\lambda}$ est définie par

$$\top_{MT_\lambda}(a,b) = \begin{cases} \max(a+b-\lambda,0) & \text{si } \lambda \in]0,1] \text{ et } (a,b) \in [0,\lambda]^2 \\ \top_M(a,b) & \text{sinon} \end{cases}$$

où $\lambda \in [0,1]$, voir la première ligne de la Fig. C.6.
La conorme triangulaire de Mayor-Torrens $\perp_{MT_\lambda}$ est définie par

$$\perp_{MT_\lambda}(a,b) = \begin{cases} \min(a+b+\lambda-1,1) & \text{si } \lambda \in]0,1] \text{ et } (a,b) \in [1-\lambda,1]^2 \\ \perp_M(a,b) & \text{sinon} \end{cases}$$

où $\lambda \in [-1,+\infty]$, voir la seconde ligne de la Fig. C.6.
On voit facilement que pour $\lambda = 0$, on obtient $(\top,\perp)_M$, pour $\lambda = 1$, $(\top,\perp)_L$. L'ensemble des normes triangulaires de Mayor-Torrens sont continues, sont duals, et forment des copules.

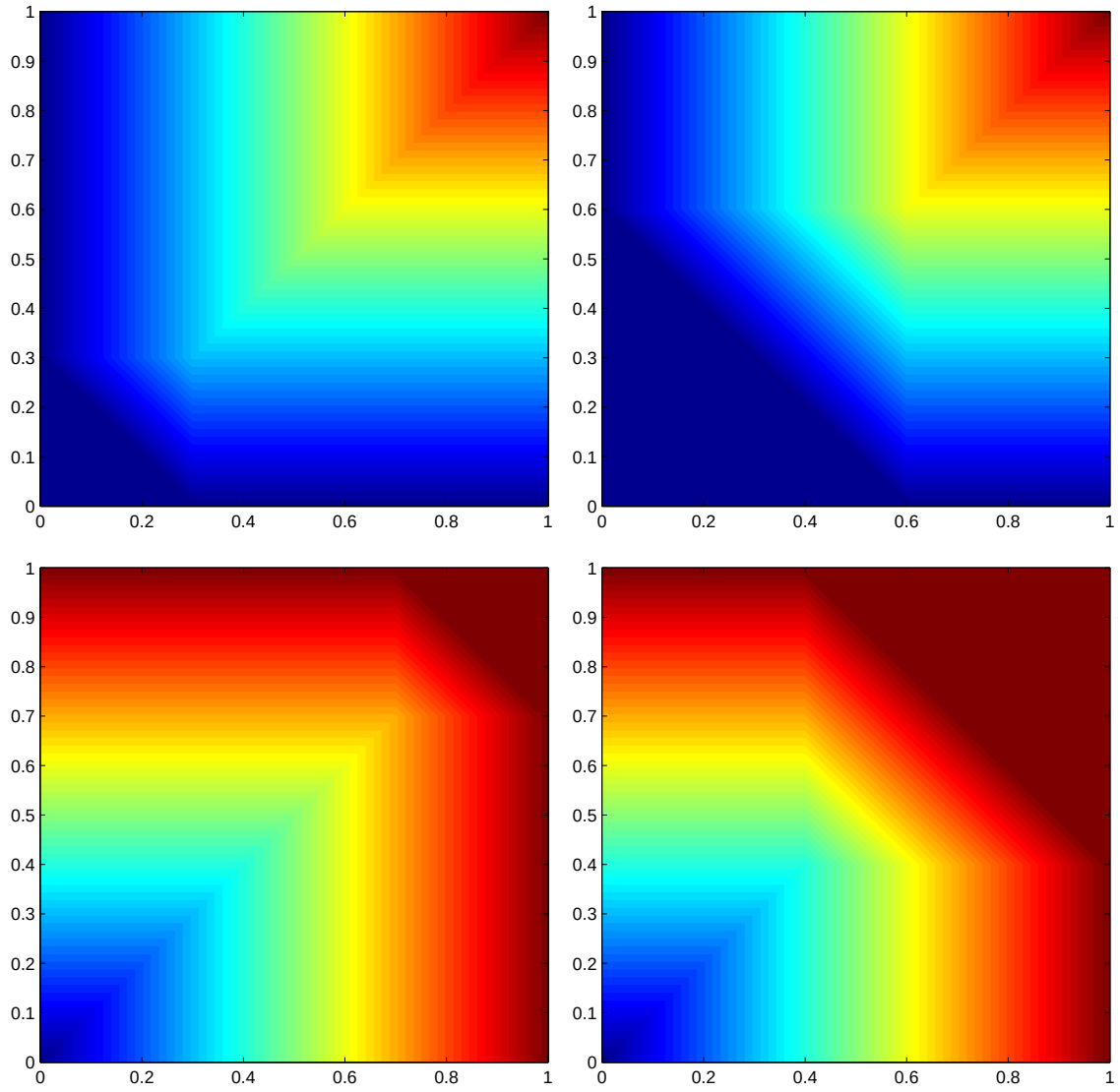

FIG. C.6: Première ligne : Iso-surface de $\top_{MT_\lambda}$ pour $\lambda = 0.3$ et $\lambda = 0.6$. Deuxième ligne : Iso-surface de $\perp_{MT_\lambda}$ pour $\lambda = 0.3$ et $\lambda = 0.6$.

C.7 Schweizer-Sklar

Très tôt, une famille comprenant l'ensemble des quatres normes triangulaires usuelles (S, A, L et D) est introduite par Schweizer & Sklar dans [Schweizer and Sklar, 1983]. La norme triangulaire de Schweizer-Sklar $\top_{SS_\lambda}$ est définie par

$$\top_{SS_\lambda}(a,b) = \begin{cases} \top_M(a,b) & \text{si } \lambda = -\infty \\ \top_P(a,b) & \text{si } \lambda = 0 \\ \top_D(a,b) & \text{si } \lambda = \infty \\ \left(\max\left(a^\lambda + b^\lambda - 1, 0\right)\right)^{\frac{1}{\lambda}} & \text{si } \lambda \in]-\infty, 0[\cup]0, \infty[\end{cases}$$

où $\lambda \in [-\infty, +\infty]$, voir la première ligne de la Fig. C.7.
La conorme triangulaire de Schweizer-Sklar $\perp_{SS_\lambda}$ est définie par

$$\perp_{SS_\lambda}(a,b) = \begin{cases} \perp_M(a,b) & \text{si } \lambda = -\infty \\ \perp_P(a,b) & \text{si } \lambda = 0 \\ \perp_D(a,b) & \text{si } \lambda = \infty \\ 1 - \left(\max\left((1-a)^\lambda + (1-b)^\lambda - 1, 0\right)\right)^{\frac{1}{\lambda}} & \text{si } \lambda \in]-\infty, 0[\cup]0, \infty[\end{cases}$$

où $\lambda \in [-\infty, +\infty]$, voir la seconde ligne de la Fig. C.7.
Les fonctions génératrices de $\top_{SS_\lambda}$ sont définies par

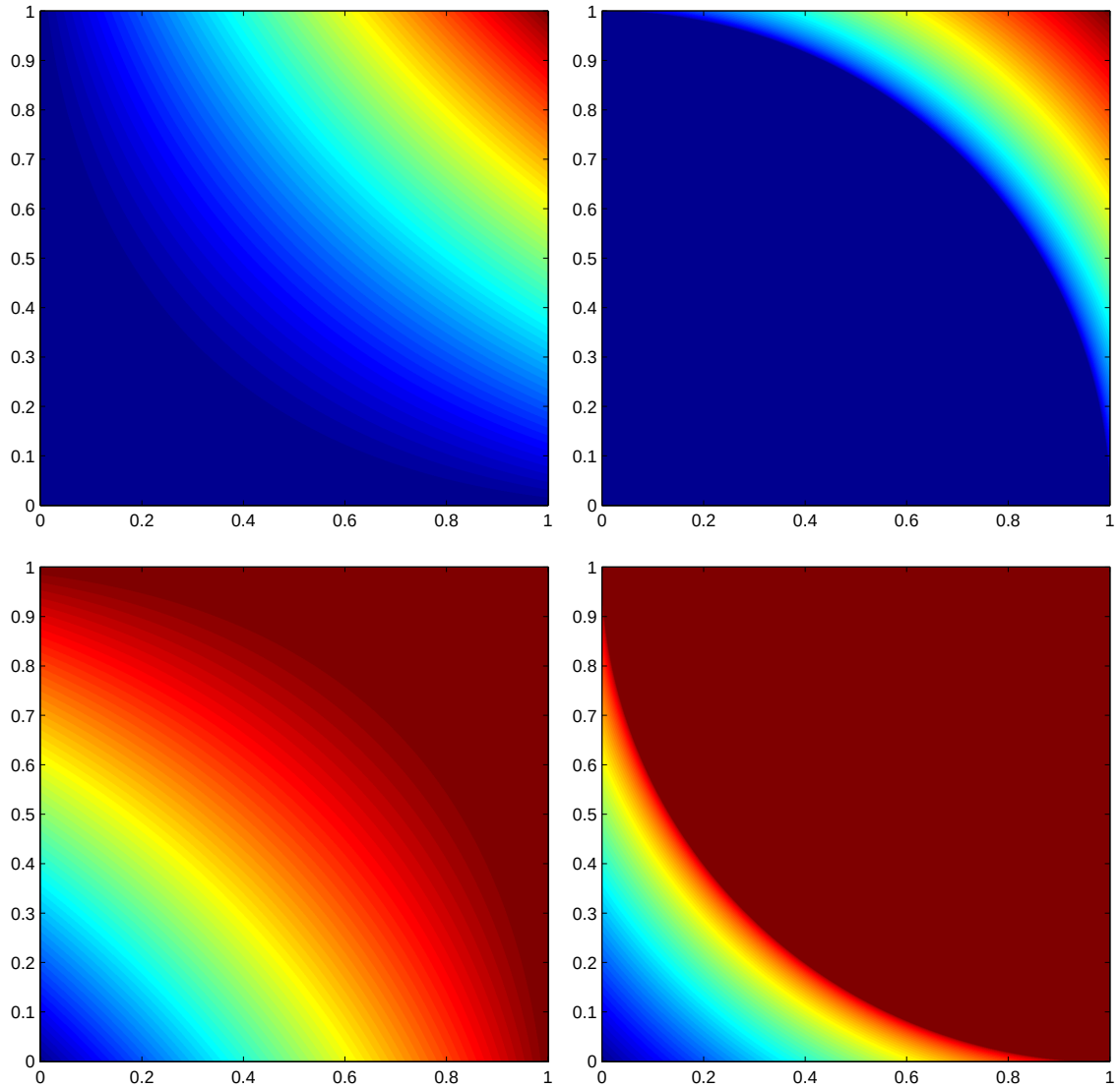

Fig. C.7: Première ligne : Iso-surface de $\top_{SS_\lambda}$ pour $\lambda = 0.5$ et $\lambda = 2$. Deuxième ligne : Iso-surface de $\perp_{SS_\lambda}$ pour $\lambda = 0.5$ et $\lambda = 2$.

$$t(a) = \begin{cases} -\log a & \text{si } \lambda = 0 \\ \frac{1-a^\lambda}{\lambda} & \text{si } \lambda \in]-\infty,0[\cup]0,\infty[\end{cases} \quad \text{si additive}$$

$$\theta(a) = \begin{cases} a & \text{si } \lambda = 0 \\ \exp(\frac{a^\lambda - 1}{\lambda}) & \text{si } \lambda \in]-\infty,0[\cup]0,\infty[\end{cases} \quad \text{si multiplicative}$$

On voit facilement que pour $\lambda = 1$, on obtient le couple $(\top,\bot)_L$. Pour tout λ réel, $(\top,\bot)_{SS_\lambda}$ sont duals, et à l'exception de $\lambda = \infty$, toutes les normes triangulaires sont continues. Pour former des copules, on doit restreindre la valeur de λ dans $[-\infty,1]$.

C.8 Weber-Sugeno

Dans [Weber, 1983], l'utilisation d'une nouvelle famille pour la modélisation d'union et d'intersection est suggérée. On remarquera que les conormes triangulaires de cette famille peuvent être vues comme une généralisation de l'addition dans le cadre des λ-mesures floues, [Sugeno, 1974]. La norme triangulaire de Weber-Sugeno $\top_{WS_\lambda}$ est définie par

$$\top_{WS_\lambda}(a,b) = \begin{cases} \top_D(a,b) & \text{si } \lambda = -1 \\ \top_P(a,b) & \text{si } \lambda = \infty \\ \max\left(\frac{a+b-1+\lambda ab}{1+\lambda},0\right) & \text{sinon} \end{cases}$$

où $\lambda \in [-1,+\infty]$, voir la première ligne de la Fig. C.8.
La conorme triangulaire de Weber-Sugeno $\perp_{WS_\lambda}$ est définie par

$$\perp_{WS_\lambda}(a,b) = \begin{cases} \perp_P(a,b) & \text{si } \lambda = -1 \\ \perp_D(a,b) & \text{si } \lambda = \infty \\ \min\left(a+b+\lambda ab,1\right) & \text{sinon} \end{cases}$$

où $\lambda \in [-1,+\infty]$, voir la seconde ligne de la Fig. C.8.
Les fonctions génératrices de $\top_{WS_\lambda}$ sont définies par

$$t(a) = \begin{cases} 1-a & \text{si } \lambda = 0 \\ -\log a & \text{si } \lambda = \infty \\ 1-\frac{\log(1+\lambda a)}{\log(1+\lambda)} & \text{si } \lambda \in]-1,0[\cup]0,\infty[\end{cases} \quad \text{si additive}$$

$$\theta(a) = \begin{cases} \exp(1-a) & \text{si } \lambda = 0 \\ a & \text{si } \lambda = \infty \\ \exp\left(\frac{\log(1+\lambda a)}{\log(1+\lambda)}-1\right) & \text{si } \lambda \in]-1,0[\cup]0,\infty[\end{cases} \quad \text{si multiplicative}$$

On voit facilement que pour $\lambda = 0$, on obtient le couple $(\top,\perp)_L$. À la différence des autres familles, $\top_{WS_\lambda}$ et $\perp_{WS_\mu}$ sont duals si et seulement si $\mu = -\frac{\lambda}{1+\lambda}$. Pour cette raison, les figures présentées sont les seules à ne pas présenter une certaine symétrie parmi l'ensemble des couples paramétriques présentés.

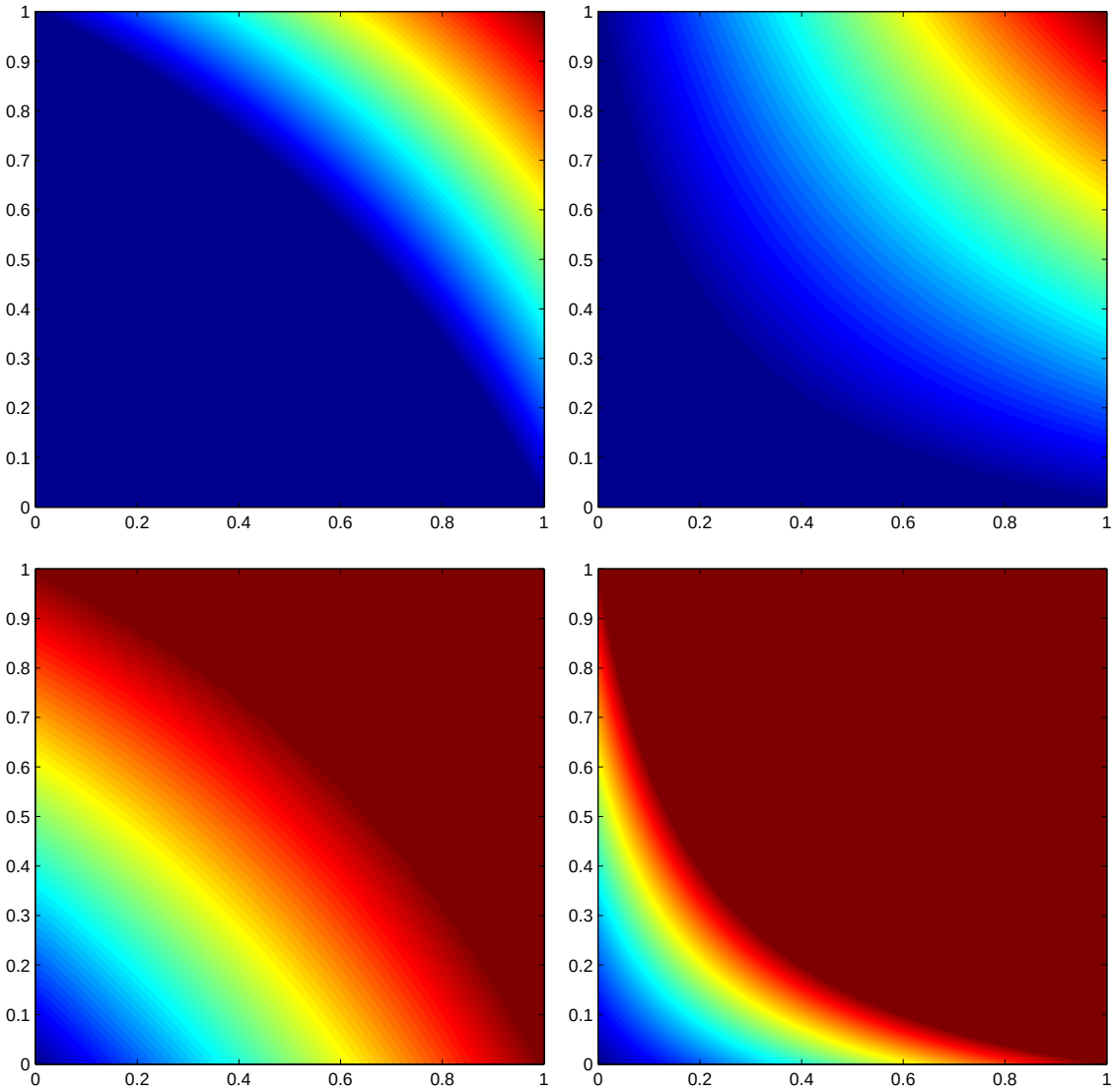

FIG. C.8: Première ligne : Iso-surface de $\top_{WS_\lambda}$ pour $\lambda = -0.5$ et $\lambda = 5$. Deuxième ligne : Iso-surface de $\perp_{WS_\lambda}$ pour $\lambda = -0.5$ et $\lambda = 5$.

C.9 Yager

Cette famille, introduite dans [Yager, 1980], figure parmi les choix les plus populaires pour la modélisation d'intersection entre ensembles flous. Cette proposition repose sur le fait d'utiliser λ comme une mesure de l'importance du **ET** logique. Ainsi, fixer $\lambda = 0$ correspond au plus faible **ET**, tandis que $\lambda = \infty$ correspondra au plus fort **ET**. La norme triangulaire de Yager $\top_{Y_\lambda}$ est définie par

$$\top_{Y_\lambda}(a,b) = \begin{cases} \top_D(a,b) & \text{si } \lambda = 0 \\ \top_M(a,b) & \text{si } \lambda = \infty \\ \max\left(1 - \left((1-a)^\lambda + (1-b)^\lambda\right)^{1/\lambda}, 0\right) & \text{si } \lambda \in]0,\infty[\end{cases}$$

où $\lambda \in [0, +\infty]$, voir la première ligne de la Fig. C.9.
La conorme triangulaire de Yager $\perp_{Y_\lambda}$ est définie par

$$\perp_{Y_\lambda}(a,b) = \begin{cases} \perp_D(a,b) & \text{si } \lambda = 0 \\ \perp_M(a,b) & \text{si } \lambda = \infty \\ \min\left((a^\lambda + b^\lambda)^{1/\lambda}, 1\right) & \text{si } \lambda \in]0,\infty[\end{cases}$$

où $\lambda \in [0, +\infty]$, voir la seconde ligne de la Fig. C.9.

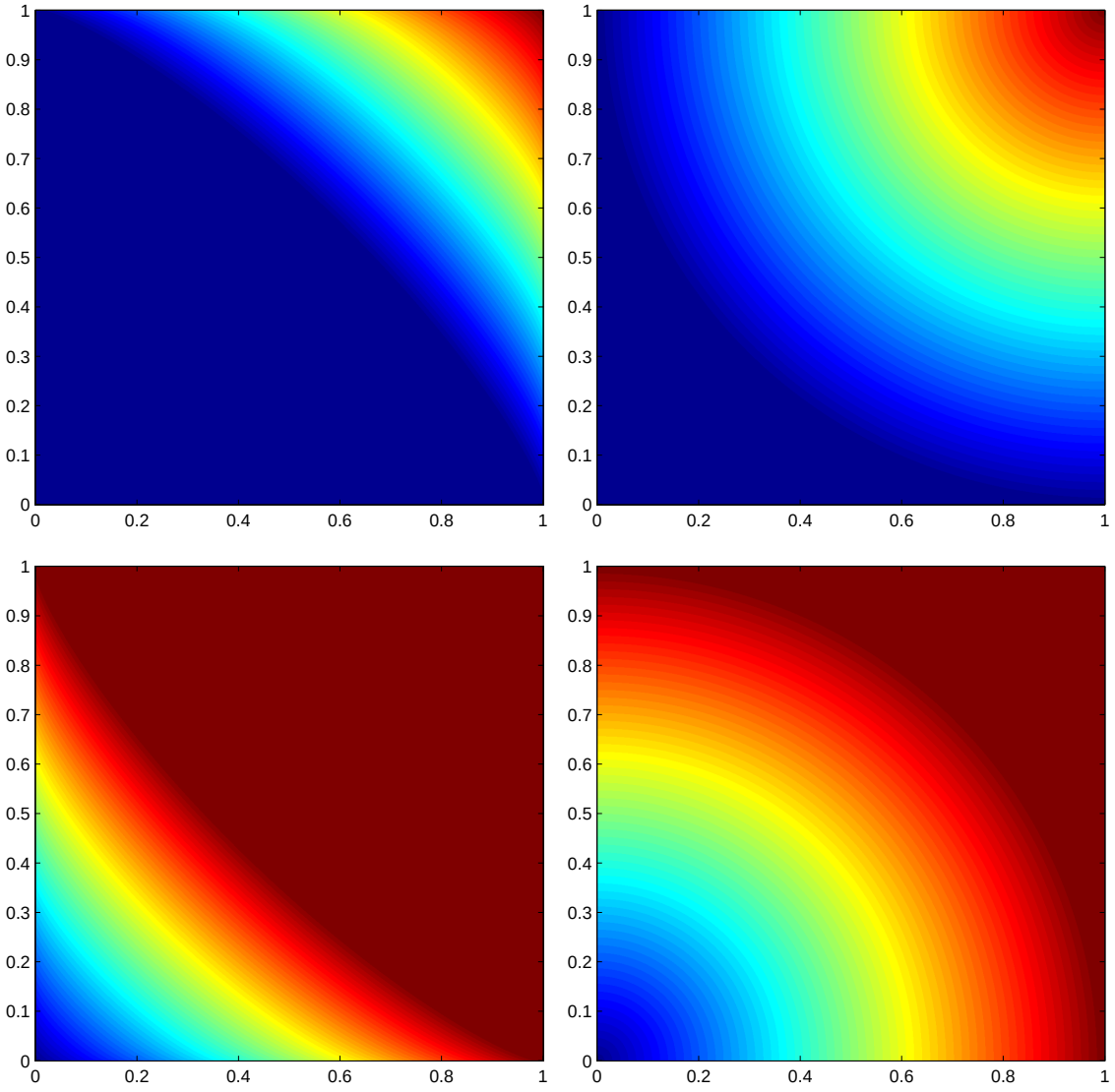

Fig. C.9: Première ligne : Iso-surface de $\top_{Y_\lambda}$ pour $\lambda = 0.8$ et $\lambda = 2$. Deuxième ligne : Iso-surface de $\perp_{Y_\lambda}$ pour $\lambda = 0.8$ et $\lambda = 2$.

Les fonctions génératrices de $\top_{Y_\lambda}$ sont définies par

$$\begin{aligned} t(a) &= (1-a)^\lambda && \text{si additive} \\ \theta(a) &= \exp(-(1-a)^\lambda) && \text{si multiplicative} \end{aligned}$$

En fixant $\lambda = 1$, on voit facilement que l'on obtient le couple $(\top,\bot)_L$. L'ensemble des couples de Yager, à l'exception de $(\top,\bot)_{Y_0}$, sont continus, et sont également des copules.

Annexe D

Résultats complémentaires

D.1 Courbes ER

Dans cette section, nous donnons les courbes Erreur-Rejet (ER) de l'ensemble des jeux de données considérés au Chapitre 4.

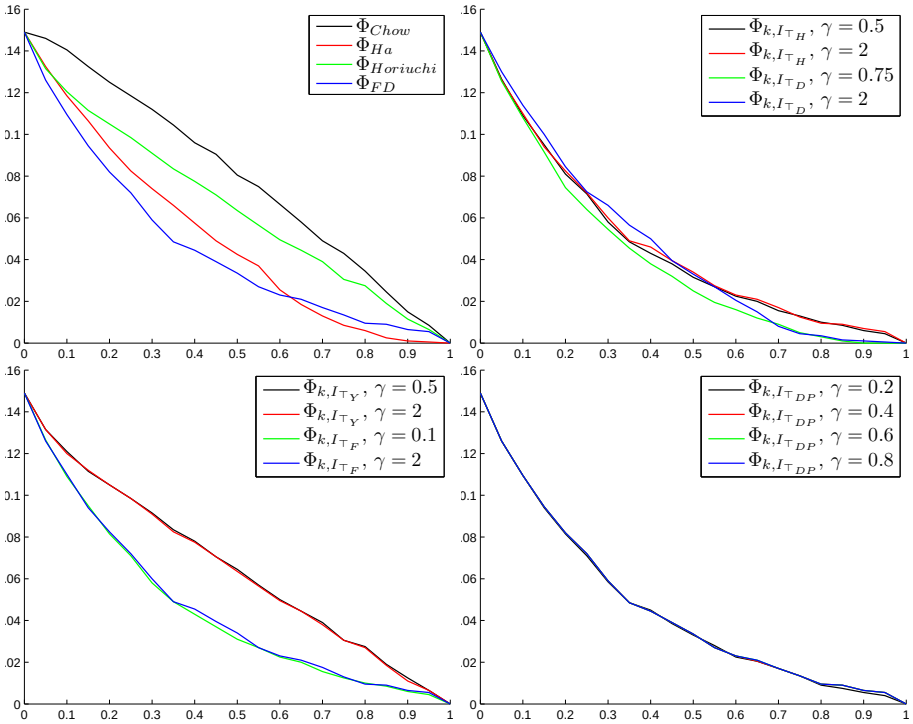

Fig. D.1: Courbes ER sur les données $D1$. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

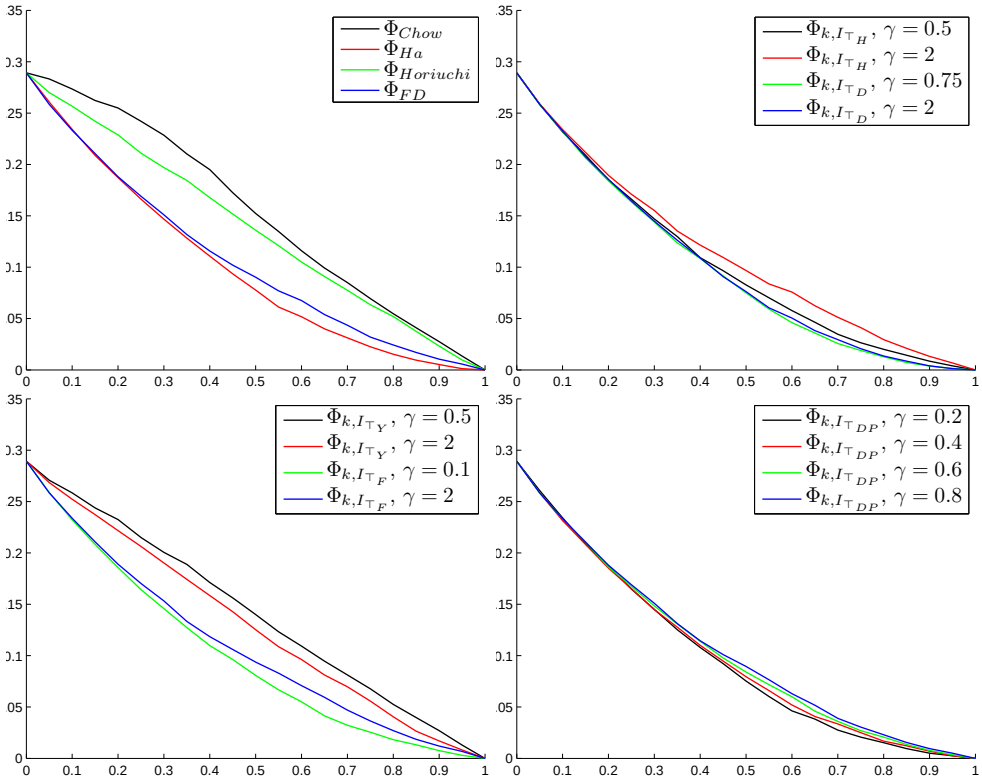

FIG. D.2: Courbes ER sur les données $D2$. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

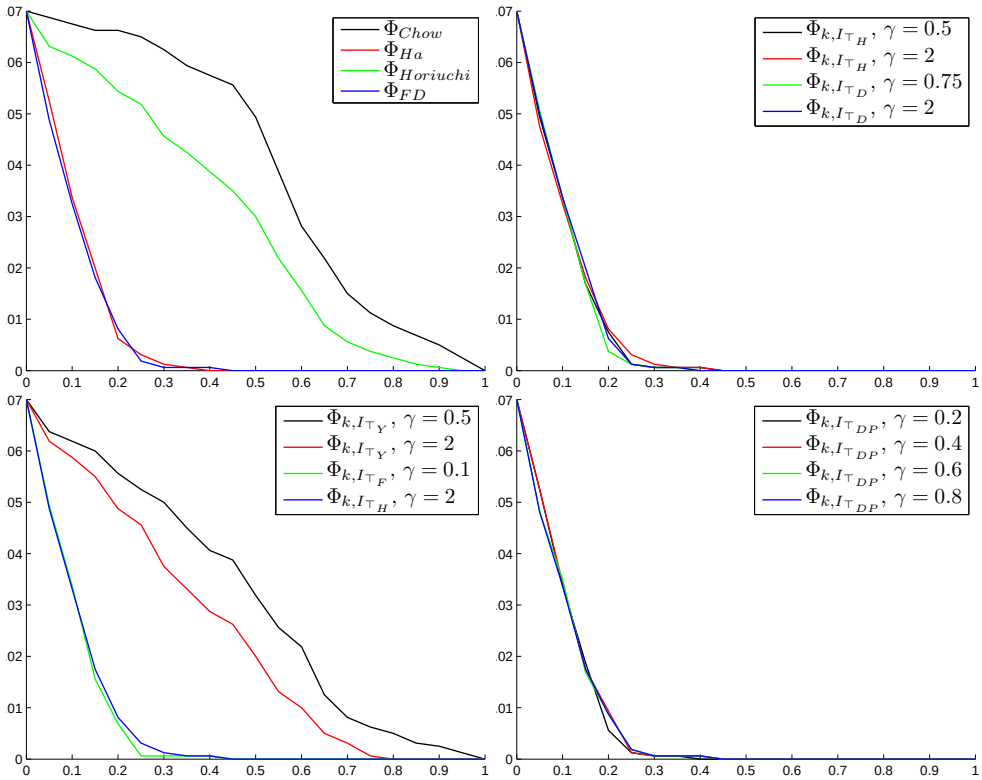

FIG. D.3: Courbes ER sur les données DH. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

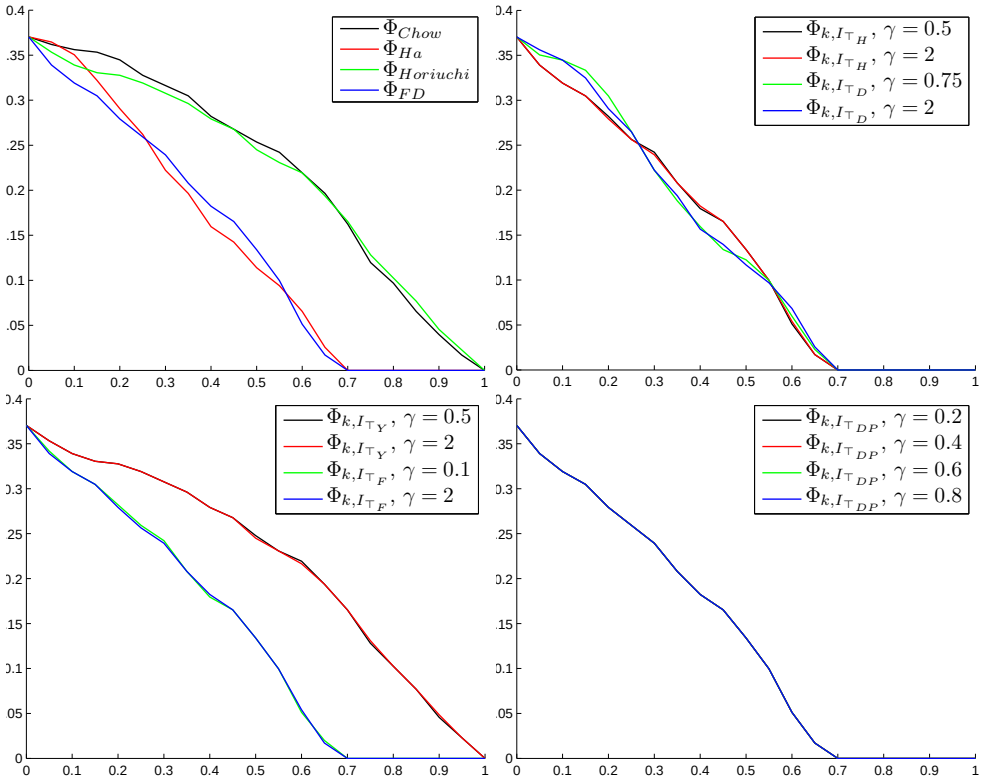

FIG. D.4: Courbes ER sur les données $Ionosphere$. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

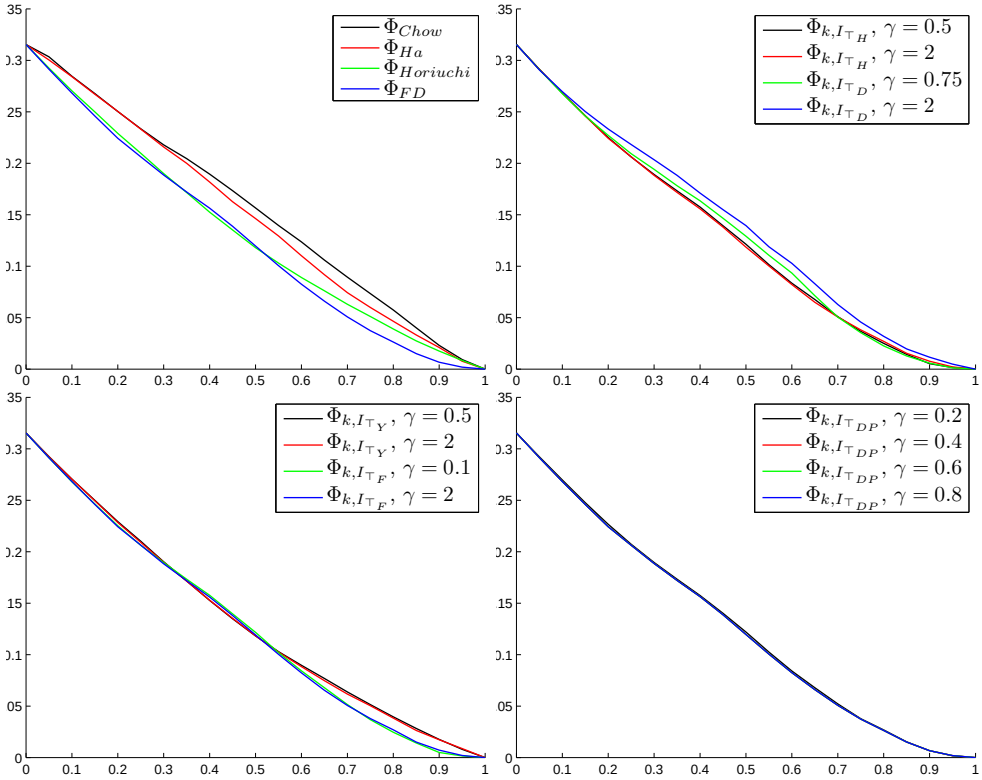

FIG. D.5: Courbes ER sur les données $Forest$. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

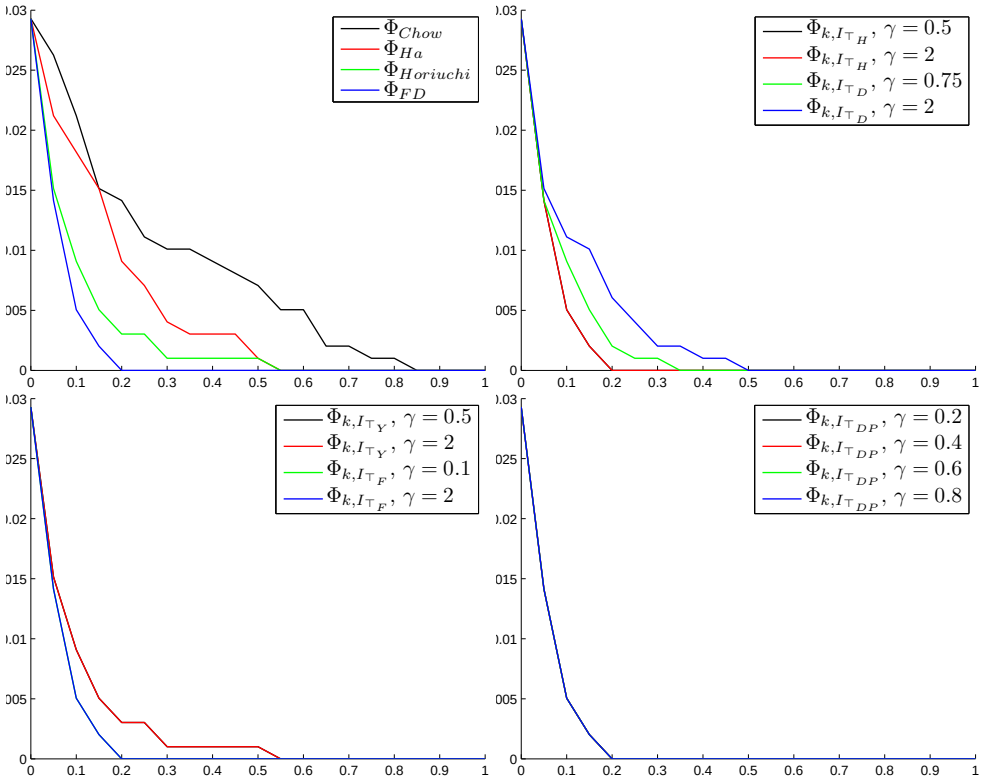

FIG. D.6: Courbes ER sur les données $Vowel$. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

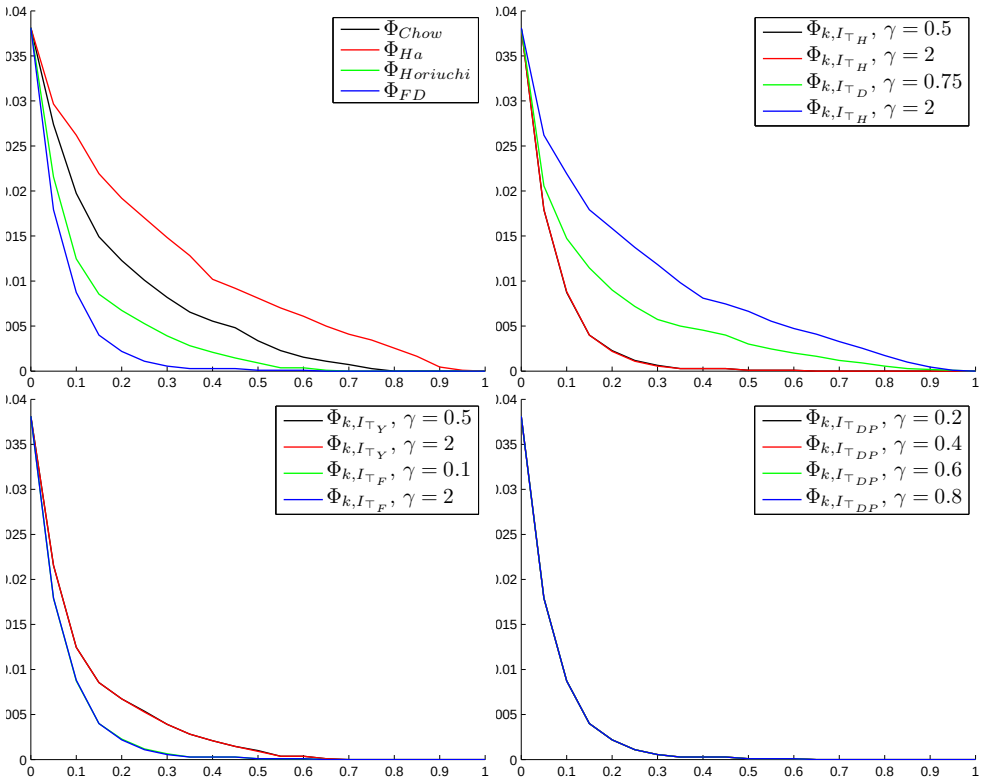

FIG. D.7: Courbes ER sur les données $Digits$. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

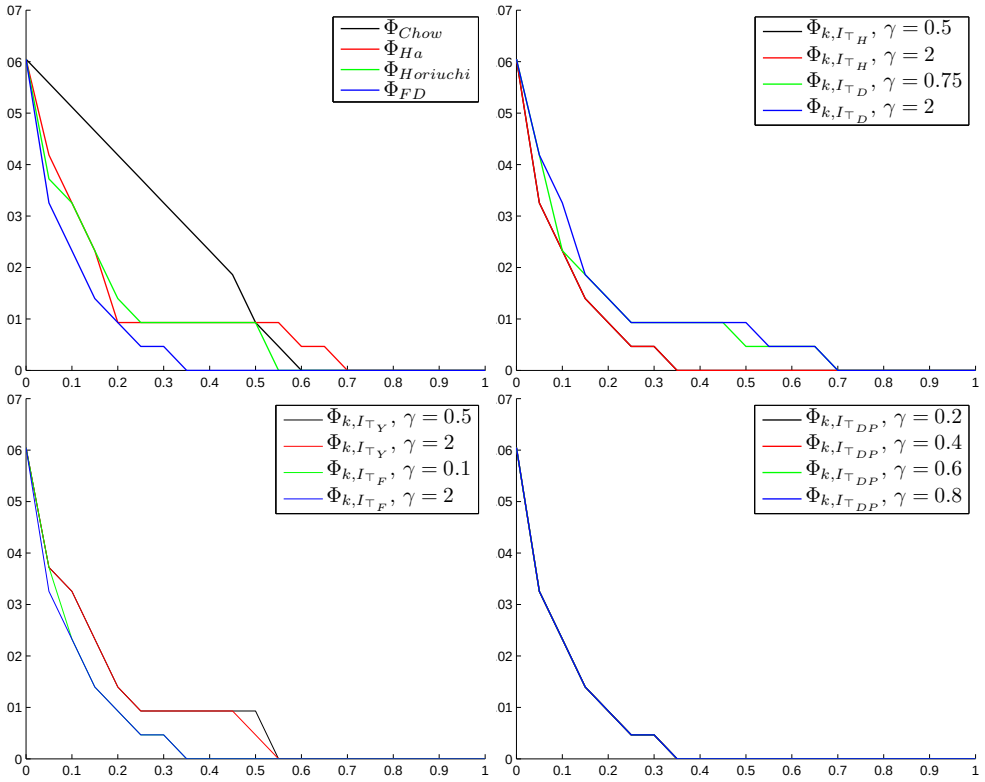

FIG. D.8: Courbes ER sur les données *Thyroid*. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

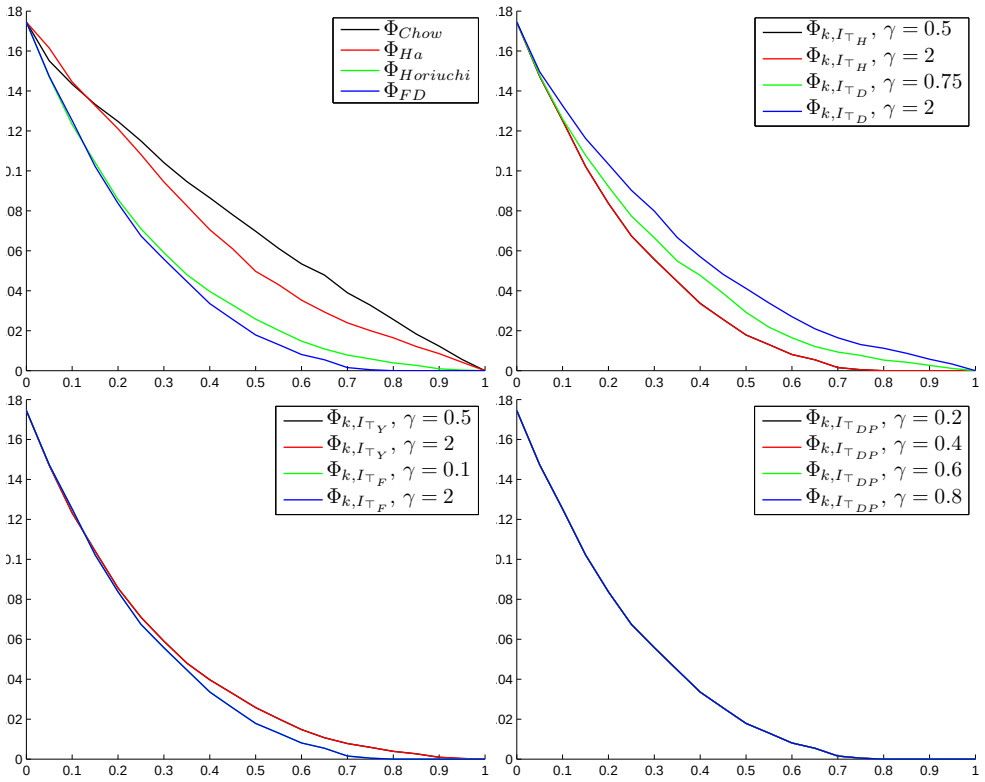

FIG. D.9: Courbes ER sur les données *Statlog*. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

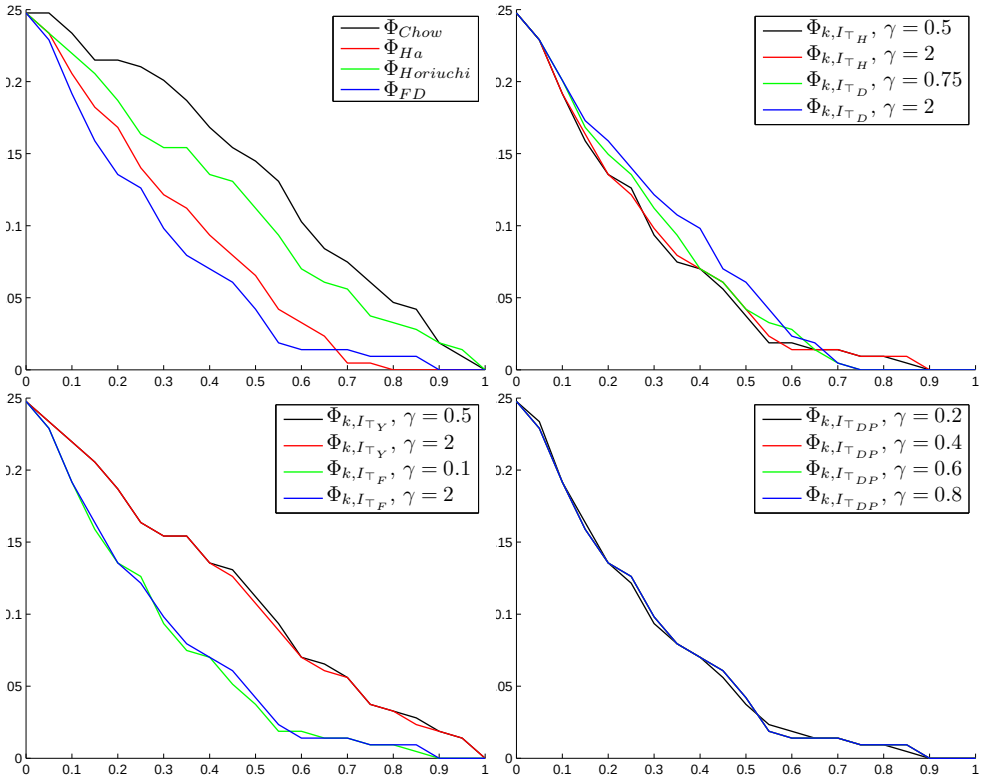

FIG. D.10: Courbes ER sur les données *Glass*. Mesures usuelles (*haut-gauche*), Mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*)

D.2 Courbes $E\overline{n}$

Dans cette section, nous donnons les courbes Erreur-Nombre Moyen de Classes ($E\overline{n}$) de l'ensemble des jeux de données considérés au CHAPITRE 4.

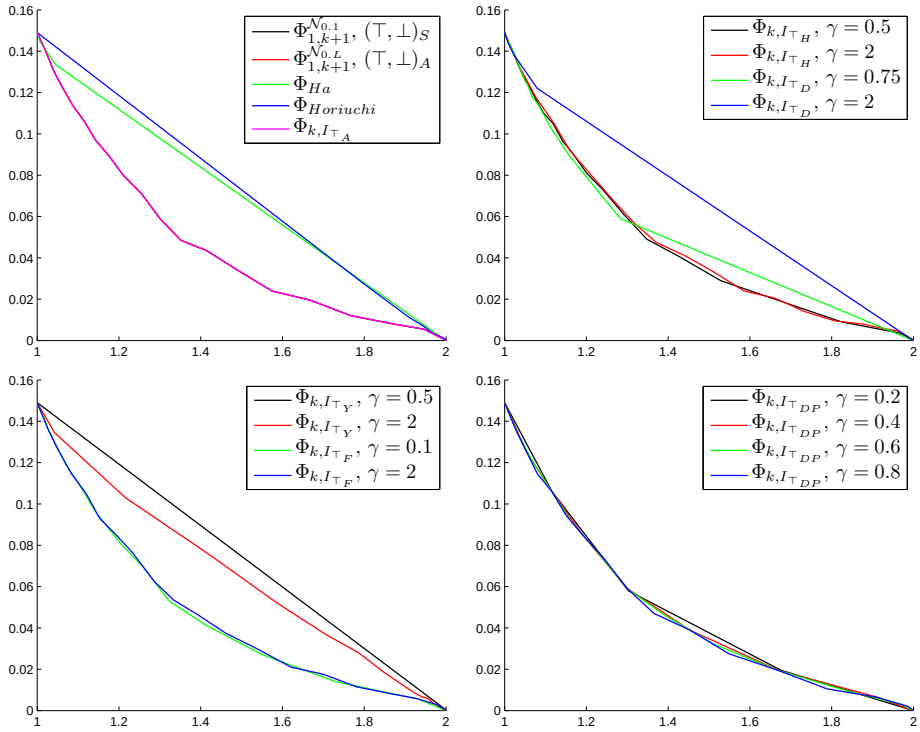

FIG. D.11: Courbes $E\overline{n}$ sur les données $D1$. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

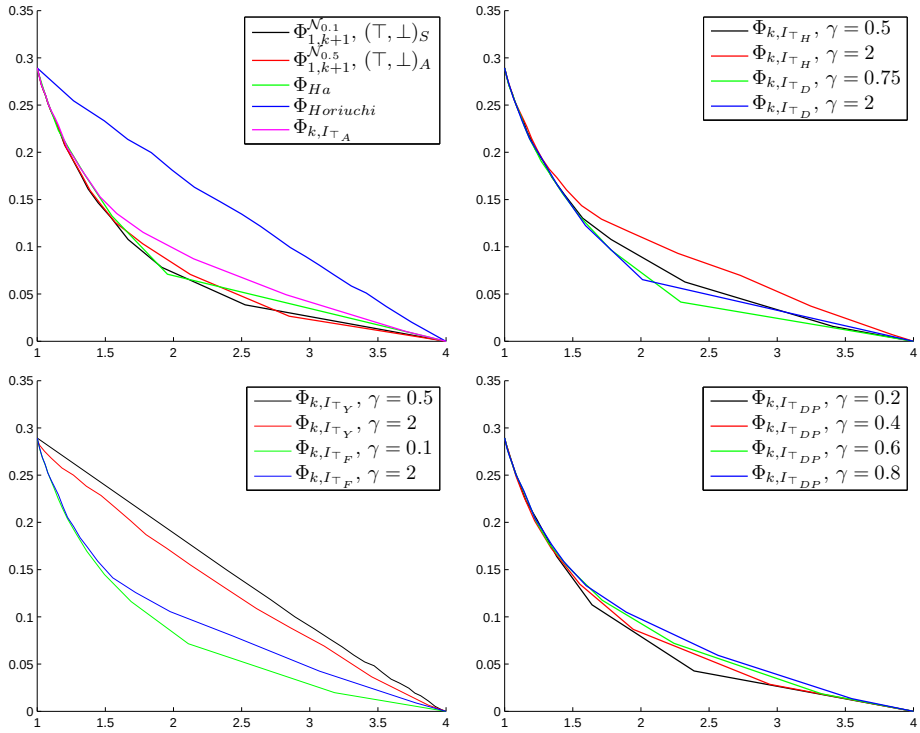

FIG. D.12: Courbes $E\overline{n}$ sur les données $D2$. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

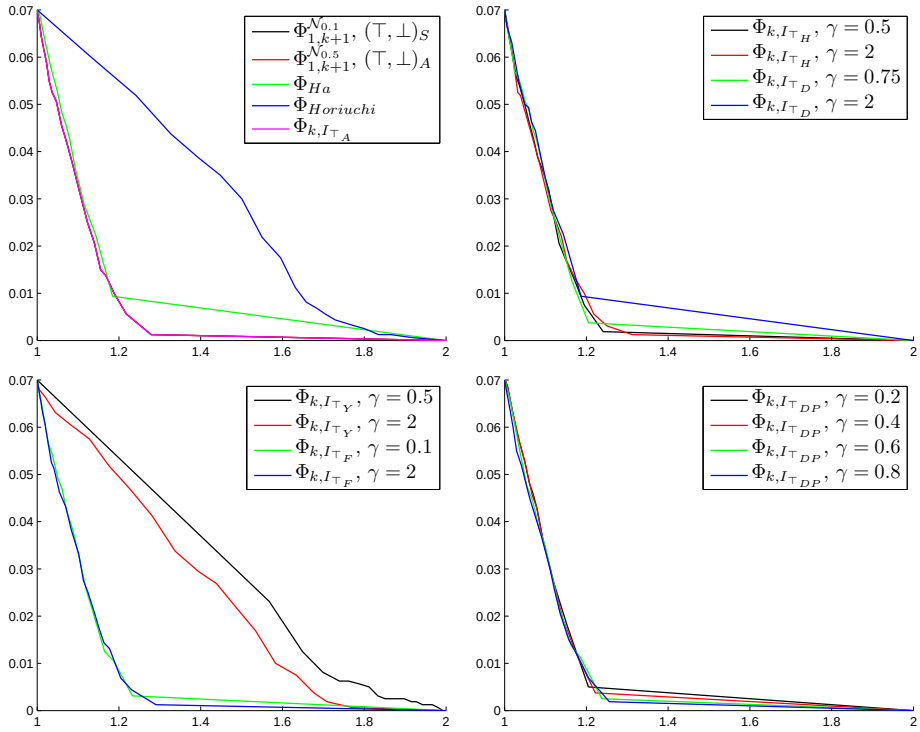

FIG. D.13: Courbes $E\overline{n}$ sur les données DH. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

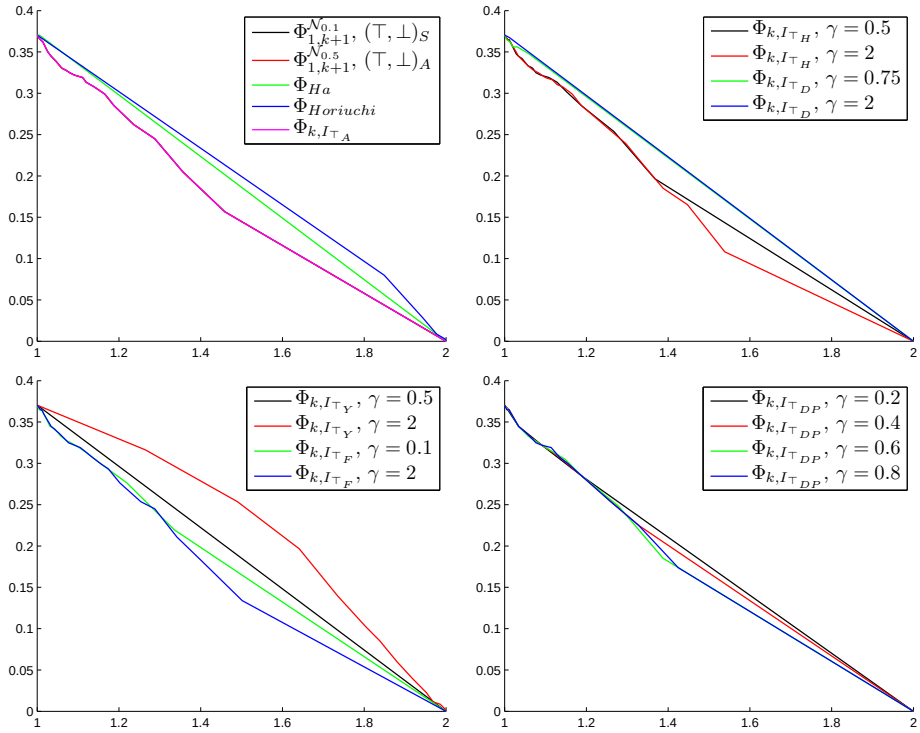

FIG. D.14: Courbes $E\overline{n}$ sur les données *Ionosphere*. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

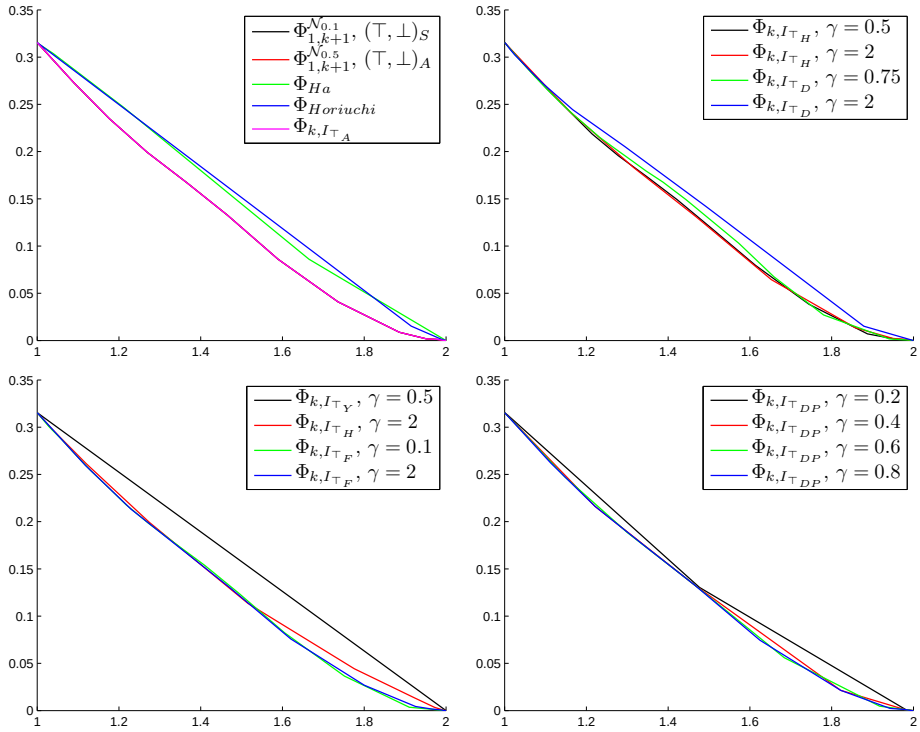

FIG. D.15: Courbes $E\overline{n}$ sur les données *Forest*. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

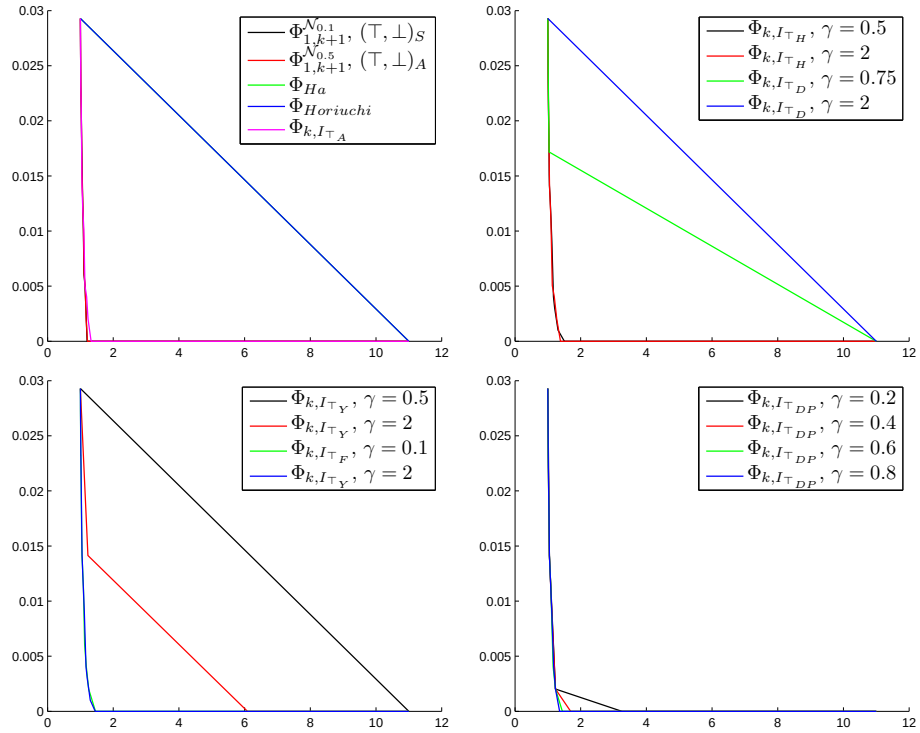

FIG. D.16: Courbes $E\overline{n}$ sur les données *Vowel*. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

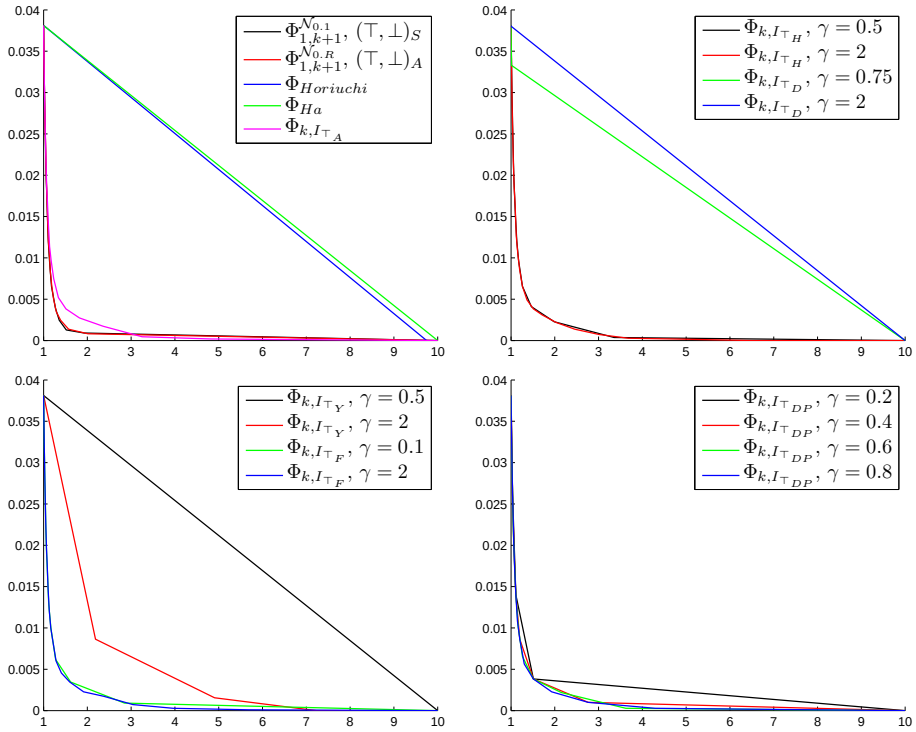

FIG. D.17: Courbes $E\overline{n}$ sur les données *Digits*. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

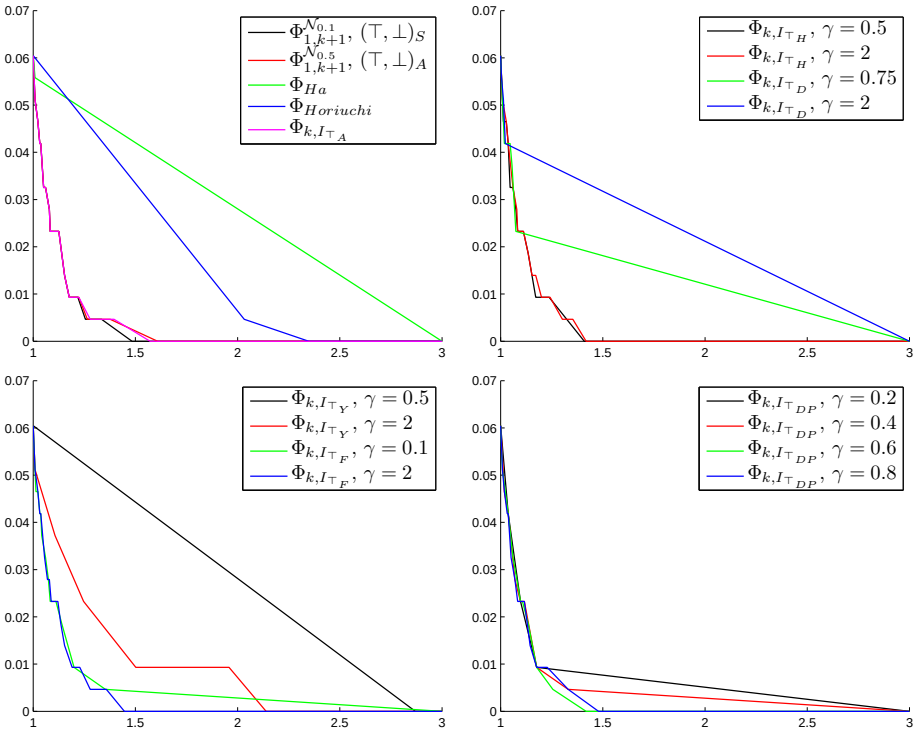

FIG. D.18: Courbes $E\overline{n}$ sur les données *Thyroid*. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

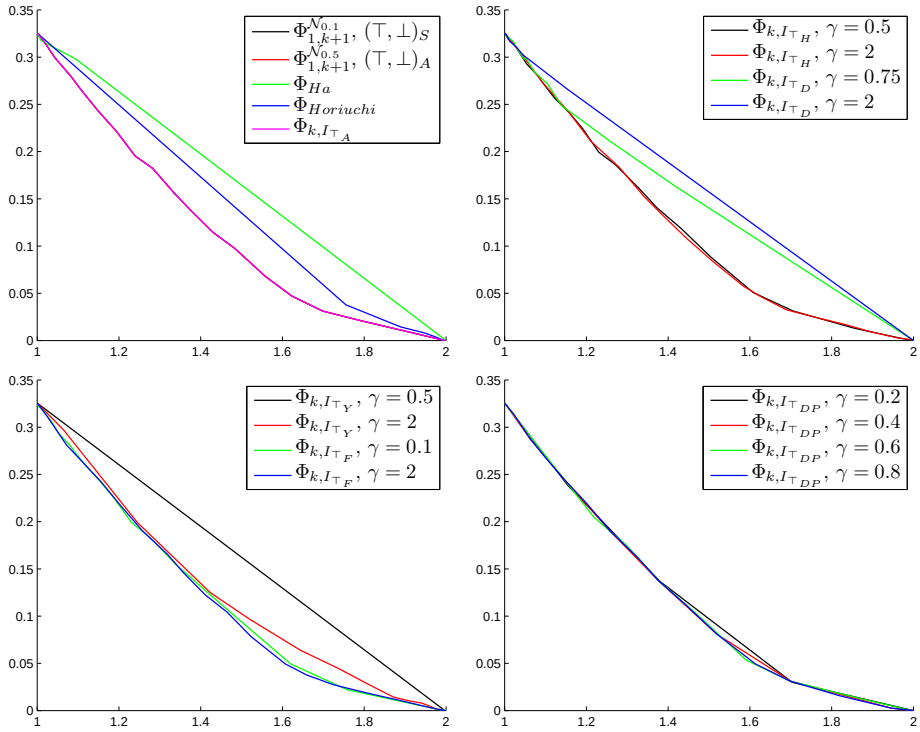

FIG. D.19: Courbes $E\overline{n}$ sur les données *Pima*. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

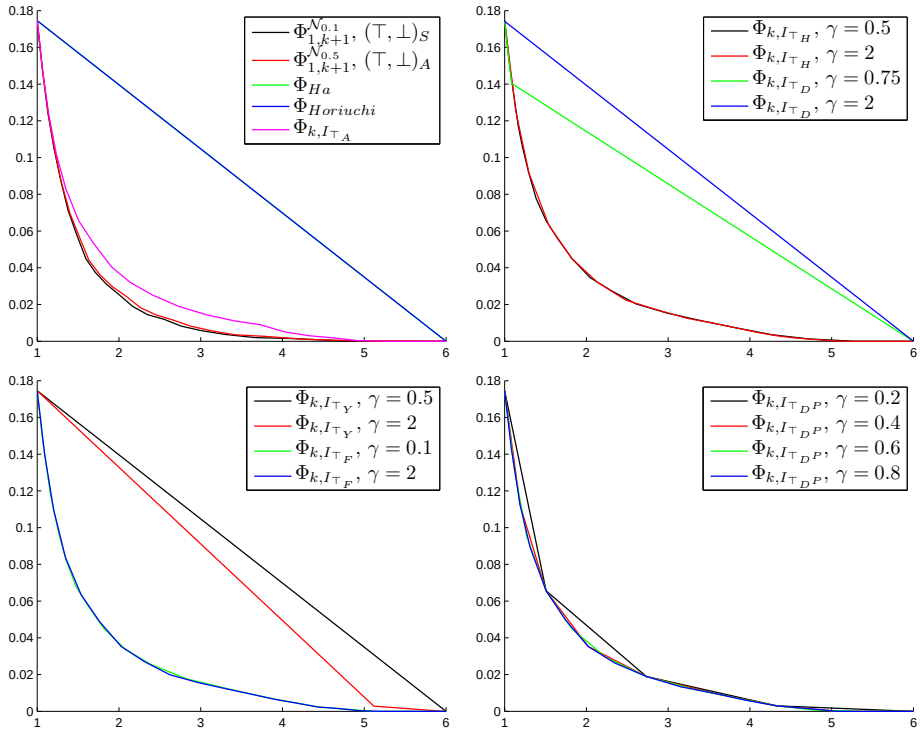

FIG. D.20: Courbes $E\overline{n}$ sur les données *Statlog*. Mesures de Ha et Horiuchi ainsi que ambiguïté par blocs (*haut-gauche*), mesures paramétrique de Hamacher et Dombi, Yager et Frank, Dubois-Prade, respectivement (*haut-droite*), (*bas-gauche*), (*bas-droite*).

Annexe E

Sélection d'articles

E.1 A fuzzy modeling approach to cluster validity

Cet article a été publié comme :
H. Le Capitaine & C. Frélicot. A fuzzy modeling approach to cluster validity. In Proc. IEEE Int. Conf. on Fuzzy Systems, pages 462-467. Jeju Island, Korea, 2009.

A Fuzzy Modeling Approach to Cluster Validity

Hoel Le Capitaine and Carl Frélicot

***Abstract*— This paper presents a new approach to find the optimal number of clusters of a fuzzy partition. It is based on a fuzzy modeling approach which combines measures of clusters' separation and overlap. Theses measures are based on triangular norms and a discrete Sugeno integral. Results on artificial and real data sets prove its efficiency compared to indexes from the literature.**

I. Introduction

Clustering aims at detecting natural groups (or clusters) in multidimensional data sets. The principle is that data points within a cluster are as similar as possible whereas data points of different clusters as dissimilar as possible. Since clusters may have different shapes and sizes, a partition resulting from this unsupervised classification process needs to be validated. A key point is the number of clusters and many Cluster Validity Indexes (CVIs) have been proposed, in particular for fuzzy clustering, see [3], [11], [19].

In this paper, we propose a new index following a fuzzy system modeling approach using measures of separation and overlap degree of the data points from the obtained clusters. Theses measures are based on dedicated operators, based on triangular norms and a discrete Sugeno integral, that aggregate the clusters fuzzy labels provided by the clustering algorithm.

II. Cluster Validity for Fuzzy Clustering

A. The Fuzzy C-Means algorithm

Let $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ be a n points data set in a p-dimensional feature space, say $\mathbb{R}^p$, with the usual euclidean norm $||.||$. The fuzzy c-means (FCM) algorithm partitions X into $c > 1$ clusters by minimizing the following objective function [2]:

$$J_m(U, V, X) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m ||\mathbf{x}_k - \mathbf{v}_i||^2 \qquad (1)$$

where $u_{ik} \in [0, 1]$ is the membership degree of $\mathbf{x}_k$ to the i^{th} cluster represented by its centroid $\mathbf{v}_i \in \mathbb{R}^p$. Centroids are gathered into a $(c \times p)$ matrix $V = [\mathbf{v}_1, \cdots, \mathbf{v}_c]$. Degrees u_{ik} are subject to a normalization constraint $\sum_{i=1}^{c} u_{ik} = 1$ for all $\mathbf{x}_k$ in X, and are elements of the fuzzy c-partition $U(c \times n)$ whose columns $\mathbf{u}_k = {}^t(u_{1k}, \cdots, u_{ck})$ are the membership vectors of each $\mathbf{x}_k$. The so-called fuzzifier $m > 1$ is a weighting exponent which makes the resulting partition more or less fuzzy: the higher m, the softer the cluster boundaries are. Minimization of (1) is obtained by iteratively updating (U, V) as follows:

$$u_{ik} = 1 \Big/ \sum_{j=1}^{c} \left(\frac{||\mathbf{x}_k - \mathbf{v}_i||}{||\mathbf{x}_k - \mathbf{v}_j||} \right)^{2/(m-1)} \qquad (2)$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{n} u_{ik}^m} \qquad (3)$$

The usual euclidian norm $||.||$ induces hyper-spherical clusters, hence FCM can only detect clusters with the same shape and orientation. Thus, such clusters description is not well suited to every possible situation, e.g.: bridges, outliers, additional noisy points. Cluster Validity (CV) is then a more challenging problem using FCM instead of other algorithms that behave better in such situations.

B. Cluster validity procedure and classical fuzzy indexes

Validating the provided clustering (U, V) of X consists in assessing whether the resulting partition reflects the data structure or not. Since clustering is unsupervised, no prior knowledge on the data is taken into account, and the number c of clusters is a user-defined parameter for clustering algorithms such as FCM. Most of works on cluster validity focus on the number of clusters problem and many CVIs have been proposed, refer to [3], [11] for comparative studies. Given a CVI, the procedure to automatically select the optimal number of clusters c_{best} in a predefined range $[c_{min}, c_{max}]$ and therefore the best partition is as follows:

(1) choose values c_{min} and c_{max}
(2) for $c = c_{min}$ to c_{max}
- run FCM
- compute CVI(c) from (X, U, V)
(3) select c_{best} such as CVI(c_{best}) is optimal and take the corresponding partition (U, V)

CVIs can be classified either according to the type of information they handle (only membership degrees to clusters vs additional information on the geometrical structure of clusters) or to cluster properties (compactness within each cluster and/or separation between clusters). Note these categories are not mutually exclusive and most indexes present advantages and drawbacks. Earliest CVIs only use partial membership degrees (U). Let us cite the *Partition Coefficient* [2], taking values in $[\frac{1}{c}, 1]$:

$$PC(c) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^2 \qquad (4)$$

or the *Partition Entropy*, taking values in $[0, \log(c)]$:

$$PE(c) = -\frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \, log(u_{ik}) \qquad (5)$$

Hoel Le Capitaine/Carl Frélicot are with the MIA Laboratory, University of La Rochelle, Avenue Michel Crépeau, France (phone: +33 546 458 322/234; email: {hlecap01},{carl.frelicot}@univ-lr.fr).

978-1-4244-3597-5/09/$25.00 ©2009 IEEE

Authorized licensed use limited to: CNUDST. Downloaded on October 18, 2009 at 08:20 from IEEE Xplore. Restrictions apply.

Both *PC* to be maximized and *PE* to be minimized are monotonic with c, as well as their bounds. Normalized versions have been proposed to overcome these drawbacks. For the experiments in section IV, we use *NPC* [18], [6] and *NPE* [7] defined by:

$$NPC(c) = \frac{c\,PC(c) - 1}{c - 1} \tag{6}$$

$$NPE(c) = \frac{n\,PE(c)}{n - 1} \tag{7}$$

The second category consists of indexes that use membership degrees but also some information about the geometrical structure of the data (U, V, X), *e.g.* the Xie-Beni index [22]:

$$XB(c) = \frac{J_m(U,V)\,/n}{\min_{i,j=1,c;j\neq i}\ ||\mathbf{v}_i - \mathbf{v}_j||^2} \tag{8}$$

or the Fukuyama-Sugeno index [8]:

$$FS(c) = J_m(U,V) - \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m\,||\mathbf{v}_i - \overline{\mathbf{v}}||^2 \tag{9}$$

where $\overline{\mathbf{v}}$ is the mean of centroids. Both *XB* and *FS* combine the FCM objective function (1) which measures how much the clusters are *compact* and an additional term which measures how much they are *separated*. Combination indicates that both indexes are to be minimized. The more compact and separated the clusters are, the less fuzzy and the more crisp the partition is, therefore the more optimal c is. In [9], Gath and Geva propose the Fuzzy Hyper Volume (*FHV*) defined by:

$$FHV(c) = \sum_{i=1}^{c} \sqrt{\det(C_i)} \tag{10}$$

where $C_i = \frac{\sum_{k=1}^{n} u_{ik}^m (\mathbf{x}_k - \mathbf{v}_i)^{\,t}(\mathbf{x}_k - \mathbf{v}_i)}{\sum_{k=1}^{n} u_{ik}^m}$ is the fuzzy covariance matrix of the i^{th} cluster. This index should be low when clusters are compact, so the optimal number of clusters can be found by minimization of (10). In order to decrease the tendency of *XB* (8) to monotonically decrease when c tends to n, Kwon add a penalty function [13], yielding to an index to be minimized:

$$K(c) = \frac{J_m(U,V) + \frac{1}{c}\sum_{i=1}^{c} ||\mathbf{v}_i - \overline{\mathbf{v}}||^2}{\min_{i,j=1,c;j\neq i}\ ||\mathbf{v}_i - \mathbf{v}_j||^2} \tag{11}$$

where $\overline{\mathbf{v}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$. Wu and Yang [21] propose a validity index defined by:

$$WY(c) = \sum_{i=1}^{c}\sum_{k=1}^{n} \frac{u_{ik}^2}{u_M} - \sum_{i=1}^{c} \exp\left(-\min_{j\neq i} \frac{||\mathbf{v}_i - \mathbf{v}_j||^2}{\beta_T}\right) \tag{12}$$

where $u_M = \min_{1\leq i\leq c}\left(\sum_{k=1}^{n} u_{ik}^2\right)$ is the compactness of the most compact cluster and $\beta_T = \frac{1}{c}\sum_{i=1}^{c} ||\mathbf{v}_i - \overline{\mathbf{v}}||^2$ is the total average distance measure for all clusters. A large value of *WY* means that the c clusters are compact and separated, so the optimal number of clusters can be found by maximizing (12). As compactness for each cluster is computed relative to u_M, *WY* gives good results in presence of outliers.

III. The Proposed Cluster Validity Index

A. Basic aggregation operators

Aggregation functions aim at associating a typical value to a number of several numerical values which are generally defined on a finite real interval or on ordinal scales. They are used in many fields, *e.g.* decision-making and pattern recognition an we assume in this paper, with no loss of generality, that they come from the unit interval. If not, a simple transformation can be found to make this true. Given c values, an aggregation operator is then a mapping Φ: $[0,1]^c \rightarrow [0,1]$, $\mathbf{a} = \{a_1, a_2, \ldots, a_c\} \mapsto \Phi(\mathbf{a})$. One finds many families, e.g.: triangular norms, OWA (*Ordered Weighted Averaging*) operators, γ-operators, or fuzzy integrals. They are classified either by mathematical properties they share or by the way the values are aggregated: conjunctives, disjunctives, compensatory, and weighted operators. An aggregation operator Φ is said to be *conjunctive* if $\Phi(\mathbf{a}) \leq min\{a_1, a_2, \ldots, a_c\}$, *disjunctive* if $\Phi(\mathbf{a}) \geq max\{a_1, a_2, \ldots, a_c\}$, and *compensatory* if $min\{a_1, a_2, \ldots, a_c\} \leq \Phi(\mathbf{a}) \leq max\{a_1, a_2, \ldots, a_c\}$, refer to [5] for a large survey on aggregation operators.

Triangular norms (t-norms) and conorms (t-conorms) are interesting conjunctive and disjunctive operators that generalize the logical AND and OR crisp operators to fuzzy sets. Briefly, a t-norm is a binary operation $\top : [0,1]^2 \rightarrow [0,1]$ which is commutative, associative, non decreasing and has 1 for neutral element. Alternatively, a t-conorm is the dual binary operation $\perp : [0,1]^2 \rightarrow [0,1]$ having the same properties except that its neutral element is 0, see [12] for a survey. Numerical results we give in in section IV are obtained using the *Standard*, *Algebraic* and *Hamacher* ($\gamma \in [0, +\infty[$) triangular norms respectively defined by:

- $a_1 \top_S a_2 = min(a_1, a_2)$ and $a_1 \perp_S a_2 = max(a_1, a_2)$
- $a_1 \top_A a_2 = a_1\,a_2$ and $a_1 \perp_A a_2 = a_1 + a_2 - a_1\,a_2$
- $a_1 \top_H a_2 = \frac{a_1\,a_2}{\gamma + (1-\gamma)(a_1 + a_2 - a_1\,a_2)}$ and $a_1 \perp_H a_2 = \frac{a_1 + a_2 - a_1\,a_2 - (1-\gamma)a_1\,a_2}{1 - (1-\gamma)a_1\,a_2}$.

The discrete Sugeno integral is a fuzzy aggregation operator which computes the mean value of a function with respect to a fuzzy measure m. It is a non-additive measure of uncertainty, i.e. more general than a possibility one and therefore a probability one. The integral of a function μ is defined by

$$\mathcal{S}_m = \bigvee_{i=1}^{n} \mu(x_i) \wedge m(A_{(i)}) \tag{13}$$

where $A_{(i)}$ is the fuzzy subset $\{x_{(i)}, \cdots, x_{(n)}\}$ with respect to a permutation so that $\mu(x_{(i)}) \leq \cdots \leq \mu(x_{(n)})$, and $\vee, \wedge$ are the maximum and minimum operators, respectively. This integral is widely used in decision making, and in particular for pattern recognition [10] because of its ability to model some kind of interaction between features describing a pattern $\mathbf{x}$.

B. Separation of clusters

Most separation measures between clusters are based on the distances between cluster centroids, *e.g.* (8-9), but it is seldom sufficient to interpret the geometrical structure of the

Authorized licensed use limited to: CNUDST. Downloaded on October 18, 2009 at 08:20 from IEEE Xplore. Restrictions apply.

data, see [11] for examples. We propose to define a rule for the separation of clusters, for each point $\mathbf{x}_k$, based on its membership degrees $\mathbf{u}_k$ (k^{th} column of U), using the $l-$order fuzzy OR operator (fOR-l) and the fuzzy exclusive OR operator (fXOR), both defined in [15] for the supervised classification problem with reject options. Let $\mathcal{P}$ be the power set of $C = \{1, 2, ..., c\}$ and $\mathcal{P}_l = \{A \in \mathcal{P} : |A| = l\}$ where $|A|$ denotes the cardinality of subset A, then the fOR-l associates to $\mathbf{u}_k$ a single value $\overset{l}{\perp}(\mathbf{u}_k) \in [0,1]$ defined by:

$$\overset{l}{\perp}(\mathbf{u}_k) = \overset{l}{\underset{i=1,\cdots,c}{\perp}} u_{ik} = \underset{A\in\mathcal{P}_{l-1}}{\top} \left(\underset{j\in C\setminus A}{\perp} u_{jk} \right) \tag{14}$$

It must be viewed as some kind of generalization of the notion of "l^{th} highest" value, with $l \in C$. Using standard t-norms, $\overset{l}{\perp}(\mathbf{u}_k)$ is exactly the "l^{th} highest" element of $\mathbf{u}_k$. Given a fuzzy complement, *e.g.* $\overline{a_1} = 1 - a_1$, the fXOR associates a single value $\underline{\perp}(\mathbf{u}_k)$ to $\mathbf{u}_k$ defined by:

$$\underline{\perp}(\mathbf{u}_k) = \underset{i=1,c}{\underline{\perp}} u_{ik} = \overset{1}{\perp}(\mathbf{u}_k) \top \overline{\overset{2}{\perp}(\mathbf{u}_k) / \overset{1}{\perp}(\mathbf{u}_k)} \tag{15}$$

The value of fXOR is "high" if the "highest" value is large enough compared to the second "highest" one. Therefore, we introduce the following rule for separation, with respect to each $\mathbf{x}_k$:

if u_{1k} *is high xor* $\cdots$ *xor* u_{ck} *is high,*
then $\mathbf{x}_k$ *belongs to a well separated cluster*

Using (14-15), it can be formally expressed by the satisfaction level $\tau^{(s)}(\mathbf{x}_k) = \underline{\perp}(\mathbf{u}_k)$, denoted $\tau_k^{(s)}$ for short.

C. Overlap of clusters

Non monotonic CVIs succeed on well separated clusters but they generally fail when some clusters naturally overlap. We propose to define a rule for the degree of overlap of clusters, for each point $\mathbf{x}_k$, based on its membership degrees $\mathbf{u}_k$ using the aggregation operator $\Phi_{i,j}$ defined by the ratio of two Sugeno integrals in [14] for the supervised classification problem with reject options. Assuming each $\mathbf{u}_k$ to be sorted in descending order, *i.e.* $u_{1k} \geq u_{2k} \geq \cdots \geq u_{ck}$, let the blockwise similarity operator $\Phi_{i,j}$ taking values in $[0,1]$, which quantifies the similarity of the block of values $\{u_{ik}, \cdots, u_{jk}\}$ in $\mathbf{u}_k$, defined by:

$$\Phi_{i,j}(\mathbf{u}_k) = \begin{cases} \dfrac{\overset{j}{\underset{l=\frac{i+j}{2}}{\perp}} u_{lk} \top \mathcal{K}_\lambda(l,j)}{\overset{i}{\underset{l=\frac{i+j}{2}}{\perp}} u_{lk} \top \mathcal{K}_\lambda(l,i)} & \text{if } j-i \text{ is even} \\ \dfrac{\overset{j}{\underset{l=\frac{i+j+1}{2}}{\perp}} u_{lk} \top \mathcal{K}_\lambda(l,j)}{\overset{i}{\underset{l=\frac{i+j-1}{2}}{\perp}} u_{lk} \top \mathcal{K}_\lambda(l,i)} & \text{if } j-i \text{ is odd} \end{cases} \tag{16}$$

where $\mathcal{K}_\lambda(l,j)$ is a symmetrical kernel at resolution level $\lambda \in \mathbb{R}^+$, *e.g.* a gaussian kernel $\mathcal{N}_\lambda(l,j)$, taking values in [0,1] such that values within the block more or less contribute. The larger λ is, the larger the contribution is, so increasing λ makes two consecutive values more similar but can increase the similarity of blocks of larger size, see [14] for proofs and details. A high value of $\Phi_{1,j}(\mathbf{u}_k)$ reveals that the j "highest" values are similar. Therefore, we introduce the following rule for degree of overlap, with respect to $\mathbf{x}_k$:

if $\Phi_{1,2}(\mathbf{u}_k)$ *is high or* $\cdots$ *or* $\Phi_{1,c}(\mathbf{u}_k)$ *is high,*
then $\mathbf{x}_k$ *belongs to overlapping clusters*

Formally, it can be expressed, using (16), by the satisfaction level $\tau^{(o)}(\mathbf{x}_k) = \overset{1}{\underset{j=1,c}{\perp}} \Phi_{1,j}(\mathbf{u}_k)$, denoted $\tau_k^{(o)}$ for short.

D. Fuzzy modeling and the new CVI

In order to build up a new CVI, we propose to follow the so-called *fuzzy system modeling technique* [23], issued from the fuzzy control community. It expresses in terms of a two rules knowledge base:

1) *if* $\mathbf{x}_k$ *belongs to a well separated cluster, then use the most satisfied corresponding measure*
2) *if* $\mathbf{x}_k$ *does not belong to a well separated cluster or belongs to overlapping clusters, then use the least satisfied corresponding measure*.

Introducing two concepts *high* and *low*, as well as a measure $SO_\top(\mathbf{u}_k)$ (for *Separation–Overlap*), with respect to each point $\mathbf{x}_k$, depending on the dual couple (t-norm,t-conorm) at hand the corresponding fuzzy model gives:
Rule 1: if $\tau_k^{(s)}$ is *high* and $\tau_k^{(o)}$ is *low*, then $SO_\top(\mathbf{u}_k)$ is B_1
Rule 2: if $\tau_k^{(s)}$ is *low* or $\tau_k^{(o)}$ is *high*, then $SO_\top(\mathbf{u}_k)$ is B_2
where B_1 and B_2 are the following singleton fuzzy subsets:

$$B_1 = \left\{ 1 \Big/ \left(\tau_k^{(s)} \perp \tau_k^{(o)} \right) \right\} \text{ and } B_2 = \left\{ 1 \Big/ \left(\tau_k^{(s)} \top \tau_k^{(o)} \right) \right\}.$$

Since t-norms and t-conorms model the *and* and *or* connectives, respectively, we get for the firing strength of *Rule 1* and *Rule 2*:

$$\rho_1 = \tau_k^{(s)} \top \overline{\tau_k^{(o)}} \text{ and } \rho_2 = \overline{\tau_k^{(s)}} \perp \tau_k^{(o)}.$$

Then, by aggregating the effective output of each rule, we obtain the following system output:

$$O = \left\{ \frac{\rho_1}{\tau_k^{(s)} \perp \tau_k^{(o)}}, \frac{\rho_2}{\tau_k^{(s)} \top \tau_k^{(o)}} \right\}.$$

Depending the defuzzification method, several operators can be obtained. Using *center of area* method, we get the measure:

$$SO_\top(\mathbf{u}_k) = \frac{\rho_1 \times \left(\tau_k^{(s)} \perp \tau_k^{(o)} \right) + \rho_2 \times \left(\tau_k^{(s)} \top \tau_k^{(o)} \right)}{\rho_1 + \rho_2}.$$

Finally, we define the new CVI, called $SOI_\top$ (*Separation–Overlap Index*), by simply averaging the resulting measure on X, taking values in $[0,1]$, as follows:

$$SOI_\top(c) = \frac{1}{n} \sum_{k=1}^{n} SO_\top(\mathbf{u}_k) \tag{17}$$

The more separated and not overlapping the clusters are, the more $SOI_\top(c)$ is. Maximization of (17) gives the optimal

Authorized licensed use limited to: CNUDST. Downloaded on October 18, 2009 at 08:20 from IEEE Xplore. Restrictions apply.

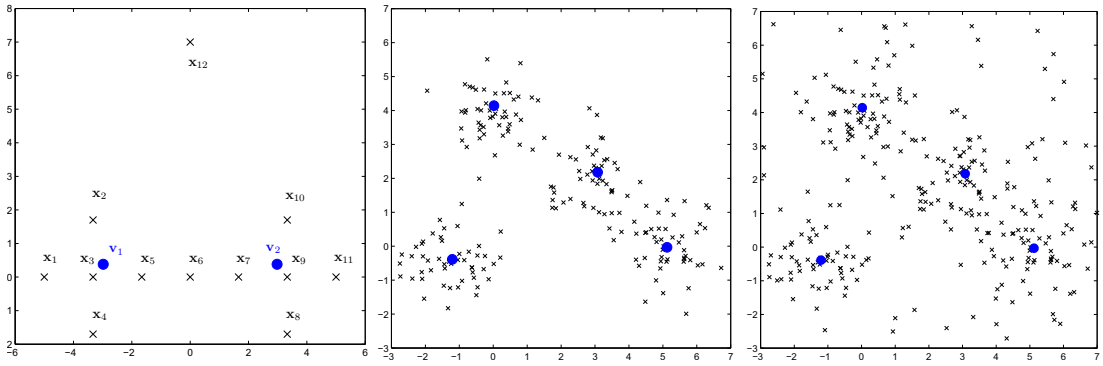

Fig. 1. Artificial data sets: *Diamond+ (left)*, D_1 *(center)* and D_2 *(right)*, and their associated centroids represented in blue dots.

TABLE I

OVERLAP AND SEPARATION MEASURES FOR THE *Diamond+* DATA SET USING ALGEBRAIC T-NORMS AND T-CONORMS

	Point	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
$c=2$	u_{1k}	.937	.957	.993	.908	.919	.500	.080	.091	.006	.042	.062	.500
	u_{2k}	.062	.042	.006	.091	.080	.500	.919	.908	.993	.957	.937	.500
	$\tau_k^{(s)}$	.883	.917	.986	.833	.852	.500	.852	.833	.986	.917	.883	.500
	$\tau_k^{(o)}$	.066	.044	.006	.100	.087	.999	.087	.100	.006	.044	.066	.999
	$SO_{\top_S}(\mathbf{u}_k, c)$	.745	.812	.967	.658	.690	.499	.690	.658	.967	.812	.745	.499
$c=3$	u_{1k}	.912	.871	.999	.912	.872	.450	.086	.056	.000	.057	.045	.000
	u_{2k}	.045	.057	.000	.056	.086	.450	.872	.912	.999	.871	.912	.000
	u_{3k}	.042	.070	.000	.030	.040	.098	.041	.030	.000	.070	.042	.998
	$\tau_k^{(s)}$	.848	.791	.998	.847	.792	.550	.792	.847	.998	.791	.848	.997
	$\tau_k^{(o)}$	.135	.206	.001	.147	.220	.999	.220	.147	.001	.206	.135	.001
	$SO_{\top_S}(\mathbf{u}_k, c)$	.668	.585	.994	.664	.585	.550	.585	.664	.994	.585	.668	.994

number of clusters. Since each rule is defined with respect to a given dual couple (t-norm,t-conorm) the proposed cluster validity index is a family of indexes.

Proposition 1: if U is a hard c-partition, then $SOI_{\top}(c) = 1$ whatever c, for all couple $(\top, \perp)$.

Proof: for all k, $u_{ik} \in \{0,1\}$ and $\sum_{i=1}^{c} u_{ik} = 1$, then one value equals 1 while the others are 0, say $u_{1k} = 1$ and $u_{2k} = \cdots = u_{ck} = 0$. In this case, we have $\tau_k^{(s)} = 1$: as 0 is the absorbing element of $\top$, it is easy to check in Eq. (14) and $l = 2$ that $\overset{2}{\perp} = 0$. Therefore, replacing in Eq. (15), we have $\tau_k^{(s)} = 1 \top 1 = 1$, whatever c and $(\top, \perp)$.
We also have $\tau_k^{(o)} = 0$: here again, since 0 is the absorbing element of $\top$, we have $\Phi_{1,j} = \mathbf{0}/(\mathbf{0} \perp 1 \top 1) = 0$, for $j \geq 2$, whatever $(\top, \perp)$. Consequently, we obtain $\rho_1 = 1 \top 1 = 1$, and $\rho_2 = 0 \perp 0 = 0$.
Finally, we get $SO_{\top}(\mathbf{u}_k) = \frac{1 \times (1 \perp 0) + 0 \times (1 \top 0)}{1+0} = 1$, which gives $SOI_{\top}(c) = 1$, whatever c, for all $(\top, \perp)$. □

Proposition 2: if U is a totally fuzzy c-partition, then $SOI_{\top_S}(c) = 0$ whatever c.

Proof: for all k, $u_{ik} = \frac{1}{c}$. By properties of $\Phi_{i,j}$, we have $\tau_k^{(o)} = 1$ whatever c and $(\top, \perp)$. Thus, we obtain $\rho_1 = 0 \top \tau_k^{(s)} = 0$, and $\rho_2 = 1 \perp \overline{\tau_k^{(s)}} = 1$.
Finally, we get $SO_{\top}(\mathbf{u}_k) = \frac{0 \times (\tau_k^{(s)} \perp 1) + 1 \times (\tau_k^{(s)} \top 1)}{1+0} = \tau_k^{(s)}$.
Using $(\top, \perp)_S$, we have $\overset{1}{\perp} \mathbf{u}_k = \overset{2}{\perp} \mathbf{u}_k = \frac{1}{c}$ by (14) and $\tau_k^{(s)} = min\{\frac{1}{c}, 1-1) = 0$, so that $SOI_{\top_S}(c) = 0$, whatever c. □

Let us illustrate the ability of the proposed index to find the right number of clusters and the right partition on an small example, inspired by [17], that we call *Diamond+*. It consists of the eleven two-dimensional points first introduced by Windham [20] and an outlier, see Figure 1-left. Besides the outlier, the correct partition is composed of $c^\star = 2$ clusters: the two touching diamond shapes. CVIs that only consider compactness and separation will select three clusters: the two touching diamond shapes and the outlier. Table I gives the detailed values of the satisfaction levels $\tau_k^{(s)}$ and $\tau_k^{(o)}$ as well as the measures $SO_{\top}(\mathbf{u}_k)$ for the twelve points, with $c = 2$ and $c = 3$. The algebraic triangular norms are used and the values of the new index are $SOI_{\top_A}(2) = 0.729$ and $SOI_{\top_A}(3) = 0.711$ showing that it recovers the natural partition in two clusters. In terms of separation and overlap, two points are of particular interest, namely $\mathbf{x}_6$ and $\mathbf{x}_{12}$.

For $c = 2$, both $\mathbf{x}_6$ and $\mathbf{x}_{12}$ are the most overlapping points (0.999) and the most isolated points (0.500) as expected. The point $\mathbf{x}_6$ is equidistant to centroids $\mathbf{v}_1$ and $\mathbf{v}_2$, so that it induces high overlapping and low separation. The point $\mathbf{x}_{12}$ is equidistant to centroids $\mathbf{v}_1$ and $\mathbf{v}_2$, but farther than $\mathbf{x}_6$. Due to the sum constraint on $\mathbf{u}_k$, we have $u_{1,12} = u_{2,12} = 0.5$ so it has a high overlap and a low separation measure. For $c = 3$, $\mathbf{x}_{12}$ becomes a well separated and not overlapping cluster, but

Authorized licensed use limited to: CNUDST. Downloaded on October 18, 2009 at 08:20 from IEEE Xplore. Restrictions apply.

TABLE II
VALIDITY INDEXES ON ARTIFICIAL DATA D_1 (OPTIMAL VALUES ARE IN BOLD).

c	NPC	NPE	XB	FS $\times 10^{-3}$	K $\times 10^{-3}$	FHV	WY	$SOI_{\top}$ S	A	H_0
2	0.78	**0.50**	0.13	-1.48	0.32	3.97	1.64	0.67	0.58	0.58
3	0.80	0.54	**0.07**	-3.25	**0.18**	2.24	1.82	0.74	0.65	0.69
4	**0.81**	0.57	0.08	-2.95	0.20	**1.83**	**2.20**	**0.79**	**0.69**	**0.72**
5	0.73	0.78	0.29	**-4.23**	0.51	2.11	0.09	0.72	0.65	0.62
6	0.68	0.93	0.37	-1.02	0.54	2.35	-0.79	0.65	0.63	0.58
7	0.64	1.07	0.46	-3.75	0.61	2.28	-1.72	0.64	0.63	0.58
8	0.62	1.16	0.27	-0.88	0.57	2.25	0.99	0.57	0.65	0.58
9	0.59	1.25	0.32	-1.75	0.59	2.35	0.18	0.57	0.61	0.57
10	0.56	1.34	0.28	-1.17	0.55	2.32	-0.05	0.56	0.62	0.54

TABLE III
VALIDITY INDEXES ON NOISY DATA D_2 (OPTIMAL VALUES ARE IN BOLD).

c	NPC	NPE	XB	FS $\times 10^{-3}$	K $\times 10^{-3}$	FHV	WY	$SOI_{\top}$ S	A	H_0
2	**0.74**	**0.58**	0.18	-1.28	0.64	6.19	1.69	0.68	0.57	0.56
3	0.72	0.72	**0.11**	**-4.10**	**0.45**	4.48	**2.24**	0.71	0.63	0.58
4	0.71	0.83	0.13	-0.74	0.54	**4.10**	1.91	**0.74**	**0.66**	**0.62**
5	0.63	1.08	0.41	-3.44	0.94	5.12	0.84	0.68	0.63	0.60
6	0.59	1.21	0.40	-1.00	1.43	5.06	-0.23	0.59	0.59	0.59
7	0.56	1.34	0.42	-2.09	1.48	5.07	-0.95	0.60	0.58	0.57
8	0.55	1.41	0.40	-3.09	1.38	5.04	0.16	0.57	0.58	0.56
9	0.54	1.50	0.31	-0.99	1.37	5.25	-0.42	0.55	0.55	0.56
10	0.54	1.51	0.28	-1.01	1.27	5.29	-1.11	0.54	0.56	0.57

$\mathbf{x}_6$ remains the most overlapping point. The point $\mathbf{x}_6$ is still equidistant to $\mathbf{v}_1$ and $\mathbf{v}_2$, but is far from the third centroid $\mathbf{v}_3$, which is near $\mathbf{x}_{12}$, so that $u_{3,6}$ is low. Thus, resulting separation and overlap measures for $\mathbf{x}_6$ are almost the same than in case of $c = 2$. The point $\mathbf{x}_{12}$ is close to $\mathbf{v}_3$, leading to a high $u_{3,12}$ value, and low membership degree to the two previous clusters. This obviously gives a high separation and a low overlap value for $\mathbf{x}_{12}$. Averaging all $SO_{\top}(\mathbf{u}_k)$ values, in particular high values for points clearly belonging to clusters 1 and 2, makes the partition in $c = 3$ clusters less desirable than the one in $c = 2$ clusters, as shown by the proposed CVI values: $SOI_{\top_A}(2) \geq SOI_{\top_A}(3)$. This result clearly depends on the $\mathbf{x}_{12}$ position. If it was farther, the membership degrees of the other points to the third cluster would become lower, resulting in a better partition in $c = 3$ clusters. Furthermore, if one adds one more point near $\mathbf{x}_{12}$, it would create a third cluster, and the mean value of $SO_{\top}(\mathbf{u}_k)$ would increase, leading to select $c = 3$ as the optimal number of clusters.

IV. NUMERICAL RESULTS

In this section, we evaluate the performance of the proposed CVI by conducting a comparison to the seven CVIs using the procedure described in section II for data sets of various structures (good separation, overlapping clusters, presence of outliers, additional noisy points) making the CV problem more or less easy. FCM is used with $m = 2$ and a termination parameter set to 10^{-3}. The optimal number of clusters is searched in the range $[c_{min} = 2, c_{max}]$ with $c_{max} = 10$ for data sets with high cardinality (n) and an integer value close to $\sqrt{n}$ for the others as usually done in the literature. For $SOI_{\top}$, we use a gaussian kernel $\mathcal{N}_{\lambda}(l, j) = \exp\left(-(l-j)^2/\lambda\right)$ in (16) with $\lambda = 5$ so that all the values in the blocks of size c have a high contribution to $\Phi_{1,c}$. Consequently, the choice of the t-norm has a little effect on the optimal value, otherwise the standard one tends to select a smaller number of clusters because it is not archimedean[1].

A. Artificial data sets

We generated an artificial data set D_1 containing $n = 200$ points composed of four bivariate gaussian distributions, two of them slightly overlapping, see Figure 1-center. Table II reports the obtained results for the tested CVIs, where bold values are the optimal ones indicating the selected number of clusters. We see that, for any t-norm, $SOI_{\top}$ finds the optimal number of cluster while some classical indexes fail.

To emphasize the robustness to noisy data, we added 100 points drawn from a uniform distribution to generate another data set D_2. This additional noisy points can make the FCM algorithm partitioning this second artificial data set into three clusters because the two least separated clusters tend to become one cluster with noise, see Figure 1-right. The corresponding results are given in Table III. Again, $SOI_{\top}$ outperforms the classical indexes, none of them except *FHV* finding the optimal number of clusters (four).

B. Benchmark data sets

Additionally to data sets *Diamond+*, D_1 and D_2, we compare $SOI_{\top}$ on artificial and real benchmark data sets

[1] a t-norm $\top$ is archimedean if $\top(a,a) < a, \forall a \in [0, 1]$

Authorized licensed use limited to: CNUDST. Downloaded on October 18, 2009 at 08:20 from IEEE Xplore. Restrictions apply.

TABLE IV

OPTIMAL NUMBER OF CLUSTERS SELECTED BY THE TESTED CVIS ON ARTIFICIAL AND BENCHMARK DATA SETS.

Data set	c_{max}	NPC	NPE	XB	FS	K	FHV	WY	$SOI_\top$			$c^\star$
									S	A	H_0	
Diamond+	4	3	2	3	3	3	4	3	2	2	2	2
D_1	10	4	2	3	5	3	4	4	4	4	4	4
D_2	10	2	2	3	3	3	4	3	4	4	4	4
X30	5	3	3	3	4	2	3	3	3	3	3	3
Bensaid	7	3	2	3	7	3	3	6	3	3	3	3
Iris	10	2	2	2	3	2	3	2	2	2	2	2 or 3
Starfield	10	2	2	6	7	3	9	3	8	8	8	8 or 9
Wine	10	3	2	2	10	2	3	3	3	3	3	3
Soybean	7	3	2	3	4	3	2	3	3	4	4	4

commonly used in the literature:

• The artificial data set *X30* introduced in [3], consisting in 30 observations in $\mathbb{R}^2$, for which 3 clusters are expected.

• The bi-dimensional artificial data set *Bensaid* [1] characterized by 3 classes of different cardinalities (6, 3 and 40).

• *Iris* [4], composed of 3 classes of 50 flowers each described by 4 physical attributes. Two classes have a substantial overlap in the feature space and the optimal number of clusters to be found is debatable: 2 or 3, see [3].

• The original *Starfield* [22], which contains the position and light intensity of 51 bright stars near Solaris. The expected number of clusters is 8 or 9, depending on the papers.

• *Wine* [4], which consists of 13 chemical attributes for 178 Italian wines belonging to 3 separable classes.

• *Soybean-small* [4], composed of four classes characterizing various diseases affecting soybean plants. Each of the 47 observations is described by 35 attributes.

The optimal number $c^\star$ of clusters for each data set and the one found by the different CVIs are given in Table IV. The proposed index always finds the optimal number of clusters except for the *Soybean-small* data set and standard t-norms because of the common value of λ, while the other do not.

V. CONCLUSION

We introduce a new family of cluster validity indexes for fuzzy partitions. They use new measures of separation and degree of overlap of the clusters based on triangular norms and a discrete Sugeno integral. Another novelty is that they lie on a fuzzy system modeling technique expressed in terms of a two rules knowledge base allowing to take into account the relative importance of each measure. Results on benchmark data sets of various structures show that the proposed indexes are more efficient than a large collection of indexes from the literature on cluster validity. Recommendations on the choice of the t-norms and the kernel function will be the subject of a longer forthcoming paper where we will show how their mathematical properties affect the indexes.

REFERENCES

[1] A. M. Bensaid et. al., "Validity-Guided (Re)Clustering with Applications to Image Segmentation", *IEEE Trans. on Fuzzy Systems*, vol. 4, no. 2, pp. 112-123, 1996.

[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, 1981.

[3] J.C. Bezdek and N.R. Pal, "Some new indexes of cluster validity", *IEEE Trans. on Systems, Man and Cybernetics*, vol. 28, no. 3, pp. 301-315, 1998.

[4] C. Blake and C. Merz, UCI Repository of machine learning databases, http ://www.ics.uci.edu/mlearn/MLRepository.html, 1998.

[5] T. Calvo and G. Mayor and R. Mesiar, Aggregation Operators: New Trends and Applications. Physica-Verlag, 2002.

[6] R.N. Davé, "Validating fuzzy partitions obtained through c-shells clustering", *Pattern Recognition Letters*, vol. 17, no. 6, pp. 613-623, 1996.

[7] J.C. Dunn, "Indices of partition fuzziness and the detection of clusters in large data sets", In: Fuzzy Automata and Decision processes, ed. by M.M. Gupta, Elsevier, 1977.

[8] Y. Fukuyama and M. Sugeno, "A new method for choosing the number of clusters for the fuzzy c-means method", *Proc. 5th Fuzzy Systems Symposium*, pp. 247-250, 1989.

[9] I. Gath and A.B. Geva, "Unsupervised optimal fuzzy clustering", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 773-780, 1989.

[10] M. Grabisch, "Fuzzy pattern recognition by fuzzy integrals and fuzzy rules", In: Pattern Recognition - From Classical to Modern Approaches, ed. by S. Pal and P. Pal, pp. 257-280, World Scientific, 2002.

[11] D-W. Kim, K.H. Lee and D. Lee, "On cluster validity index for estimating the optimal number of fuzzy clusters", *Pattern Recognition*, vol. 37, no. 3, pp. 2009-2025, 2004.

[12] E.P. Klement and R. Mesiar (Eds), Logical, algebraic, analytic and probabilistic aspects of triangular norms. Elsevier, 2005.

[13] S.H. Kwon, "Cluster validity index for fuzzy clustering", *Electronic Letters*, vol. 34, no. 22, pp. 2176-2177, 1998.

[14] H. Le Capitaine and C. Frélicot, "A class-selective rejection scheme based on blockwise similarity of typicality degrees", *Proc. 19th Int. Conf on Pattern Recognition*, 2008.

[15] L. Mascarilla, M. Berthier and C. Frélicot, "A k-order fuzzy OR operator for pattern classification with k-order ambiguity rejection", *Fuzzy Sets and Systems*, vol. 159, no. 2, pp. 2011-2029, 2008.

[16] L. Mascarilla and C. Frélicot, "A class of reject-first possibilistic classifiers based on dual triples", *Proc. 9th Int. Fuzzy Systems Association World Congress*, pp. 743-747, 2001.

[17] M-H. Masson and T. Denoeux, "ECM: An evidential version of the fuzzy c-means algorithm", *Pattern Recognition*, vol. 41, pp. 1384-1397, 2008.

[18] M. Roubens, "Pattern classification problems and fuzzy sets", *Fuzzy Sets and Systems*, vol. 1, pp. 239-253, 1978.

[19] W. Wang and Y. Zhang, "On fuzzy cluster validity indices", *Fuzzy Sets and Systems*, vol. 158, no. 19, pp. 2095-2117, 2007.

[20] M.P. Windham, "Numerical classification of proximity data with assignment measure", *J. of Classification*, vol. 2, pp. 157-172, 1985.

[21] K.L.Wu and M.S. Yang, "A cluster validity index for fuzzy clustering", *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1275-1291, 2005.

[22] X.L. Xie and G. Beni, "A validity measure for fuzzy clustering", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 841-847, 1991.

[23] R.R. Yager and D.P. Filev, Essentials of Fuzzy Modeling and Control. Wiley, 1994.

[24] H-J. Zimmerman and P. Zysno, "Quantifying vagueness in decision models", *European J. of Operational Research*, vol. 22, no. 2, pp. 148-158, 1985.

Authorized licensed use limited to: CNUDST. Downloaded on October 18, 2009 at 08:20 from IEEE Xplore. Restrictions apply.

E.2 Segmentation d'images couleur par des mesures de chevauchement et de séparation fondées sur l'agrégation de partition floue

Cet article a été publié comme :
H. Le Capitaine & C. Frélicot. Segmentation d'images couleur par des mesures de chevauchement et de séparation fondées sur l'agrégation de partition floue. To appear In Rencontres Francophones sur la Logique Floue et ses Applications LFA09. Annecy, France, 2009.

Segmentation d'images couleur par des mesures de chevauchement et de séparation fondées sur l'agrégation de partition floue

Color image segmentation using an overlap and a separation measures based on the aggregation of a fuzzy partition

Hoel Le Capitaine ⋆ Carl Frélicot ⋆
⋆ Laboratoire Mathématiques, Images et Applications (MIA)
Université de La Rochelle, Avenue M. Crépeau, 17042 La Rochelle Cedex1

Résumé :

Cet article traite de la segmentation d'images par régions à partir d'une partition floue des pixels dans un espace colorimétrique. Quelle que soit la méthode de partitionnement utilisée, ici les Fuzzy C-Means avec contraintes spatiales, le nombre optimal de classes induisant le nombre de régions de couleur homogène doit être choisi ou déterminé. Nous proposons de le sélectionner à l'aide d'un indice de validité de partition floue fondé sur des mesures de chevauchement et de séparation qui agrègent les degrés d'appartenance des pixels de l'image aux classes de couleur. Les résultats obtenus, comparés à des méthodes bien connues fondées sur un partitionnement, et dont la sophistication est comparable, montrent l'intérêt de l'approche proposée.

Mots-clés :

segmentation, image couleur, agrégation, partition floue.

Abstract:

This article addresses the region-based image segmentation problem using the fuzzy partition of the pixels in a color space. Whatever the partitioning method, here the Fuzzy C-Means with spatial constraints, the optimal number of clusters inducing the number of homogeneous regions must be chosen or found. We propose to select it by means of a cluster validity index for fuzzy partitions based on an overlap and a separation measures which aggregate the membership degrees of the image pixels to the color clusters. Results we obtained, when compared to well-known partitioning-based methods, with similar degree of sophistication, show the interest of the proposed approach.

Keywords:

segmentation, color image, aggregation, fuzzy partition.

1 Introduction

La segmentation consiste à partitionner une image en zones homogènes selon les composantes colorimétriques décrivant l'image. C'est une étape importante en traitement d'image, puisqu'elle conduit à une caractérisation plus concise et plus sûre de l'image, et permet ainsi d'améliorer et d'accélérer les traitements ultérieurs [18]. De nombreuses approches ont été proposées pour la segmentation d'images. Parmi celles-ci, on distinguera les approches contours et les approches régions. L'approche région de la segmentation d'images consiste à diviser une image en régions distinctes de telle sorte que des pixels d'une même région sont homogènes et des pixels de régions différentes ne le sont pas. Les algorithmes de partitionnement comme les C-Moyennes (K-means, [17]), ou les C-Moyennes Floues (FCM, [3]) sont largement utilisés dans ce but [1, 2] pour des images à niveaux de gris, ou pour des images couleurs [16]. Comme pour toute méthode de classification non supervisée, le nombre de classes pour FCM doit être fixé par l'utilisateur, rendant l'algorithme inefficace pour la segmentation de grandes banques d'images à moins de le déterminer de manière automatique. Les Indices de Validité de Classification (CVI) permettent d'obtenir le nombre optimal de classes à partir d'une partition, la partition correspondante étant alors réputée optimale [22, 4, 12]. A ce jour, l'utilisation d'indices de validité pour la segmentation d'image a concerné les images en niveaux de gris [8], ou se basent sur des processus de fusion de clusters par méthodes agglomératives [15]. Nous proposons dans cet article un indice de validité calculé sur une version modifiée de l'algorithme FCM incluant une contrainte spatiale, et permettant la segmentation d'images couleurs.

2 Partitionnement flou et segmentation d'image couleur

2.1 Algorithmes de partitionnement flou

Le partitionnement est une approche de la classification non supervisée ayant pour but de trouver une certaine structure de groupes dans un ensemble $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ de n points en dimension p. Dans ce cadre, les étiquettes des points $\mathbf{x}_1 \in X$ sont inconnues et les algorithmes de partitionnement ont pour objectif de les produire à partir de X. Par exemple, le très usité algorithme des *C-Moyennes Floues* (FCM, [3]) partitionne de manière itérative X en $c > 1$ groupes en minimisant la fonctionnelle suivante :

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m \, ||\mathbf{x}_k - \mathbf{v}_i||^2 \qquad (1)$$

où u_{ik} est le degré d'appartenance du point $\mathbf{x}_k$ au $i^{\text{ème}}$ groupe représenté par son centre $\mathbf{v}_i \in \mathbb{R}^p$, $||.||$ est la norme euclidienne usuelle, et $m > 1$ est un paramètre de pondération qui rend la partition résultante plus ou moins floue [22]. Plus m est grand, plus la partition résultante est floue ; la valeur $m = 2$ est la plus couramment utilisée. Les degrés sont calculés sous les contraintes $\sum_{i=1}^{c} u_{ik} = 1$ $(\forall k = 1, n)$ et $0 < \sum_{k=1}^{n} u_{ik} < n$ $(\forall i = 1, c)$. En considérant les matrices $U = [\mathbf{u}_1, ..., \mathbf{u}_n]$ et $V = [\mathbf{v}_1, ..., \mathbf{v}_c]$, FCM est donc une application : $X \mapsto (U, V)$. La minimisation de (1) est réalisée par itérations successives où (U, V) sont mises à jour selon :

$$u_{ik} = 1 \Big/ \sum_{j=1}^{c} \left(\frac{||\mathbf{x}_k - \mathbf{v}_i||}{||\mathbf{x}_k - \mathbf{v}_j||} \right)^{2/(m-1)} \qquad (2)$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m \, \mathbf{x}_k}{\sum_{k=1}^{n} u_{ik}^m} \qquad (3)$$

Notons que FCM fournit des groupes de forme hyper-sphérique de volumes égaux. Des variantes existent, en particulier celle consistant à introduire dans le processus de minimisation le calcul de matrices A_i induisant des normes permettant de détecter des groupes de forme hyper-ellipsoïdale, d'orientation et volumes différentes [10].

Si on considère chaque pixel d'une image comme un point $\mathbf{x}_k$ dans un espace colorimétrique [21], disons $\mathbb{R}^3$ de manière générique, alors (U, V) est une partition en r régions de l'image X, où $r \geq c$ puisque plusieurs régions non adjacentes de l'image peuvent avoir le même centre dans l'espace colorimétrique. Lorsque seuls les attributs colorimétriques sont utilisés, FCM ne donne pas toujours de bons résultats (ex. image bruitée) et une procédure de filtrage en aval est nécessaire. Une alternative consiste à ajouter des contraintes spatiales afin de lisser l'effet de pixels non homogènes dans chaque région, comme dans l'algorithme sFCM [11] où les degrés des pixels au voisinage de $\mathbf{x}_k$ sont utilisés pour modifier u_{ik} à chaque itération selon :

$$u_{ik} = \frac{u_{ik}^p h_{ik}^q}{\sum_{j=1}^{c} u_{jk}^p h_{jk}^q} \qquad (4)$$

où (p, q) sont deux paramètres contrôlant l'importance relative de u_{jk} et h_{jk}, et h est une fonction dont la forme générale est :

$$h_{ik} = \sum_{j \in \mathcal{V}(\mathbf{x}_k)} u_{ij} \qquad (5)$$

où $\mathcal{V}(\mathbf{x}_k)$ est l'ensemble des pixels voisins de $\mathbf{x}_k$), par exemple les pixels d'une fenêtre centrée en $\mathbf{x}_k$. Notons que sFCM se réduit à FCM si on fixe $(p, q) = (1, 0)$. C'est l'algorithme de partitionnement flou que nous utilisons pour la segmentation d'images couleur dont les résultats sont donnés à la section 4, avec $(p, q) = (1, 1)$.

2.2 Validation du partitionnement flou

Indépendamment de l'application visée (ici, la segmentation d'images couleur), tout algorithme de partitionnement (ici, sFCM) produit un résultat : $X \mapsto (U, V)$. Se pose alors le problème de sa validation. Mis à part la question de l'existence d'une structure de groupes pour X, il s'agit généralement de déterminer les paramètres optimaux de l'algorithme (ex : m,p et q pour sFCM) mais surtout le nombre c de groupes que doit fixer l'utilisateur. La

plupart des travaux concernent ce problème et la littérature regorge d'*Indices de Validation de Classification* (CVI) pour FCM et ses dérivés depuis plus de trois décennies. Ils sont souvent classifiés selon le type d'information qu'ils manipulent : U pour les indices les plus anciens [7, 23], (U, V) ou même (U, V, X) [25, 9, 1, 14][1], ou alors, et ce de manière non antinomique, selon la/les propriété(s) de la structure qu'ils privilégient : *compacité* C et/ou *séparation* S des groupes. Étant donné un CVI, la procédure de validation se réduit généralement à :

(1) choisir les bornes $c_{min} \geq 2$ et $c_{max} \leq n$ d'un intervalle de valeurs possibles pour c
(2) pour $c = c_{min}$ to c_{max} : exécuter FCM et calculer CVI(c) à partir de (X, U, V)
(3) sélectionner c_{opt} tel que CVI(c_{opt}) est optimal et valider le partitionnement correspondant (U, V).

Pour obtenir une image couleur segmentée en r_{opt} régions homogènes, il suffit alors de remplacer les trois composantes de chaque pixel $\mathbf{x}_k$ de l'image d'origine par celles du centre $\mathbf{v}_i$ optimal le plus proche ($i = argmax_{j=1:c_{opt}} u_{ik}$).

3 Un indice de validité fondé sur des mesures de chevauchement et de séparation

3.1 L'agrégation des degrés d'appartenance

Pour l'application visée, nous avons pris le parti d'utiliser uniquement l'information contenue dans la matrice partition U. Nous nous intéressons donc aux fonctions d'agrégation qui, à une collection $\mathbf{u}$ de c valeurs dans $[0, 1]$, associe une valeur dans $[0, 1]$, formellement : $\mathbf{u} = {}^t(u_1, \cdots, u_c) \mapsto \mathcal{A}(\mathbf{u})$. Parmi les fonctions les plus utilisées, on trouve, outre les moyennes, les normes et conormes triangulaires (t-norme et t-conorme, respectivement). Introduites afin de caractériser les opérateurs logiques multi-valués *ET* et *OU*, elles sont utilisés en logique floue afin d'implémenter des opérations conjonctives et disjonctives, respectivement. Une t-norme est une fonction commutative, associative et monotone $\top$ ayant pour élément neutre 1. Son opérateur dual, la t-conorme est une fonction commutative, associative et monotone $\perp$ ayant pour élément neutre 0. Voici par exemple, les normes *Standard*, *Algébrique* et de *Hamacher* ($\gamma \in [0, +\infty[$) définies respectivement par :

- $u_1 \top_S u_2 = min(u_1, u_2)$ et
  $u_1 \perp_S u_2 = max(u_1, u_2)$,
- $u_1 \top_A u_2 = u_1\, u_2$ et
  $u_1 \perp_A u_2 = u_1 + u_2 - u_1\, u_2$,
- $u_1 \top_H u_2 = \frac{u_1\, u_2}{\gamma+(1-\gamma)(u_1+u_2-u_1\, u_2)}$ et
  $u_1 \perp_H u_2 = \frac{u_1+u_2-u_1\, u_2-(1-\gamma)u_1\, u_2}{1-(1-\gamma)u_1\, u_2}$;

le lecteur pourra se référer à [13] pour une revue complète.

Étant donné un partitionnement (U, V) de X en c groupes, chaque u_{ik} définit la similarité de $\mathbf{x}_k$ au prototype $\mathbf{v}_i$. La t-conorme standard (opérateur max) est généralement appliqué aux composantes du vecteur $\mathbf{u}_k$ pour sélectionner le groupe auquel $\mathbf{x}_k$ doit être associé[2]. Malheureusement, les valeurs plus petites interagissent avec la plus grande et une telle association exclusive n'est pas efficace. Ainsi, il s'avère opportun d'évaluer à quel point $\mathbf{x}_k$ peut appartenir à plusieurs groupes et à combien. Nous proposons d'utiliser l'opérateur OU flou d'ordre l (fOU-l) défini dans [20] dans le cadre de la classification supervisée avec options de rejet. Soit $\mathcal{P}_l$ l'ensemble des parties de $\mathcal{C} = \{1, 2, ..., c\}$ de cardinalité l, alors fOU-l associe à $\mathbf{u}_k$ une valeur $\overset{l}{\perp}(\mathbf{u}_k) \in [0, 1]$ définie par :

$$\overset{l}{\perp}(\mathbf{u}_k) = \bigperp_{i=1,\cdots,c}^{l} u_{ik} = \bigtop_{A \in \mathcal{P}_{l-1}} \left(\bigperp_{j \in \mathcal{C} \setminus A} u_{jk} \right) \quad (6)$$

On peut montrer que :

- $\overset{1}{\perp}(\mathbf{u}_k) = \perp(\mathbf{u}_k)$ et $\overset{c}{\perp}(\mathbf{u}_k) = \top(\mathbf{u})$, pour tous $(\top, \perp)$ et $\forall c \in \mathcal{C}$

[1] pour ne citer que ceux auxquels le domaine se réfère sans cesse

[2] il en résulte une partition *stricte*, c'est-à-dire qu'elle correspond à une matrice de partition U telle que $u_{ik} \in \{0, 1\}$

- si on utilise les t-normes standard, alors $\overset{1}{\perp}_S(\mathbf{u}_k) = u_{(l)k}$, la l-ème plus grande valeur[3] dans $\mathbf{u}_k$; par exemple, prenons $\mathcal{C} = \{1, 2, 3\}$ et $l = 2$, alors $\mathcal{P}_{l-1} = \{\{1\}, \{2\}, \{3\}\}$ et $\overset{2}{\perp}_S(\mathbf{u}_k) = min(max(u_{2k}, u_{3k}), max(u_{1k}, u_{3k}), max(u_{1k}, u_{2k}))$, de sorte que si $u_{2k} < u_{1k} < u_{3k}$ alors $\overset{2}{\perp}_S(\mathbf{u}_k) = u_{1k}$.

3.2 Les mesures et l'indice proposés

Comme il a été rappelé à la section 2.2, les CVIs de la littérature sont généralement fondés sur une mesure de *compacité* C (à minimiser) ou sur une mesure de *séparation* S (à maximiser) des groupes. Si on considère seulement la première, la meilleure partition est celle où il y a autant de groupes que de points. Inversement, si seule la séparation est prise en compte, la partition optimale consiste en un seul groupe. C'est pourquoi, il est est admis depuis fort longtemps qu'il est préférable de les combiner. Il a été établi que les mesures de séparation qui ne prennent en compte que les distances entre centres conduisent à des difficultés d'interprétation de la structure, voir exemples dans [12]. Par ailleurs, bien que la capacité à détecter des chevauchements entre groupes est aujourd'hui considérée comme un critère important pour un CVI [5], la plupart des travaux sont fondés sur des représentations intuitives des chevauchements, sans aucune quantification de leur degré puisque ceux-ci sont implicitement donnés par U.

Ces raisons nous conduisent à proposer la définition, pour chaque point $\mathbf{x}_k \in X$, une mesure C_k évaluant le degré de chevauchement d'un nombre l de groupes et une mesure de séparation S_k quantifiant le degré de chevauchement du groupe le plus probable, c'est-à-dire celui correspondant au degré d'appartenance le plus élevé par rapport aux $(c-1)$ autres. De manière évidente, une valeur faible de S_k indique une grande séparation du groupe le plus probable par rapport aux autres. Par calculs successifs du fOU-l pour différentes valeurs de l, nous obtenons une combinaison d'ordre de degrés de chevauchement pour $\mathbf{x}_k$. Afin de déterminer le degré total de chevauchement C_k pour un point donné, il suffit d'apprécier quel ordre (≥ 2) implique le plus grand chevauchement, par exemple à l'aide d'une disjonction floue : $\forall \mathbf{u}_k \in U$, $c \geq 2$,

$$C_k(\mathbf{u}_k) = \overset{1}{\bigperp}_{l=2,c} \left(\overset{l}{\bigperp}_{i=1,c} u_{ik} \right) \tag{7}$$

Pour mesurer la séparation de chaque point $\mathbf{x}_k$, nous proposons d'utiliser le fOU-1, qui évalue le chevauchement d'un seul groupe, c'est à dire sa séparation vis-à-vis des autres, puisque $\mathbf{u}_k$ est normalisé. Cette agrégation, qui correspond à la disjonction floue des degrés d'appartenance, sélectionne donc le groupe le plus probable : $\forall \mathbf{u}_k \in U$, $c \geq 2$,

$$S_k(\mathbf{u}_k) = \overset{1}{\bigperp} \left(\underbrace{\overset{1}{\bigperp}_{i=1,c} u_{ik}, \cdots, \overset{1}{\bigperp}_{i=1,c} u_{ik}}_{c-1 \text{ fois}} \right) \tag{8}$$

Ainsi, étant donnée une t-norme $\top$, une faible valeur de C_k et une valeur élevée de S_k indiquent que $\mathbf{x}_k$ appartient à un groupe bien séparé et non chevauché par aucun des autres.

In fine, il faut combiner, puis agréger ces mesures pour tous les points pour obtenir un CVI. Ainsi, nous définissons la famille d'indices OSI (*Overlap and Separation Index*) d'ordre l à valeurs dans $[0, 1]$, comme la moyenne arithmétique des rapports des deux mesures : étant donnés une t-norme $\top$ et une matrice de partition floue U de X en $c \geq 2$ groupes,

$$OSI_{\perp}(U) = \frac{1}{n} \sum_{k=1}^{n} \frac{C_k(\mathbf{u}_k)}{S_k(\mathbf{u}_k)} \tag{9}$$

Cet indice, qu'il faut minimiser, possède des propriétés intéressantes.

[3] habituellement $u_{(l)k}$ représente la l-ème valeur de $\mathbf{u}_k$ dans l'ordre croissant mais l'ordre décroissant est plus adapté ici

Propriété 1. *Si on utilise les t-normes standard (min,max), $OSI_{\perp}(U)$ est la moyenne des rapports du $2^{ème}$ plus grand des degrés d'appartenance sur le plus grand.*

Démonstration. Avec $(\top = min, \perp = max)$, on a $\overset{l}{\perp}(\mathbf{u}_k) = u_{(l)k}$, $\forall \mathbf{u}_k$. Alors, d'après (7-8), on a $C_k(\mathbf{u}_k) = u_{(2)k}$ et $S_k(\mathbf{u}_k) = u_{(1)k}$ et par conséquent, $OSI_{\perp_S}(U) = \frac{1}{n}\sum_{k=1}^{n} u_{(2)k}/u_{(1)k}$. □

Propriété 2. *Si la partition est stricte, alors $OSI_{\perp}(U) = 0$, pour tout couple $(\top, \perp)$.*

Démonstration. Pour une partition stricte, on a $u_{ik} \in \{0, 1\}$, $\forall \mathbf{u}_k$, et on écrit que $u_{(1)k} = 1$ et $u_{(2)k} = ... = u_{(c)k} = 0$. Alors, d'après (7), on a : quel que soit $(\top, \perp)$

$$C_k(\mathbf{u}_k) = \left(\overset{2}{\perp}(1,0,...,0)\right) \overset{1}{\perp} \left(\overset{3}{\perp}(1,0,...,0)\right) \overset{1}{\perp} ... \overset{1}{\perp} \left(\overset{c}{\perp}\left(\frac{1}{c},...,\frac{1}{c}\right)\right)$$
$$= \underbrace{0 \overset{1}{\perp} ... \overset{1}{\perp} 0}_{c-1 \text{ fois}} = 0$$

car 0 est absorbant pour $\top$.
De même, d'après (8), on a $S_k(\mathbf{u}_k) = 1$ car 1 est absorbant pour $\perp$, et par conséquent $OSI_{\perp}(U) = 0$. □

Propriété 3. *Si la partition est complètement floue*[4]*, alors $OSI_{\perp}(U) = b_{\perp} \leq 1$.*

La démonstration doit être faite pour chaque couple $(\top, \perp)$ car la borne $b_{\perp}$ en dépend. Montrons, par exemple dans le cas des normes standard, que $b_{\perp_S} = 1$:

Démonstration. Avec $(\top = min, \perp = max)$, on a $\overset{l}{\perp}(\mathbf{u}_k) = u_{(l)k}$, $\forall \mathbf{u}_k$. Alors, d'après (7) :

$$C_k(\mathbf{u}_k) = max\left(\overset{2}{\perp}\left(\frac{1}{c},...,\frac{1}{c}\right), \overset{3}{\perp}\left(\frac{1}{c},...,\frac{1}{c}\right), ..., \overset{c}{\perp}\left(\frac{1}{c},...,\frac{1}{c}\right)\right)$$
$$= max\underbrace{\left(\frac{1}{c},...,\frac{1}{c}\right)}_{c-1 \text{ fois}} = \frac{1}{c}.$$

De même, $S_k(\mathbf{u}_k) = \frac{1}{c}$, et par conséquent $OSI_{\perp_S}(U) = 1$ □

[4] une partition est *complètement floue* si sa matrice de c-partition U est telle que $u_{ik} = \frac{1}{c}, \forall i = 1, c$ et $\forall k = 1, n$

4 Résultats de segmentation

Nous comparons notre méthode de segmentation sur la base de Berkeley [19] constituée de 300 images couleur pour lesquelles une segmentation humaine est donnée. L'objet de cet étude n'étant pas la sélection du meilleur espace colorimétrique, nous utilisons les coordonnées CIE-lab, voir [21]. De nombreuses méthodes non supervisées existent, plus ou moins sophistiquées (fusion d'espaces colorimétriques, métriques adaptées, pré– et post–traitement, etc [15, 21]) de sorte que nous limitons les résultats à des méthodes comparables, à savoir utilisant de l'information similaire à une simple partition des pixels dans l'espace colorimétrique : *Mean-shift* [6] et *N-cuts* [24]. Pour chacune d'elles, nous avons pris les valeurs de paramètres par défaut. Pour OSI, nous ne présentons ici que les résultats obtenus avec les t-normes standard $(\top = min, \perp = max)$ qui ont l'avantage d'agréger de manière prudente par rapport aux autres qui, nous l'avons constaté, peuvent sur– ou sous–segmenter les images. La performance est donnée par comparaison des images segmentées aux segmentations humaines. Nous utilisons les scores classiques suivants : le *Probabilistic Rand Index* (PRI) qui compte le nombre de pixels dont les étiquettes sont les mêmes pour les deux segmentations, le *Variation of Information* (VoI) qui moyenne l'entropie conditionnelle d'une segmentation étant donnée l'autre, et le *Global Consistency Error* (GCE) qui évalue à quel point une segmentation peut être vue comme le raffinement de l'autre. Comme ces mesures ont des intervalles de variation très différents, nous les normalisons par le score de la segmentation humaine de telle sorte que la méthode est meilleure si la valeur normalisée est grande (Tableau 1). Le score de la segmentation humaine est évidemment calculé avec les mêmes indices sur les images segmentées fournies dans la base, constituant de ce fait une vérité terrain (parmi d'autres). Les

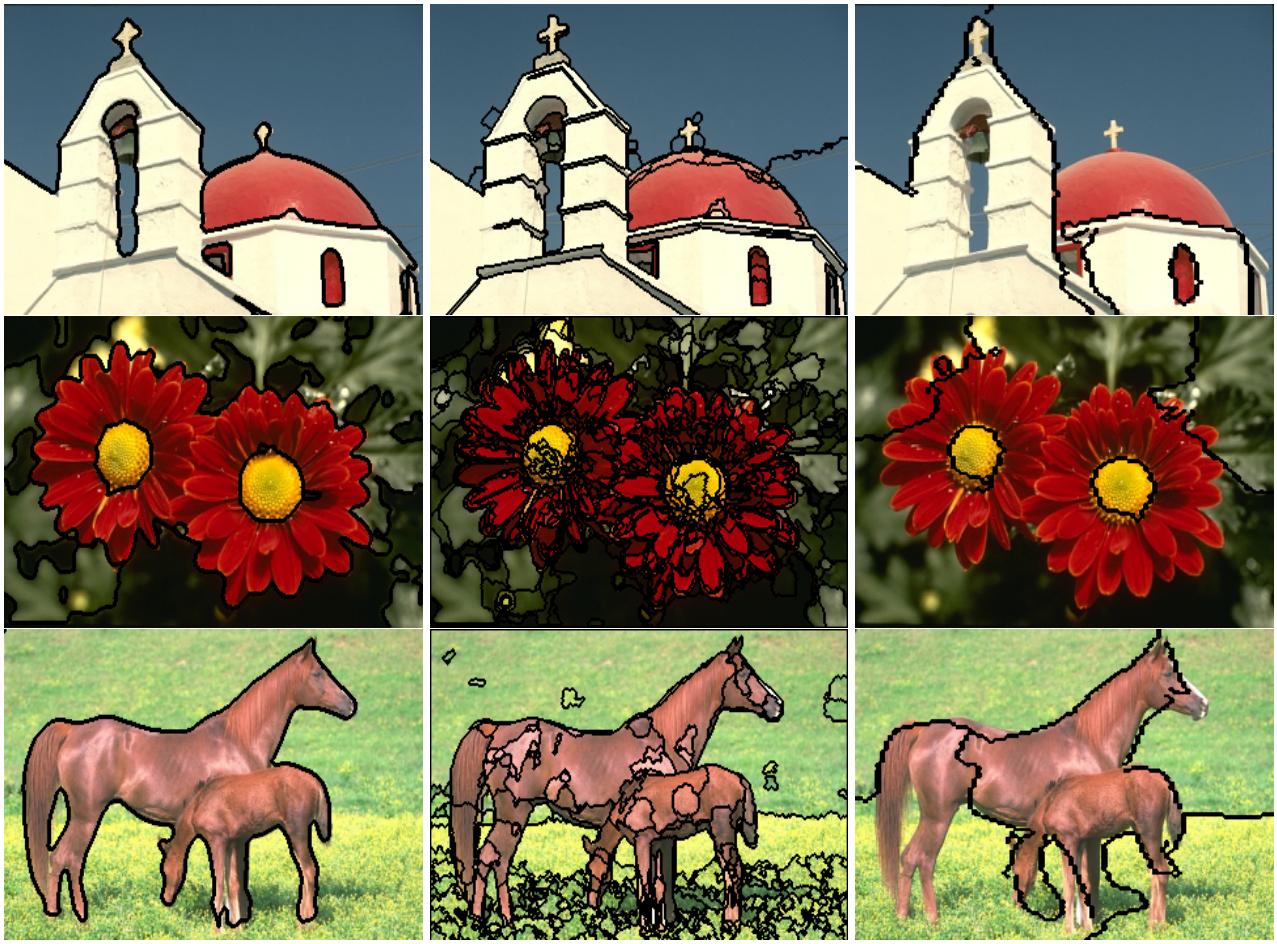

Figure 1 – Exemples d'images segmentées par (sFCM,$OSI_{\perp=max}$) (gauche), Mean-Shift (centre) et N-cuts (droite).

résultats montrent que l'approche proposée est satisfaisante par rapport aux méthodes comparables : pour deux des trois indices de performances, sFCM_{OSI} est supérieur à Mean-shift et N-cuts.

Tableau 1 – Performance des méthodes testées (la valeur en gras est le meilleur score).

Méthode	**PRI**	**VoI**	**GCE**
sFCM_{OSI}	82.86%	**48.36%**	**39.90%**
Mean-shift [6]	**86.29%**	44.57%	30.50%
Ncuts [24]	82.51%	37.65%	36.24%

Des images segmentées sont données à la Fig. 1 à des fins d'illustration (cette page et la suivante). Chaque frontière entre régions homogènes est représentée par une bordure noire. Comme pour les résultats obtenus sur la base entière (Tableau 1), on peut voir que la méthode proposée fournit des images segmentées de meilleure qualité. En particulier, on notera que les images ne sont pas sur-segmentées, et que les contours sont relativement bien localisés, à la différence de Mean-shift et N-cuts. Bien sûr, un paramétrage plus fin des autres méthodes aurait permis une segmentation meilleure que des images comme celle des fleurs ou de l'étoile de mer mais il aurait fallu un autre réglage pour ne pas détériorer la segmentation des images comme celle du paysage vallonné. Hormis le choix de la t-norme, la même pour toutes les images ici, la méthode proposée ne nécessite aucun paramétrage.

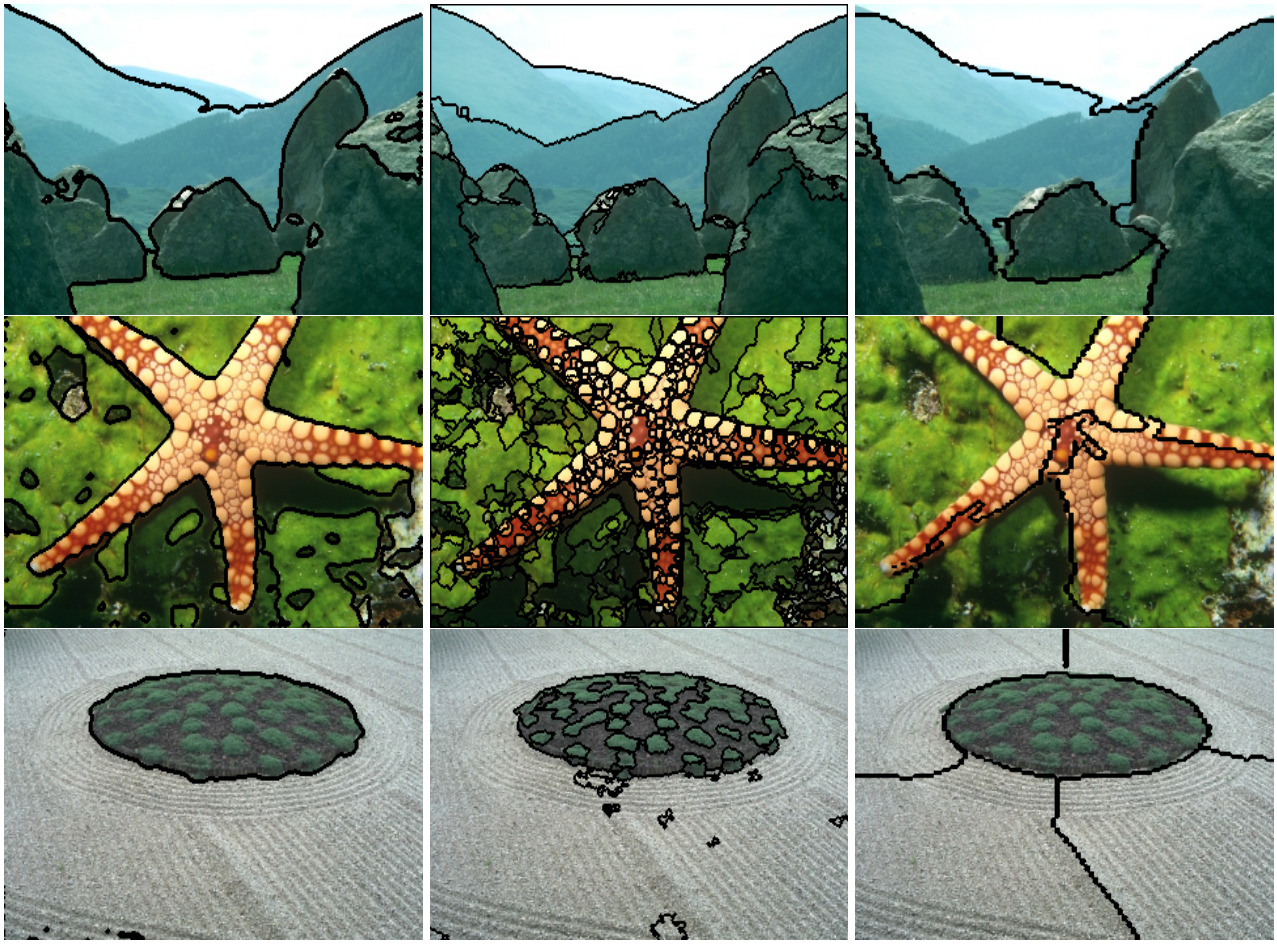

Figure 1 – (suite) Exemples d'images segmentées par (sFCM,$OSI_{\perp=max}$) (gauche), Mean-Shift (centre) et N-cuts (droite).

5 Conclusion

Nous proposons une méthode de segmentation non paramétrique fondée sur un indice de validité de partitions floues. Cet indice permet de sélectionner le nombre optimal de groupes de pixels homogènes en termes de plus grande séparation et moindre chevauchement flous et par suite de segmenter une image en régions homogènes. Afin d'améliorer les partitions floues en amont, il est associé à la version avec contraintes spatiales de l'algorithme FCM. Les résultats obtenus, comparés à des méthodes bien connues montrent la validité de l'approche.

Dans un futur proche, nous proposerons une étude de l'influence du choix des t-normes sur lesquelles l'indice est fondé. En effet, l'emploi de normes triangulaires autres que les normes standard ($\top = min$, $\perp = max$) peut conduire à une sur– ou sous–segmentation des images. Nous montrerons que ceci est lié aux propriétés mathématiques des normes triangulaires.

Remerciements :

Ce travail a reçu le soutien du Ministère de l'Enseignement Supérieur et de la Recherche sous la forme d'une Allocation de Recherche.

Références

[1] A. M. Bensaid, L. O. Hall, J. C. Bezdek, L. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh. Validity-guided (re)clustering with applications to image segmentation. *IEEE Transactions on Fuzzy Systems*, 4(2) :112–123, 1996.

[2] J. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer Academic, 1999.

[3] J. C. Bezdek. *Pattern Recognition with fuzzy objective function algorithm*. Plenum Press, 1981.

[4] J.C. Bezdek and N.R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics*, 23(3) :301–315, 1998.

[5] M. Bouguessa, S. Wang, and H. Sun. An objective approach to cluster validation. *Pattern Recognition Letters*, 27(13) :1419–1430, 2006.

[6] D. Comaniciu and P. Meer. Mean shift : A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5) :603–619, 2002.

[7] J. C. Dunn. *Fuzzy Automata and Decision processes*, chapter Indices of partition fuzziness and the detection of clusters in large data sets. Elsevier, NY, 1977.

[8] M. El-Melegy, E. Zanaty, W. Abd-Elhafiez, and A. Farag. On cluster validity indexes in fuzzy and hard clustering algorithms for image segmentation. In *IEEE Int. Conf. on Image Processing*, 2007.

[9] Y. Fukuyama and M. Sugeno. A new method for choosing the number of clusters for the fuzzy c-means method. In *Proc. 5th Fuzzy Systems Symposium*, pages 247–250, 1989.

[10] D.E. Gustafson and W.C. Kessel. Fuzzy clustering with fuzzy covariance matrix. In *Proc. IEEE Conference on Decision and Control*, pages 761–766, San Diego, California, 1979.

[11] S. Chen J. Wu K-S. Chuang, H-L. Tzeng and T-J. Chen. Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*, 30(1) :9–15, 2006.

[12] D-W. Kim, K.H. Lee, and D. Lee. On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition*, 37(10) :2009–2025, 2004.

[13] E.P Klement and R. Mesiar. *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*. Elsevier, 2005.

[14] S. H. Kwon. Cluster validity index for fuzzy clustering. *Electronic Letters*, 34(22) :2176–2177, 1998.

[15] P. Lambert and H. Grecu. A quick and coarse color image segmentation. In *IEEE Int. Conf. on Image Processing*, 2003.

[16] Y.W. Lim and S.U. Lee. On the color image segmentation algorithm based on the thresholding and fuzzy c-means techniques. *Pattern Recognition*, 23(9) :935–652, 1990.

[17] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1*, 1967.

[18] H. Maitre. *Le traitement d'images*. Lavoisier, 2003.

[19] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int. Conf. Computer Vision*, volume 2, pages 416–423, 2001.

[20] L. Mascarilla, M. Berthier, and C. Frélicot. A k-order fuzzy or operator for pattern classification with k-order ambiguity rejection. *Fuzzy Sets and Systems*, 159(15) :2011–2029, 2008.

[21] M. Mignotte. Segmentation by fusion of histogram-based k-means clusters in different color spaces. *IEEE Transactions on Image Processing*, 17(5) :780–787, 2008.

[22] N.R. Pal and J.C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*, 3(3) :370–379, 1995.

[23] M. Roubens. Pattern classification problems and fuzzy sets. *Fuzzy Sets and Systems*, 1(4) :239–253, 1978.

[24] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8) :888–905, 2000.

[25] X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8) :841–847, 1991.

E.3 A new fuzzy 3-rules pattern classifier with reject options based on aggregation of membership degrees

Cet article a été publié comme :

H. Le Capitaine & C. Frélicot. A new fuzzy 3-rules pattern classifier with reject options based on aggregation of membership degrees. In Proc. 12th Int. Conf. on Information Processing and Management of Uncertainty, pages 473-480. Spain, 2008.

A new fuzzy 3-rules pattern classifier with reject options based on aggregation of membership degrees

Hoel Le Capitaine **Carl Frélicot**
MIA Laboratory, University of La Rochelle, France
{hlecap01,cfrelico}@univ-lr.fr

Abstract

In this paper, we address the problem of fuzzy rule-based pattern recognition with reject options. These options are made possible thanks to simple rules whose satisfaction level is expressed by the value of dedicated operators that aggregate degrees of typicality. Results obtained with the proposed classifier on articicial and real data are given.
Keywords: Pattern recognition, reject options, fuzzy rules, aggregation.

1 Introduction

In many decision-making systems, we face the problem of aggregating collections of numerical or ordinal data to obtain a typical value. Aggregation operators are used to obtain an overall score for each alternative, which is exploited to establish a decision. In the context of pattern recognition, such a decision consists in assigning objects to a given class. Since the publication of L. A. Zadeh's paper on fuzzy sets, this theory has evolved into powerful tools for managing uncertainty in decision-making systems. Many research works have been carried out for applications to pattern recognition, e.g. fuzzy rule-based classifiers such as Takagi-Sugeno-Kang ones that approximate classification boundaries [9]. Fuzzy classifiers generally have faster training capabilities and comparable generalization abilities to other ones. This paper deals with the design of fuzzy rules whose inputs are the membership degrees of objects to the classes at hand instead of features describing them like in classical fuzzy if-then classifiers. By including possible rejection of extraneous and ambiguous objects, we allow to significantly improve the performance of such a classification or decision-making system.
The paper is organized as follows. In section 2, we recall the formal definition of aggregation operators and some special functions. In section 3, a brief overview of fuzzy classifier design and fuzzy rule-based classifiers is given. Section 4 briefly describes the principles of pattern rejection and the existing strategies leading to the different options: exclusive classification, ambiguity or distance rejection. Then, the new approach for obtaining classification boundaries using simple fuzzy rules is presented and discussed. Section 5 present results obtained on both artificial and real data sets. Concluding remarks and ideas for future work are finally given in section 6.

2 Aggregation Operators

The aggregation problem is of major importance in decision-making systems, where values to be aggregated are generally defined on a finite real interval or on ordinal scales. In this paper, we assume with no loss of generality that they come from the unit interval. If not, a simple transformation can be found to make this assumption true. Among the frequently used aggregation operators, in addition to the mean operators, one can cite: triangular norms [10], OWA (*Ordered Weighted Averaging*) operators [16], γ-operators [19], or fuzzy integrals [14]. These operators are divided in several categories, depending on the way the values are aggregated: conjunctives, disjunctives, compensatory, compen-

sative, and weighted operators. An aggregation operator $\mathcal{A}$ on the unit interval is said to be conjunctive if $\mathcal{A}(x_1, \cdots, x_n) \leq min(x_1, \cdots, x_n)$. The values are combined by a fuzzy logical *AND*, which means that the overall score is high if and only if all partial scores are high. If we add properties of non decreasingness, commutativity and associativity, we obtain the family of the triangular norms (t-norms). A triangular norm is a commutative, associative and monotone function $T : [0,1]^2 \to [0,1]$ having for neutral element 1, i.e. $T(x,1) = x$ for $x \in [0,1]$. The minimum operator is a t-norm and $T(x,y) \leq x \wedge y$, with $\wedge = min$, so that the minimum is the greatest t-norm. An aggregation operator on the unit interval is said to be disjunctive if $\mathcal{A}(x_1, \cdots, x_n) \geq max(x_1, \cdots, x_n)$. The values are combined by a fuzzy logical *OR*, which means that the overall score is low if and only if all partial scores are low. If we add properties of non decreasingness, commutativity and associativity, we obtain the family of the triangular conorms (t-conorms). A triangular conorm is a commutative, associative and monotone function $S : [0,1]^2 \to [0,1]$ having for neutral element 0, i.e. $S(x,0) = x$ for $x \in [0,1]$. The maximum operator is a t-conorm and $S(x,y) \geq x \vee y$, with $\vee = max$, so that the maximum is the weakest t-conorm. Table 1 shows the three basic t-norms ($\top$) and t-conorms ($\perp$), but there exists a large panel of triangular norms (indeed infinite since the combination of two t-norms is a t-norm), see [8] for a survey.

Table 1: Basic t-norms and t-conorms couples

Standard	$a \top_S b = \min(a,b)$
	$a \perp_S b = \max(a,b)$
Algebraic	$a \top_A b = a\,b$
	$a \perp_A b = a + b - a\,b$
Lukasiewicz	$a \top_L b = \max(a+b-1, 0)$
	$a \perp_L b = \min(a+b, 1)$

An aggregation operator on the unit interval is said to be compensatory if $min(x_1, \cdots, x_n) \leq \mathcal{A}(x_1, \cdots, x_n) \leq max(x_1, \cdots, x_n)$. Here, a high value (resp. low) can be compensated by a low value (resp. high). If we add properties of non decreasingness and idempotency, we obtain the family of mean operators. Conjunctives, disjunctives and compensatory operators form a large part of aggregation operators on $[0,1]$, but one can find some operators that do not belong to any of these categories, e.g. the symmetric sum [13] and compensative operators [19]. Both compensative and compensatory operators tend to express a compromise within the values, but the output of the former operators do not necessarily lie between the minimum and the maximum values whereas the latter do.

To conclude this section, let us mention the existence of weighted operators. Multi-criteria decision often need to establish the importance of each evaluated criterion, which implies an extension of the usual non weighted operators. These weights cause the loss of neutrality from the decision system, but could allow to perform better results. OWA operators are used to adjust the terms AND and OR, and allow an easier semantic interpretation of the linguistic quantifiers. Fuzzy integrals compute the mean value of a given function with respect to a fuzzy measure and thus can be seen as aggregation operators in the discrete case.

3 Fuzzy Pattern Classification

3.1 Classifier Design

Let $x = {}^t(x_1 \;\; x_2 \;\; ... \;\; x_p)$ be a pattern in a feature space, to be classified with respect to a set $\Omega = \{\omega_1, \cdots, \omega_c\}$ of c classes. A conventional hard classifier is a rule aiming at assigning an unknown pattern x to a particular class ω_i, thanks to the aggregation of class-labels $u_j(x)$, e.g. posterior probabilities that x belongs to the classes or membership degrees to fuzzy sets associated with the classes. We do not address the labelling problem in this paper, so we will use a measure of typicality:

$$u_i(x) = \frac{\alpha}{\alpha + d^2(x, p_i)} \qquad (1)$$

where α is a user-defined parameter, d a distance, and p_i a prototype of ω_i obtained from a learning set of patterns. The most popular agregation operator is the standard t-conorm, defining the so-called *max classifier* (MC):

1. compute class-labels $u_j(x)$ $(j = 1, c)$
2. aggregate labels $\mathcal{A}(u_1, \cdots, u_c)$
3. rule: if $\mathcal{A}$ is u_i then assign x to ω_i

Such an exclusive classification rule is not so efficient because it supposes that classes do not significantly overlap (*separability*) and that Ω is exhaustively defined (*closed-world*). These assumptions are generally not valid in practice.

3.2 Fuzzy Rule-Based Classifiers

Fuzzy systems are meant to be models understandable for the end-user. They use if-then rules and a mechanism which should correspond to the expert knowledge for a given problem. A fuzzy if-then classifier consists in a model with fuzzy rules of the form:

$$\text{if } A_i^1 \text{ AND } A_i^2 \text{ AND } \cdots \text{ AND } A_i^n \text{ then } B_i \quad (2)$$

where A_i^k is a fuzzy set with membership function $a_i^k : \mathbb{R} \to [0,1]$, $i = 1, \cdots, m$, $k = 1, \cdots, p$ and $B_i \in \mathbb{R}$. Among the most popular models, the Takagi-Sugeno-Kang (*TSK*) model [15] is characterized by:

1. a set of m fuzzy rules.
2. a connective operator $\mathcal{A}$ whose output provides the satisfaction τ_i, or firing strength, of the rule i:
$$\tau_i(x) = \mathcal{A}(a_i^1(x_1), \cdots, a_i^p(x_p)) \quad (3)$$
3. a defuzzification method allowing the final assignment.

For instance, by choosing the product for the connective operator and the COA (*Center Of Area*) defuzzification method, we obtain a TSK2 [9] classifier for the object x:

$$C(x) = \frac{\sum_{i=1}^{m} B_i \prod_{k=1}^{p} a_i^k(x_k)}{\sum_{i=1}^{m} \prod_{k=1}^{p} a_i^k(x_k)} \quad (4)$$

Let us consider a simple two-classes problem as first example. We generated 100 two-dimensional samples equally arising from two normal distributions $\omega_1 \sim \mathcal{N}\left(p_1 = [1,3]^T, I\right)$ and $\omega_2 \sim \mathcal{N}\left(p_2 = [5,1]^T, I\right)$ where I is the identity matrix. Two rules are enough to define a TSK2 classifier for this problem:

- rule R1:
  if x_1 is about 1 AND x_2 is *about 3*, then b_1
- rule R2:
  if x_1 is about 5 AND x_2 is *about 1*, then b_2

where *about c* can be modelled by the following membership function:

$$a^k(x_k) = \exp\left(-(x_k - c)^2/2\right) \quad (5)$$

Thus, the satisfaction τ_i ($i = 1, 2$) of the rules are:

$$\tau_i(x) = \exp\left(-(x - p_i)^T (x - p_i)/2\right)$$

If we set b_1 to 1 and b_2 to 0, we obtain:

$$C(x) = \frac{\tau_1}{\tau_1 + \tau_2} \quad (6)$$

and the decision areas shown in Fig. 1 that vary from black, corresponding to ω_2 ($b_2 = 0$), to white corresponding to ω_1 ($b_1 = 1$). The classification boundary is given by $\tau_1 = \tau_2$, whose solution is the following hyperplan equation:

$$x_2 = \frac{3}{2} x_1 - 7 \quad (7)$$

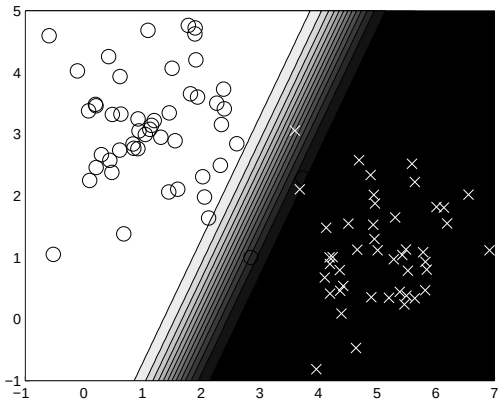

Figure 1: A simple two-classes $(\circ, \times)$ problem with normal distribution in $\mathbb{R}^2$

Many membership functions from $\mathbb{R}$ to $[0,1]$ can be used, then too much flexibility arises, leading to untractable learning. So we restrict the study to positive definite functions. Since classes may have ellipsoidal shapes, Abe and Thawonmas introduced a fuzzy classifier, based on neural networks and fuzzy rules, which provides ellipsoidal decision areas [1]. They use the following membership function:

$$a_i(x) = \exp(-h_i^2(x))$$
$$h_i^2(x) = \frac{d_i^2(x)}{\alpha_i}$$
$$d_i^2(x) = (x - p_i)^T \Sigma_i^{-1} (x - p_i)$$

where Σ_i is the covariance matrix of the class ω_i and α_i is a user-defined parameter.
Furthermore, it has been proved [9] that some TSK classifiers with adequate membership functions are equivalent to non-parametrical classifiers:

- TSK3 with $a_i(x) = \exp\Big(-\frac{1}{2}(x-p_i)^2\Big)$ is equivalent to the nearest neighbor classifier,
- TSK4 with $a_i(x) = \exp\Big(-\frac{1}{2h^2}(x-p_i)^2\Big)$ is equivalent to a Parzen classifier.

In [7], the authors proved that a monotonic function $f(x)$ can model the classification boundary for two classes in $\mathbb{R}^2$, and they propose to use other t-norms and t-conorms than minimum and maximum because they fail in $p > 2$-dimensional problems, even for linear separable problems.

4 The New Classifier

4.1 Reject Options

We said in subsection 3.1 that the separability and the closed-world assumptions are generally not valid in practice. Reject options have been proposed to overcome these difficulties and to reduce the misclassification risk. The first one, called distance reject option, is dedicated to outlying patterns. If x is far from all the class-prototypes, the option allows to assign it to no class. The second one allows to assign inlying patterns to several or all the classes. If x is close to two or more prototypes, it is associated with the corresponding classes. Finally, a pattern is exclusively classified when its maximum membership degree is significantly higher than the others. Different strategies can be adopted to handle the options at hand [4, 5] but they all lead to a three types decision system: exclusive classification, ambiguity rejection, distance rejection.

4.2 Reject Options Using Fuzzy Rules

In this section, we detail the construction of reject operators with the help of fuzzy rules. Step by step, by combining triangular norms, we define what the suitable operator should satisfy. In [18], Yager and Rybalov showed that t-norms and t-conorms provide downward and upward reinforcement, respectively. Downward (respectively upward) reinforcement corresponds to the fact that if the values are all low (respectively high), then they reinforce each other and the overall value will be low (respectively high). From these properties, they build a full reinforcement operator by the use of a fuzzy system modeling technique. We follow this approach and define new rules. According to the rules defined, the operator will have a conjunctive, disjunctive or a compensative behavior.

- rule $R1$: there is only one high value, (so exclusive classification is possible)
- rule $R2$: several values are high, (so ambiguity rejection is needed)
- rule $R3$: all the values are low (so distance rejection is needed)

We propose to formalize this set of rules.

4.3 The Fuzzy 3-Rules based Classifier with Reject Options (3-RCRO)

Let L, respectively H, be the fuzzy subset defined on $[0,1]$ corresponding to the concept *low*, respectively *high*. Furthermore, let (T,S) be any dual couple (t-norm,t-conorm). As mentioned in the second section, the t-conorm S is the fuzzy equivalent of the logical operator OR. We define $H(u_i) = u_i$ using a linear membership function and $L(u_i) = 1 - u_i$ as the negation of H. In [6], the authors built a fuzzy exclusive OR operator that extends the crisp XOR operator to the fuzzy context:

$$\underset{i=1,c}{\overline{\perp}} u_i = \left(\overset{1}{\underset{i=1,c}{\perp}} u_i\right) \top \left(\overline{\overset{2}{\underset{i=1,c}{\perp}} u_i \Big/ \overset{1}{\underset{i=1,c}{\perp}} u_i}\right) \tag{8}$$

$$\text{where} \qquad \overset{k}{\underset{i=1,c}{\perp}} u_i = \underset{A\in\mathcal{P}_{k-1}}{\top}\left(\underset{j\in C\setminus A}{\perp} u_j\right) \tag{9}$$

with $\mathcal{P}$ the powerset of $C = \{1, 2, ..., c\}$ and $\mathcal{P}_k = \{A \in \mathcal{P} : |A| = k\}$ where $|A|$ denotes the cardinality of subset A. Assusming u to be a sorted c-tuple, i.e. $u_1 \geq u_2 \cdots \geq u_c$, we defined in [12] an operator based on triangular norms and the Sugeno integral which quantifies the similar-

ity of the block of values $\{u_j, \ldots, u_k\}$:

$$\Phi_{j,k}(u) = \begin{cases} \dfrac{\displaystyle\top_{i=\frac{k+j}{2}}^{k} u_i \top \mathcal{K}_\lambda(i,k)}{\displaystyle\top_{i=\frac{k+j}{2}}^{j} u_i \top \mathcal{K}_\lambda(i,j)} & \text{if } k-j \text{ is even} \\ \dfrac{\displaystyle\top_{i=\frac{k+j+1}{2}}^{k} u_i \top \mathcal{K}_\lambda(i,k)}{\displaystyle\top_{i=\frac{k+j-1}{2}}^{j} u_i \top \mathcal{K}_\lambda(i,j)} & \text{if } k-j \text{ is odd} \end{cases} \tag{10}$$

where $\mathcal{N}_\lambda(i,l)$ is a gaussian kernel defined by:

$$\mathcal{N}_\lambda(i,l) = \exp \frac{-(i-l)^2}{\lambda} \tag{11}$$

The new fuzzy 3-rules classifier with reject options we propose derive as follows:

- rule $R1$: if u_1 is high XOR u_2 is high XOR $\cdots$ XOR u_c is high, then $x \mapsto \omega_{\text{argmax}_j(u_j)}$
- rule $R2$: if $\Phi_{1,2}$ is high OR $\cdots$ OR $\Phi_{1,c}$ is high, then reject x for ambiguity
- rule $R3$: if u_1 is low AND u_2 is low AND $\cdots$ AND u_c is low, then reject x for distance.

We define the firing strengths of the 3 rules by:

$$\tau_1(x) = \underset{i=1,c}{\perp} H(u_i) = \underset{i=1,c}{\perp} u_i \tag{12}$$

$$\tau_2(x) = \underset{i=2,c}{\perp} H(\Phi_{1,i}(u)) = \underset{i=2,c}{\perp} \Phi_{1,i}(u) \tag{13}$$

$$\tau_3(x) = \underset{i=1,c}{\top} L(u_i) = \underset{i=1,c}{\top} \overline{u_i} \tag{14}$$

Finally, a simple *winner takes all* strategy applied to the triplet $\{\tau_1(x), \tau_2(x), \tau_3(x)\}$ activates the corresponding rule which gives the classification result. It is worthnoting that the proposed classifier does not involve any threshold, compared to most of other classifiers with reject options. Depending on his interest, the user can select only the reject rules, and choose a conjunction, for instance a triangular norm, for the aggregation process of these firing strengths. In this case, a threshold can be applied on $\underset{i=1,3}{\top} \tau_i(x)$. By changing the characteristics, various discrimination procedures are possible, see section 5 for examples.

Examples of label vectors $u(x)$ for a $c = 3$ classes problem and resulting firing strengths are given in Table 2 for different dual couples: standard (S), algebraic (A), and the parameterized Hamacher family (H_γ) defined by:

$$\begin{aligned} \top_{H,\gamma}(a,b) &= \frac{a\,b}{\gamma + (1-\gamma)\,(a+b-a\,b)} \\ \perp_{H,\gamma}(a,b) &= \frac{a+b+(\gamma-2)\,a\,b}{1-(1-\gamma)\,a\,b} \end{aligned}$$

which reducts to $(\top_A, \perp_A)$ when $\gamma = 1$.

The given label vectors are representative of various situations: ambiguity between three and two classes, exclusive classification, and distance rejection, respectively. Whatever the dual couple, the Winning Firing Strength (*WFS*) gives the correct classification result. However, $\tau_2(x)$ does not exhibit selective ambiguity rejection, e.g. ambiguity between three or two classes in the Table. This refinment, corresponding to the k-order ambiguity concept, can obviously be obtained by indexing $\tau_2(x)$ by k instead of c in (13).

Table 2: Examples of firing strengths

$u(x)$	$\tau_i(x)$ and $(\top, \perp)$			
$\begin{pmatrix}0.85\\0.90\\0.75\end{pmatrix}$		$\tau_1(x)$	$\tau_2(x)$	$\tau_3(x)$
	$S:$	0.05	0.94	0.10
	$A:$	0.07	0.99	0.00
	$H_0:$	0.16	0.95	0.05
$\begin{pmatrix}0.85\\0.90\\0.10\end{pmatrix}$		$\tau_1(x)$	$\tau_2(x)$	$\tau_3(x)$
	$S:$	0.05	0.94	0.1
	$A:$	0.21	0.96	0.01
	$H_0:$	0.20	0.94	0.06
$\begin{pmatrix}0.15\\0.90\\0.10\end{pmatrix}$		$\tau_1(x)$	$\tau_2(x)$	$\tau_3(x)$
	$S:$	0.83	0.16	0.10
	$A:$	0.72	0.30	0.07
	$H_0:$	0.70	0.32	0.09
$\begin{pmatrix}0.15\\0.00\\0.10\end{pmatrix}$		$\tau_1(x)$	$\tau_2(x)$	$\tau_3(x)$
	$S:$	0.15	0.66	0.85
	$A:$	0.19	0.40	0.80
	$H_0:$	0.17	0.45	0.81

5 Experimental Results

5.1 Artificial Data

This example aims at showing the classfication boundaries that can result from the triplet of fir-

ing strengths $\{\tau_1(x), \tau_2(x), \tau_3(x)\}$ using a particular dual couple: $(\top_A, \perp_A)$, but others give similar results. The data set consists of four classes composed of sixty points in $\mathbb{R}^2$ each. Classes slightly overlap in such a way that k-order ambiguity areas can appear. Membership degrees in the feature space are computed by equation (1) with $\alpha = 3$ and

$$d^2(x, p_i) = (x - p_i)^T \Sigma_i^{-1} (x - p_i)$$

where the mean vector p_i and covariance matrix Σ_i of each class are estimated from the data set.
First, we choose to define the 3-RCRO with no distance reject option. Thus, it simply consists in ambiguity rejecting patterns for which $\top_{i=1,3} \tau_i(x) = \tau_1(x)\,\tau_2(x)\,\tau_3(x)$ is higher than a threshold t, else exclusively assigning them to the class of maximum degree. Fig. 2 shows the contour plot of $\top_{i=1,3} \tau_i(x)$ varying from high values (white) to low ones (black).
The second result we give is obtained by allowing the three rules to be activated by the *WFS* classifier. As mentioned in the previous section, no threshold is needed. Classification boundaries are shown in Fig. 3, where white, grey ands black areas correspond to distance rejection, ambiguity rejection and exclusive classification respectively. Such a classification procedure results in a a very low error rate (nearly zero) and a high reject rate, therefore is more suitable for applications where the cost of misclassification is very high. Remind that other combinations of part of all the firing strengths leading to a dedicated 3-RCRO could allow the user to tune the different rates, but a threshold should be used.

5.2 Real Data Sets

In this final subsection, we present some results that show the classification performance of the 3-RCRO on well-known real data. The iris data set [3] contains $n = 150$ observations from $c = 3$ four-dimensional classes (iris species) of 50 points each. It is one of the most used benchmarks in pattern recognition, especially for cluster validity because two classes, numbered 2 and 3, present a substantial overlap in the feature space the class 1 is well separated from the others. For visualization purpose, only the third and fourth

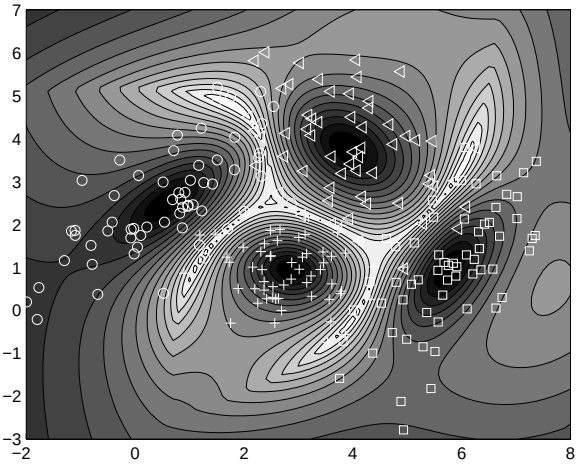

Figure 2: Contour plot of $\tau_1(x) \top \tau_2(x) \top \tau_3(x)$ with algebraic norms $(\top_A, \perp_A)$ – artificial data

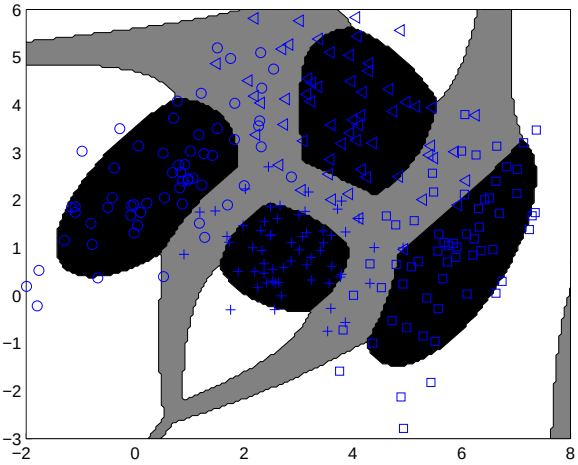

Figure 3: Classification areas using the winning firing strength – artificial data.

features were considered, as many authors do [2]. The membership degrees were computed using the same equations than for the artificial data. Contour plot of $\top_{i=1,3} \tau_i(x)$ using $(\top_A, \perp_A)$ is shown in Fig. 4 varying from high values (white) to low ones (black). We obtained similar results with other dual couples. Table 3 shows the error rates obtained with classical classifiers using a resubstitution procedure: Quadratic Bayes (*QB*), Nearest Neighbor (*NN*) and the Max Classifier (*MC*, see section 3.1).

The performance of the *WFS* classifier and the 3-RCRO for different values of the threshold on $\top_{i=1,3} \tau_i(x)$ using $(\top_A, \perp_A)$ are reported in Table

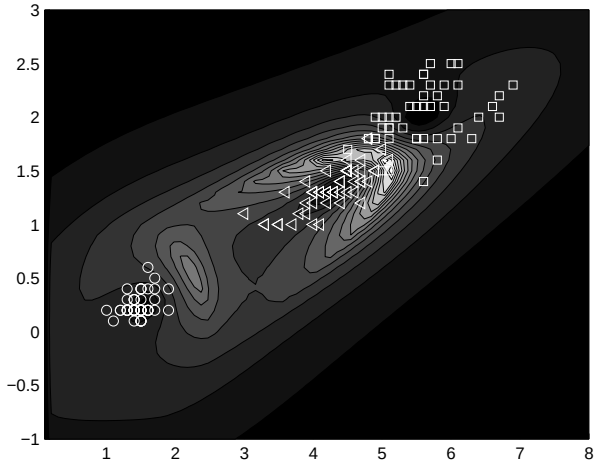

Figure 4: Contour plot of $\tau_1(x) \top \tau_2(x) \top \tau_3(x)$ with algebraic norms $(\top_A, \perp_A)$ – iris data

Table 3: Error rates without reject options – iris data

%	*QB*	*NN*	*MC*
$error$	2.00	4.67	2.67

4. Depending on the different strategies the user has in mind, the threshold can be set in order to minimize the error rate (e.g. $t = 0.08$), maximize the correct classification rate with a low error rate (e.g. $t = 0.15$), or maximize the correct classification rate with no reject options (e.g. $t = 1$) – this setting making the 3-RCRO coincide with the *MC* classifier as one could expect. Obviously, the reject rate decreases as t increases and its tuning is a keypoint as for every classifier involving a threshold. Note that rejected patterns were all rejected for ambiguity with $t = 0.15$ and not surprisingly some patterns were distance rejected with $t = 0.08$.

Table 4: Performance rates – iris data

%	*WFS*	$t = 0.08$	$t = 0.15$	$t = 1$
error	0.00	0.00	1.33	2.67
reject	25.33	12.00	2.00	0.00
correct	74.67	88.00	96.67	97.33

The second data set is the Forest Cover Type obtained from the UCI repository [3]. This is a very large GIS data set representing the forest cover type of a region, which contains $n = 581,012$ observations and $c = 7$ classes described by 54 attributes. Following [11], we only consider the $p = 10$ numeric valued attributes and the 495,141 points, belonging to classes 1 and 2. These two classes are equally distributed and have a significant overlap in the feature space. Error rates with usual supervised classifiers are shown in Table 5 and performance results of the proposed method are reported in Table 6. Compared to classical classifiers such as QB and NN without reject options, the proposed method reduces the error rate.

Table 5: Error rates without reject options – forest data

%	*QB*	*NN*	*MC*
$error$	25.56	29.83	24.91

Table 6: Performance rates – forest data

%	*WFS*	$t = 0.08$	$t = 0.15$	$t = 1$
error	9.58	3.18	19.14	24.91
reject	34.13	75.99	13.91	0.00
correct	56.29	20.82	66.94	75.09

6 Conclusion

In this paper, we propose a new approach to the classication problem including reject options. It is based on three fuzzy rules whose inputs are the membership degrees of objects to the classes at hand instead of features describing them and aggregation operators. These operators, based on combination of triangular norms and the Sugeno integral, are especially designed to give one of the three possible results. According to the firing strength of each rule, the pattern is either classified in a single class, or ambiguity rejected between several classes or distance rejected from all the classes. Two solutions have been proposed for the different decision boundaries. The first one consists in taking the most satisfied rule (*WFS*) and the second one consists in thresholding the conjonction of the three firing strengths (*3-RCRO*) or only part of them. Other combinations are under investigation. Experimental results show that the proposed approach is able to detect patterns that must be rejected and therefore gives satisfactory decison boundaries. We also

proposed to extend the approach to k-order ambiguity rejection by indexing the second rule in a proper way.
Among the future works we have in mind, let us cite the use of compensative and compensatory operators via combination of different couples for pattern classifiers with reject options, the use of uninorms instead of triangular norms couples as universal approximation [17] of fuzzy systems in this context.

References

[1] S. Abe and R. Thawonmas. A fuzzy classifier with ellipsoidal regions. *IEEE Transactions on Fuzzy Systems*, 5(3):358–368, 1997.

[2] J. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer Academic, 1999.

[3] C. Blake and C. Merz. Uci repository of machine learning databases, 1998. `http://www.ics.uci.edu/mlearn/MLRepository.html`.

[4] B. Dubuisson and M. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26(11):155–165, 1993.

[5] C. Frélicot and L. Mascarilla. A third way to design pattern classifiers with reject options. In *21th International Conference of the North American Fuzzy Information Processing Society*, 2002.

[6] C. Frélicot, L. Mascarilla, and M. Berthier. A new cluster validity index for fuzzy clustering based on combination of dual triples. In *IEEE International Conference on Fuzzy Systems*, 2006.

[7] F. Klawonn and E. P. Klement. Mathematical analysis of fuzzy classifiers. *Lecture Notes on Computer Sciences*, 1280:359–370, 1997.

[8] E.P Klement and R. Mesiar. *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*. Elsevier, 2005.

[9] L. I. Kuncheva. How good are fuzzy if-then classifiers? *IEEE Transactions on Systems, Man and Cybernetics*, 30(4):501–509, 2000.

[10] K. Menger. Statistical metrics. *Proc. National Academy of Science USA*, 28:535–537, 1942.

[11] P. Mitra, C. A. Murthy, and S.K. Pal. A probabilistic active support vector learning algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):413–418, 2004.

[12] H. Le Capitaine, T. Batard, C. Frélicot, and M. Berthier. Blockwise similarity in [0,1] via triangular norms and sugeno integrals - application to cluster validity. In *IEEE International Conference on Fuzzy Systems*, pages 835–840, London, UK, 2007.

[13] W. Silvert. Symmetric summation: a class of operations on fuzzy sets. *IEEE Transactions on Systems, Man and Cybernetics*, 9(10):659–667, 1979.

[14] M. Sugeno. *Theory of fuzzy integrals and its applications*. PhD thesis, Tokyo Institute of Technology, 1974.

[15] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics*, 15(1):116–132, 1985.

[16] R. R. Yager. Ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 18(1):183–190, 1988.

[17] R. R. Yager and V. Kreinovich. Universal approximation theorem for uninorm-based fuzzy systems modeling. *Fuzzy Sets and Systems*, 140(2):331–339, 2003.

[18] R. R. Yager and A. Rybalov. Full reinforcement operators in aggregation techniques. *IEEE Transactions on Systems, Man and Cybernetics*, 28(6):757–769, 1998.

[19] H.-J. Zimmermann and P. Zysno. Latent connectives in human decision making. *Fuzzy Sets and Systems*, 4(1):37–51, 1980.

E.4 A family of cluster validity index based on a l-order fuzzy OR operator

Cet article a été publié comme :

H. Le Capitaine & C. Frélicot. A family of cluster validity indexes based on a l-order fuzzy OR operator. Joint IAPR Int. Workshops on Structural and Syntactic Pattern Recognition and Statistical Techniques in Pattern Recognition, SSPR&SPR 2008, LNCS 5342, pp. 622-631, Orlando, Florida, 2008.

A family of cluster validity indexes based on a l-order fuzzy OR operator

Hoel Le Capitaine and Carl Frélicot

MIA Laboratory, University of La Rochelle, La Rochelle 17042 Cedex, France,
hoel.le_capitaine@univ-lr.fr , carl.frelicot@univ-lr.fr

Abstract. Clustering is one of the most important task in pattern recognition. For most of partitional clustering algorithms, a partition that represents as much as possible the structure of the data is generated. In this paper, we adress the problem of finding the optimal number of clusters from data. This can be done by introducing an index which evaluates the validity of the generated fuzzy c-partition. We propose to use a criterion based on the fuzzy combination of membership values which quantifies the l-order overlap and the intercluster separation of a given pattern.

1 Introduction

The objective of fuzzy clustering is to partition the data set into c distinct clusters. The fuzzy c-means (FCM) algorithm proposed by Bezdek [3] and its variations [8] are probably the most commonly used fuzzy clustering methods. However, these algorithms require the user to set the number c of clusters although the user do not always know it. A fuzzy c-partition obtained by FCM has to be validated because its quality depends on this number. Many cluster validity indexes have been proposed that evaluate each fuzzy c-partition and determines the optimal number $c^\star$ of clusters allowing to obtain the optimal partition of the data.

Compactness and separation of clusters are often considered to validate a partition [4, 12, 7]. In this paper, we propose a new family of indexes that combine an overlap measure and a separation measure both based on aggregation of membership values. Section 2 provides background information on basic fuzzy operators and the l-order fuzzy OR operator our work is based on. Next, in section 3, we briefly describe the FCM algorithm and recall some well-known cluster validity indexes we will use for comparison. The measures we propose to use and the definition of the new index are described in section 4. Experimental results on both synthetic and real data sets that show its efficiency and concluding remarks are given in section 5 and 6 respectively.

2 Mathematical Background

For the applications we have in mind, e.g. cluster validity, we are particulary interested in aggregation functions that map a collection $\mathbf{u}$ of c values in $[0, 1]$

(i.e. a vector $\mathbf{u} = [u_1, \cdots, u_c]^T$) to a value in $[0, 1]$. Among the frequently used aggregation operators, one can cite the classes of triangular norms (t-norms) and triangular conorms (t-conorms). They have been introduced to characterize the general multivalued logic *AND* and *OR* operations and are widely used in fuzzy logic and fuzzy set theory to implement conjunctive and disjunctive operator respectively. A triangular norm is a commutative, associative and monotone function $\top$ having for neutral element 1. Alternatively, its dual operator, the triangular conorm, is a commutative, associative and monotone function $\perp$ having for neutral element 0. Examples of triangular norms couples are given in Table 1 for two operands, see [10] for a large survey. Note that various parameterized families have been introduced, e.g. the Hamacher family defined by $u_1 \top_H u_2 = \frac{u_1\, u_2}{\gamma + (1-\gamma)\,(u_1 + u_2 - u_1\, u_2)}$ and $u_1 \perp_H u_2 = \frac{u_1 + u_2 - u_1\, u_2 - (1-\gamma)\, u_1\, u_2}{1-(1-\gamma)\, u_1\, u_2}$ where $\gamma \in [0, +\infty[$. The dual couple is generally associated with a fuzzy negation defined as $N(u_1) = 1 - u_1$ and mentionned as the triple $(\top, \perp, N)$.

Table 1. Basic t-norms and t-conorms couples

Standard	$u_1 \top_S u_2 = \min(u_1, u_2)$
	$u_1 \perp_S u_2 = \max(u_1, u_2)$
Algebraic	$u_1 \top_A u_2 = u_1\, u_2$
	$u_1 \perp_A u_2 = u_1 + u_2 - u_1\, u_2$
Łukasiewicz	$u_1 \top_L u_2 = \max(u_1 + u_2 - 1, 0)$
	$u_1 \perp_L u_2 = \min(u_1 + u_2, 1)$

Assume that the values u_i $(i = 1, \cdots, c)$ to be aggregated represent the degree to which an object $\mathbf{x}$ satisfies each group description, i.e. its similarity to the prototypes describing each group. Using this knowledge contained in $\mathbf{u}$, clustering consists in selecting the most appropriate group that the objects will be assigned to. The maximum operator is commonly used in this situation, but we may be interested in the lower values, which interact with the greatest value. In particular, if an object satisfies more than one group description, such an exclusive partitioning is not efficient. A fundamental issue becomes the determination of the overall degree of exclusive belongingness to a group or cluster. In [11], the authors define the l-order fuzzy OR operator. This operator evaluates degrees of satisfaction at a given order by combination of triangular norms. Let $\mathcal{P}$ be the powerset of $C = \{1, 2, ..., c\}$ and $\mathcal{P}_l = \{A \in \mathcal{P} : |A| = l\}$ where $|A|$ denotes the cardinality of subset A, then the fOR-l is defined by:

$$\bigperp^{l}_{i=1,\cdots,c} u_i = \bigtop_{A \in \mathcal{P}_{l-1}} \left(\bigperp_{j \in C \setminus A} u_j \right) \tag{1}$$

It must be viewed as some kind of generalization of the notion of "l^{th} highest" value, with l in C. In particular, with standard triangular norms, $\bigperp^{l}(\mathbf{u})$ is ex-

actly the "l^{th} highest" element of $\mathbf{u}$. For instance, let us take $C = \{1, 2, 3\}$, $l = 2$ and use standard triangular norms. We have $\mathcal{P}_1 = \{\{1\}, \{2\}, \{3\}\}$ and

$$\bigsqcup_{i=1,\cdots,3}^{2} u_i = \min(\max(u_2, u_3) \max(u_1, u_3) \max(u_1, u_2))$$

so that if $u_2 < u_1 < u_3$, $\bigsqcup_{i=1,\cdots,3}^{2} u_i = u_1$.

This operator satisfies nice mathematical properties such as monotony, symmetry, see [11] for proofs and details.

3 Cluster Validity for Fuzzy Clustering

3.1 Fuzzy *c*-means algorithm

Clustering is an instance of unsupervised classification which aims at finding a structure of groups in set of n p-dimensional patterns $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. In this framework, the label vectors $\mathbf{u}_k = u(\mathbf{x}_k)$ do not exist and clustering algorithms can be used to obtain them from X. For instance, the *fuzzy c-means* (FCM) algorithm partitions X into $c > 1$ clusters by minimizing the following objective function [3]:

$$J_m(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \, ||\mathbf{x}_k - \mathbf{v}_i||^2 \tag{2}$$

where u_{ik} is the membership degree of $\mathbf{x}_k$ to the i^{th} cluster represented by its centroid $\mathbf{v}_i \in \mathbb{R}^p$. Centroids are gathered into a $(c \times p)$ matrix $V = [\mathbf{v}_1, ..., \mathbf{v}_c]$. Degrees u_{ik} are subject to $\sum_{i=1}^{c} u_{ik} = 1$ for all $\mathbf{x}_k$ in X and to $0 < \sum_{k=1}^{n} u_{ik} < n$ $(\forall i = 1, \cdots, c)$. In addition, they are elements of the fuzzy $c-$partition matrix U $(c \times n)$. The so-called *fuzzifier* $m > 1$ is a weighting exponent which makes the resulting partition more or less fuzzy [12]. The higher m is, the softer the clusters' boundaries are. Minimization of (2) is obtained by iteratively updating (U, V) as follows:

$$u_{ik} = 1 \Big/ \sum_{j=1}^{c} \left(\frac{||\mathbf{x}_k - \mathbf{v}_i||}{||\mathbf{x}_k - \mathbf{v}_j||} \right)^{2/(m-1)} \tag{3}$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m \, \mathbf{x}_k}{\sum_{k=1}^{n} u_{ik}^m} \tag{4}$$

The usual euclidian norm $||.||$ induces hyperspherical clusters, hence FCM can only detect clusters with the same shape and orientation.

3.2 Classical indexes

Validating the provided clustering of X consists in assessing whether the resulting partition reflects the data structure or not. Since c is a user-defined parameter of clustering algorithms such as FCM, most of works on cluster validity focus on the number of clusters problem. Many validity indexes have been proposed for fuzzy clustering (refer to [4, 9, 14, 13, 16] for comparative studies). They can be classified in two main categories. The first one is composed of indexes that only use membership degrees (U). Let us cite the *Partition Coefficient* [3], taking values in $[\frac{1}{c}, 1]$:

$$PC(c) = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2 \tag{5}$$

or the *Partition Entropy* [2], taking values in $[0, \log(c)]$:

$$PE(c) = -\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}\, log(u_{ik}) \tag{6}$$

Both PC to be maximized and PE to be minimized are monotonic with c, as well as their bounds. Normalized versions have been proposed to reduce this monotonic tendency, *e.g.* in [6]. We will use these normalized versions in the experiments: $NPC(c) = \frac{cPC(c)-1}{c-1}$ and $NPE(c) = \frac{nPE(c)}{n-c}$. The second category consists of indexes that use membership degrees but also some information about the geometrical structure of the data (U, V, X), *e.g.* the Xie-Beni index [12, 15]:

$$XB(c) = \frac{J_m(U,V)\,/n}{\min_{i,j=1,\cdots,c;j\neq i}\ ||\mathbf{v}_i - \mathbf{v}_j||^2} \tag{7}$$

or the Fukuyama-Sugeno index [7]:

$$FS(c) = J_m(U,V) - \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m\, ||\mathbf{v}_i - \overline{\mathbf{v}}||^2 \tag{8}$$

where $\overline{\mathbf{v}}$ is the mean of centroids. Both XB and FS combine the FCM objective function (2) which measures the degree of compactness of the clusters and an additional term which measures the degree of their separation. Combination indicates that both indexes are to be minimized. The more compact and separated the clusters are, the more optimal c is.

4 The Proposed Index

Let $U = [\mathbf{u}_1, ..., \mathbf{u}_c]$ be a fuzzy c-partition provided by a fuzzy clustering algorithm, e.g. FCM. An overlap measure between l fuzzy clusters for each point $\mathbf{x}_k$ in X described by its membership degrees can be obtained by Eq. (1). By successively computing $\overset{l}{\perp}(\mathbf{u}_k)$ for different values of c, we get a combination

of l-order overlap measure for $\mathbf{x}_k$. The most satisfied order is obtained by the disjunction of these measures, and we define the overall overlap measure as:

$$O(\mathbf{u}_k, c) = \mathop{\perp}_{l=2,c}^{1} \left(\mathop{\perp}_{i=1,\cdots,c}^{l} u_{ik} \right) \tag{9}$$

In [4], Bezdek and Pal show that intercluster separation plays a more important role in cluster validity than cluster diameters. We propose to introduce such a measure by quantifying the separation of each point $\mathbf{x}_k$ with $\mathop{\perp}^{1}(\mathbf{u}_k)$ which is the overlap measure between 1 fuzzy cluster, i.e. its separation from the other fuzzy clusters, since $\mathbf{u}_k$ sum up to one. Note that if standard triangular norms ar used, $\mathop{\perp}^{1}(\mathbf{u}_k)$ is the maximum coordinate of $\mathbf{u}_k$. Finally, we define the family of l-order Fuzzy OR Indexes as:

$$lFORI(c, \top, \perp) = \frac{1}{n} \sum_{k=1}^{n} \frac{O(\mathbf{u}_k, c)}{\mathop{\perp}_{i=1,\cdots,c}^{1} u_{ik}} \tag{10}$$

Given U, the less overlaping and separated clusters are, the lower the value of $lFORI$ is expected to be and minimizing Eq. (10) will give the optimal number $c^\star$ of clusters.
If U is hard, i.e. $u_{ik} \in \{0, 1\}$, then one value equals 1 while the others are 0, say $u_{1k} = 1$ and $u_{2k} = \cdots = u_{ck} = 0$. Since 0 is the absorbing element of $\top$, it is easy to verify that $O(\mathbf{u}_k, c) = 0$ for all $\mathbf{u}_k$, whatever $(\top, \perp)$, therefore $lFORI(c, \top, \perp) = 0$:

$$\mathop{\perp}_{l=2,c}^{1} \left(\mathop{\perp}_{i=1,\cdots,c}^{l} u_{ik} \right) = \left(\mathop{\perp}_{i=1,\cdots,c}^{2} (1, 0, \cdots, 0) \right) \mathop{\perp}^{1} \cdots \mathop{\perp}^{1} \left(\mathop{\perp}_{i=1,\cdots,c}^{c} (1, 0, \cdots, 0) \right) \tag{11}$$

$$= \mathop{\perp}^{1} (\underbrace{0, \cdots, 0}_{c-1 \text{ times}}) = 0. \tag{12}$$

and $\mathop{\perp}^{1}(1, 0, \cdots, 0) = 1$, since 1 is the absorbing element of $\perp$.

On the other hand, if U is totally fuzzy, i.e. $u_{ik} = \frac{1}{c}$ $(\forall i = 1, \cdots, c)$, the resulting $lFORI$ value depends on the couple $(\top, \perp)$ because Eq. (9) only reduces to

$$\mathop{\perp}_{l=2,c}^{1} \left(\mathop{\perp}_{i=1,\cdots,c}^{l} u_{ik} \right) = \left(\mathop{\perp}_{i=1,\cdots,c}^{2} \left(\frac{1}{c}, \cdots, \frac{1}{c} \right) \right) \mathop{\perp}^{1} \cdots \mathop{\perp}^{1} \left(\mathop{\perp}_{i=1,\cdots,c}^{c} \left(\frac{1}{c}, \cdots, \frac{1}{c} \right) \right) \tag{13}$$

in the general case. If standard t-norms are used, we have:

$$\overset{1}{\underset{l=2,c}{\perp}}\left(\overset{l}{\underset{i=1,\cdots,c}{\perp}} u_{ik}\right) = \overset{1}{\perp}\left(\underbrace{\frac{1}{c},\cdots,\frac{1}{c}}_{c-1 \text{ times}}\right) \tag{14}$$

$$= \frac{1}{c} \tag{15}$$

which is the value of $\overset{1}{\perp}(\mathbf{u}_k)$ for all $\mathbf{u}_k$, therefore $lFORI(c, \top_S, \perp_S) = 1$ for all c. Unfortunately, it is not possible to give a simple value for $lFORI$ for other couples $(\top, \perp)$, but an upper bound can be found. Due to lack of space, the proof is postponed to a forthcoming long paper, as well as properties that could help the user to choose the couple $(\top, \perp)$.

5 Experiments

5.1 Behavior according to clusters' separability

First, we generated a serie of 10 data sets, each composed of 800 points drawn from a mixture of $c = 4$ bivariate normal distributions. The covariance matrix of each component is the same $\Sigma_i = I$ $(\forall i = 1, \cdots, c)$ and the mean vectors are: $\mu_1 = \alpha \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mu_2 = \alpha \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\mu_3 = \alpha \begin{pmatrix} -1 \\ -1 \end{pmatrix}$ and $\mu_4 = \alpha \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, for increasing values of $\alpha = 1, 2, \ldots, 10$. This successively moves the clusters in opposite directions, creating less overlap as the clusters become more and more separated. The first and last data sets are shown in Figure 1-*left*. Each data set was then clustered using FCM with $c = 4$, providing a fuzzy partition matrix U_α. Corresponding values of $lFORI$ for the different basic norms are plotted in Figure 1-*right* as a function of α. As expected, the proposed validity index decreases towards 0 as α increases whatever the couple $(\top, \perp)$.

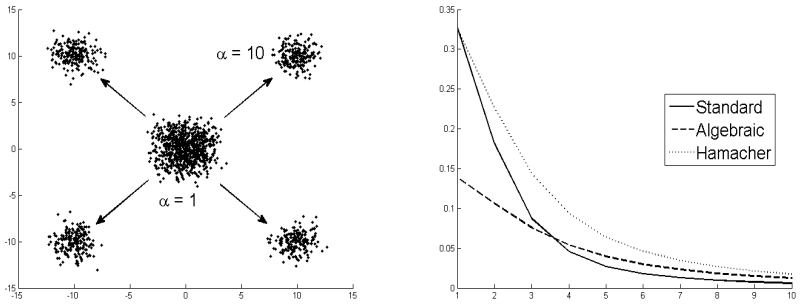

Fig. 1. α-separated data sets – $\alpha = 1$ and 10 (*left*) and values of $lFORI(c = 4)$ as a function of α for various t-norms (*right*).

5.2 Robustness to outliers

In order to compare the proposed index to the classical ones, we generated a data set D_1 containing $n = 200$ points consisting of 50 points each drawn from a mixture of $c = 4$ bivariate normal distributions, see Figure 2-*left*. FCM was used with $m = 2$ for c varying from $c_{min} = 2$ to $c_{max} = 10$. A second artificial data set D_2 was generated. It is similar to D_1 except that 100 points drawn from a uniform distribution were added, as shown in Figure 2-*right*. These additional points act as noise and can make the FCM algorithm partitioning the data set into three clusters because the less separated groups in the right-lower area tend to become only one cluster. Values of the tested validity indexes on D_1 and D_2 are given in Table 2 and Table 3 respectively, where optimal values are bold faced and acceptable ones are italicized. Even if the classical indexes are known to be efficient, most of them fail in giving the right number of clusters on D_1 and with even stronger reason in presence of noise (D_2), whereas $lFORI$ always gives the right number $c^\star = 4$ whatever the couple $(\top, \perp)$ for both data sets.

Moreover, multiple runs of FCM with random initializations on data set D_2

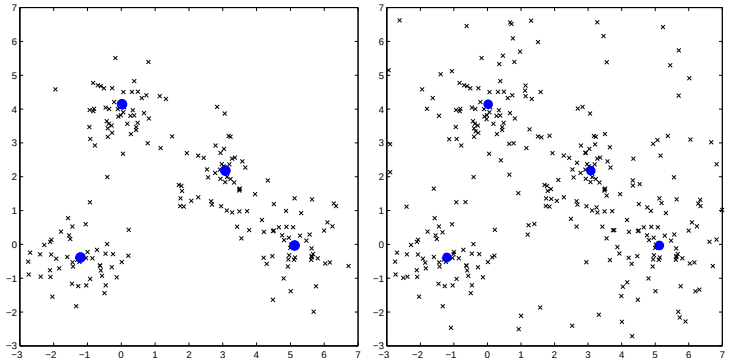

Fig. 2. Centroids of optimal clusters for data sets D_1 (*left*) and D_2 (*right*).

showed us that the proposed index gives more stable results compared to the others, thus demonstrating a significantly higher robustness to noisy data.

5.3 Sensitivity to the fuzzifier m

The FCM objective function depends on the fuzzy exponent m, see Eq. (2). Since the resulting fuzzy c-partition U is sensitive to this parameter, a cluster validity index can also be analysed with respect to m. In [4], Pal and Bezdek have shown that FCM provides best results for m lying in $[1.5, 2.5]$. We compare the different indexes in this range on data sets D_1 and D_2. Again, c is varying from $c_{min} = 2$ to $c_{max} = 10$. The selected number of clusters are reported in Table 4. As it can be seen, the proposed index $lFORI$ is less sensitive to m whatever the couple $(\top, \perp)$ for both data sets.

Table 2. Validity indexes on data set D_1.

c	NPC	NPE	XB	FS $\times 10^{-3}$	$lFORI$		
					$(\top,\perp)_S$	$(\top,\perp)_A$	$(\top,\perp)_H$
2	0.57	**0.52**	0.13	-1.57	0.17	0.14	0.15
3	0.70	0.56	**0.07**	-3.39	0.12	0.12	0.14
4	**0.74**	0.59	*0.08*	-0.74	**0.07**	**0.07**	**0.12**
5	0.67	0.80	0.31	**-4.25**	0.12	0.11	0.17
6	0.61	1.00	0.45	-2.98	0.18	0.13	0.18
7	0.56	1.15	0.37	-1.53	0.19	0.10	0.20
8	0.530	1.27	0.30	-1.60	0.23	0.09	0.19
9	0.532	1.37	0.31	-2.43	0.29	0.10	0.19
10	0.50	1.47	0.28	-0.15	0.28	0.11	0.17

Table 3. Validity indexes on data set D_2 .

c	NPC	NPE	XB	FS $\times 10^{-3}$	$lFORI$		
					$(\top,\perp)_S$	$(\top,\perp)_A$	$(\top,\perp)_H$
2	0.49	**0.58**	0.23	-0.85	0.25	0.17	0.19
3	**0.60**	0.71	**0.11**	-2.33	0.23	0.11	0.20
4	*0.59*	0.87	0.13	-0.75	**0.17**	**0.09**	**0.15**
5	0.55	1.08	0.24	**-5.21**	0.18	0.13	0.21
6	0.50	1.26	0.35	-0.06	0.21	0.14	0.22
7	0.50	1.36	0.36	-0.12	0.21	0.17	0.22
8	0.46	1.50	0.44	-2.41	0.26	0.16	0.21
9	0.45	1.61	0.46	-2.61	0.30	0.16	0.19
10	0.44	1.70	0.36	-1.81	0.30	0.15	0.18

5.4 Benchmark data sets

We finally compare the different indexes on benchmark data sets:

- *Iris* [5], composed of three classes of 50 flowers each described by 4 physical attibutes. Two classes have a substantial overlap in the feature space and the optimal number of clusters to be found is debatable: 2 or 3, see [4].
- *Wine* [5], which consists of 13 chemical attributes for $n = 178$ italian wines, divided into three classes. Classes are well separable, so the indexes found in the literature generally give the right number of clusters.
- *Wisconsin Breast Cancer* [5], composed of $n = 699$ malignant/benign cells described by 9 features computed from digitized images.
- The artificial data set $X30$ introduced in [4], consisting in $n = 30$ observations in $\mathbb{R}^2$, for which three clusters are expected.
- The bidimensional artificial data set *Bensaid* [1] characterized by 3 classes of very different cardinalities (6, 3 and 40).
- The original *Starfield* [15], which contains the position and light intensity of $n = 51$ bright stars near Solaris. The expected number of clusters is 8 or 9, depending on the papers.

Table 4. Selected number of clusters in data sets D_1 and D_2 for different values of m.

data set	m	NPC	NPE	XB	FS	$lFORI$		
						$(\top,\perp)_S$	$(\top,\perp)_A$	$(\top,\perp)_H$
D_1	1.5	4	4	3	5	4	4	4
	1.7	4	4	4	4	4	4	4
	1.9	4	2	3	4	4	4	4
	2.1	4	2	3	5	4	4	4
	2.3	4	2	3	5	4	4	4
	2.5	3	2	4	6	4	3	4
D_2	1.5	5	4	3	5	4	4	4
	1.7	4	4	3	3	4	4	4
	1.9	4	2	3	4	4	4	4
	2.1	3	2	3	5	4	4	4
	2.3	4	2	3	3	4	3	4
	2.5	4	2	4	4	4	3	4

Table 5 summarizes the results obtained on these data sets. The $c^\star$ column gives the expected number of clusters and the other columns show the optimal number of clusters obtained using the validity indexes. The proposed index always finds the optimal number of clusters whatever the couple $(\top,\perp)$ while some others do not.

Table 5. Selected number of clusters in benchmark data sets.

Data set	$c^\star$	NPC	NPE	XB	FS	$lFORI$		
						$(\top,\perp)_S$	$(\top,\perp)_A$	$(\top,\perp)_H$
Iris	2 or 3	2	2	2	3	2	2	2
Wine	3	3	2	3	5	3	3	3
Breast	2	2	2	2	3	2	2	2
X30	3	3	2	3	7	3	3	3
Bensaid	3	3	2	3	7	3	3	3
Starfield	8 or 9	2	2	6	7	8	8	8

6 Conclusion

A new family of cluster validity indexes for fuzzy partitions has been proposed. These indexes combine a new measure of overlap of clusters and a separation measure. The novelty of the approach is that, for each data point to be clustered, the relative importance of each membership degree and the relationship of the degrees are taken into account through a combination of triangular norms $(\top,\perp)$. Results obtained on artificial and benchmark data sets have shown that the proposed family of indexes is most of time more efficient than well-known

ones, less sensitive to the fuzzifier exponent and particularly robust in noisy environments.
Further results about properties of the proposed indexes based on the mathematical properties of the agregation operators involved as well as guidelines to choose a member $(\top, \perp)$ of the family appropriated to specific situations will come soon.

References

1. A. M. Bensaid, L. O. Hall, J. C. Bezdek, L. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh. Validity-guided (re)clustering with applications to image segmentation. *IEEE Transactions on Fuzzy Systems*, 4(2):112–123, 1996.
2. J. C. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3:58–72, 1974.
3. J. C. Bezdek. *Pattern Recognition with fuzzy objective function algorithm.* Plenum Press, 1981.
4. J.C. Bezdek and N.R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics*, 23(3):301–315, 1998.
5. C. Blake and C. Merz. Uci repository of machine learning databases, 1998. http ://www.ics.uci.edu/mlearn/MLRepository.html.
6. R. N. Dave. Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters*, 17(6):613–623, 1996.
7. Y. Fukuyama and M. Sugeno. A new method for choosing the number of clusters for the fuzzy c-means method. In *Proc. 5th Fuzzy Systems Symposium*, pages 247–250, 1989.
8. D.E. Gustafson and W.C. Kessel. Fuzzy clustering with fuzzy covariance matrix. In *Proc. IEEE Conference on Decision and Control*, pages 761–766, San Diego, California, 1979.
9. D-W. Kim, K.H. Lee, and D. Lee. On cluster validity index for estimating the optimal number of fuzzy clusters. *Pattern Recognition*, 37(3):2009–2025, 2004.
10. E.P Klement and R. Mesiar. *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms.* Elsevier, 2005.
11. L. Mascarilla, M. Berthier, and C. Frélicot. A k-order fuzzy or operator for pattern classification with k-order ambiguity rejection. *Fuzzy Sets and Systems*, 159(15):2011–2029, 2008.
12. N.R. Pal and J.C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*, 3(3):370–379, 1995.
13. W. Wang and Y. Zhang. On fuzzy cluster validity indices. *Fuzzy Sets and Systems*, 158(19):2095–2117, 2007.
14. K.L. Wu and M.S. Yang. A cluster validity index for fuzzy clustering. *Pattern Recognition Letters*, 26(9):1275–1291, 2005.
15. X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991.
16. M. H. F. Zarandi, E.Neshat, and I. B. Türksen. A new cluster validity index for fuzzy clustering based on similarity measure. In *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, 11th International Conference*, pages 127–135, 2007.

E.5 A class-selective rejection scheme based on blockwise similarity of typicality degrees

Cet article a été publié comme :

H. Le Capitaine & C. Frélicot. A Class-Selective Rejection Scheme based on Blockwise Similarity of Typicality Degrees. In Proc. 19th Int. Conf. on Pattern Recognition, ICPR 08. Tampa, Florida, 2008.

A Class-Selective Rejection Scheme based on Blockwise Similarity of Typicality Degrees

Hoel Le Capitaine and Carl Frélicot
MIA Laboratory, University of La Rochelle, France
{hlecap01},{cfrelico}@univ-lr.fr

Abstract

Overlapping classes and outliers can significantly decrease a classifier performance. We adress here the problem of giving a classifier the ability to reject some patterns either for ambiguity or for distance in order to improve its performance. Given a set of typicality degrees for a pattern to be classified, we use an operator based on triangular norms and a discrete Sugeno integral to quantify their blockwise similarities. We propose a new class-selective rejection scheme which uses this operator outputs. We present the resulting algorithm which allows to assign a pattern to zero, one or several classes, and show its efficiency on real data sets.

1. Introduction

The problem of aggregating collections of numerical or ordinal data to obtain a typical value is present in many decision systems. Aggregation operators are used to obtain an overall value for each alternative, which is exploited to establish a final decision. In the context of supervised pattern classification, such a decision consists in assigning objects (or patterns) to one class based on the aggregation of degrees related to the given classes (posterior probabilities, membership values, ...). It has been proved that the misclassification risk significantly be reduced by allowing a classifier to reject extraneous and/or ambiguous patterns [2, 3, 5]. Thus, a classifier with reject options allows to assign a pattern to zero (distance rejection), one (exclusive classification) or several (ambiguity rejection) classes. Among the possible approaches, the fuzzy one has received more attention in the last few decades because of its ability to manage imprecise and/or incomplete data [4]. In this framework, we propose a new classification scheme with reject options, based on an operator which aggregates class-degrees of typicality of a pattern to be classified.

2. Fuzzy Aggregation Operators

Let us recall basic definitions of fuzzy operators that will be used to combine the values of interest, i.e. the pattern class-degrees of typicality. Depending on properties, aggregation functions can be classified into several categories: conjunctive, disjunctive, compensatory, and so on. We restrict on conjunctive and disjunctive functions. By definition, the output of a conjunctive operator is lower or equal than the minimum value, whereas the output of a disjunctive operator is greater or equal than the maximum value. Beyond these operators, we choose to use the triangular norms because of their ability to generalize the logical AND and OR crisp operators to fuzzy sets, see [7] for a survey. Briefly, a triangular norm (or t-norm) is a binary operation on the unit interval $\top : [0,1]^2 \to [0,1]$ which is commutative, associative, non decreasing and has 1 for neutral element. Thus, a t-norm $\top$ is conjunctive and the minimum operator $\wedge$ is the greatest t-norm. Alternatively, a trianglar conorm (or t-conorm) is the dual binary operation $\perp : [0,1]^2 \to [0,1]$ having the same properties except the latter: its neutral element is 0. Thus, a t-conorm $\perp$ is disjunctive and the maximum operator $\vee$ is the lowest t-conorm. Typical examples of dual couples (t-norm, t-conorm) that will be used in the sequel are given in Table 1.

Table 1. Typical triangular norm couples

Standard	$a \top_S b = \min(a,b)$
	$a \perp_S b = \max(a,b)$
Algebraic	$a \top_A b = a\,b$
	$a \perp_A b = a + b - a\,b$
Hamacher	$a \top_H b = \frac{ab}{\gamma+(1-\gamma)(a+b-ab)}$
	$a \perp_H b = \frac{a+b+(\gamma-2)ab}{1+(\gamma-1)ab}$

We will use another fuzzy aggregation operator, the Sugeno integral in its discrete form. It computes the mean value of a function with respect to a fuzzy mea-

sure m, which is a non-additive measure of uncertainty, i.e. more general than a possibility one and therefore a probability one. The integral of a function μ is defined by

$$\mathcal{S}_m = \bigvee_{i=1}^{n} \mu(x_i) \wedge m(A_{(i)}) \tag{1}$$

where $A_{(i)} = \{x_{(i)}, \cdots, x_{(n)}\}$ with respect to a permutation so that $\mu(x_{(i)}) \leq \cdots \leq \mu(x_{(n)})$. This integral is widely used in decision making, and in particular for pattern recognition [4] because of its ability to model some kind of interaction between features describing a pattern x.

3. The Class-selective Rejection Scheme

3.1. Classifier design

Let $\Omega = \{\omega_1, \cdots, \omega_c\}$ be a set of c classes and x an unknown pattern described by p features. Classifier design aims at defining rules that can associate $x \in \mathbb{R}^p$ with one class of Ω. It generally consists of two steps L (*labeling*) and H (*hardening*):

- $L : x \mapsto \mu(x) = {}^t(\mu_1(x), \ldots, \mu_c(x)) \in \mathcal{L}_{\bullet c}$, depending on the mathematical framework the classifier relies on, e.g. $\mathcal{L}_{pc} = [0,1]^c$ for degrees of typicality or $\mathcal{L}_{fc} = \{\mu(x) \in \mathcal{L}_{pc} | \sum_{i=1}^{c} \mu_i(x) = 1\}$ for posterior probabilities and membership degrees.

There exists many ways to compute labels, but we do not address the labelling problem in this paper and we will use the typicality measure defined as:

$$\mu_i(x) = \frac{\alpha}{\alpha + d^2(x, v_i)} \tag{2}$$

where α is a user-defined parameter, d a distance, and v_i a prototype of the class ω_i obtained from a learning set of patterns. It has been shown through empirical studies [9] that (2) is a good model for membership functions that model vague concepts or classes.

- $H : \mu(x) \mapsto h(x) = {}^t(h_1(x), \ldots, h_c(x)) \in \mathcal{L}_{hc}$, where $\mathcal{L}_{hc} = \{h(x) \in \mathcal{L}_{fc} | h_i(x) \in \{0,1\}\}$.

We adress the hardening problem because this step, which often reduces to the class of maximum label selection, is in charge of the decision making.

3.2. Reject options and the proposed class-selective scheme

As defined, H is an exclusive rule which is not efficient in practice because it supposed that: *i*) Ω is exhaustively defined (closed-world assumption), *ii*) classes do not ovelap (separability assumption).

Such untrue assumptions can lead to very undesired decisions. In many real applications, it is more convenient to with-hold making a decision than making a wrong assignment, e.g. in medical diagnosis where a false negative outcome can be much more costly than a false positive. Reject options have been proposed to overcome these difficulties and to reduce misclassification risk. The first one, called *distance rejection* [3], is dedicated to outlying patterns. If x is far from all the class prototypes, this option allows to assign it to no class. The second one, called *ambiguity rejection*, allows to assign inlying patterns to several or all classes [2, 5]. If x is close to two or more class prototypes, it is associated with the corresponding classes. Formally, including reject options consists in modifying H such that $h(x)$ can take values in the set of vertices of the unit hypercube $\mathcal{L}_{hc}^c = \{0,1\}^c$ instead of the exclusive subset $\mathcal{L}_{hc} \subset \mathcal{L}_{hc}^c$. Different strategies can be adopted to handle these options at hand, but they all lead to a three types decision system: distance rejection when $h(x) = {}^t(0, \ldots, 0) = \underline{0}$, exclusive classification when $h(x) \in \mathcal{L}_{hc}$, ambiguity rejection when $h(x) \in \mathcal{L}_{hc}^c \backslash \{\mathcal{L}_{hc} \cup \underline{0}\}$.

For any pattern x to be classified, given its label vector $\mu(x)$ from L by (2), sorted in descending order $\mu_1(x) \geq \mu_2(x) \geq \cdots \geq \mu_c(x)$, we propose a two-steps class-selective scheme for H as follows :

1) test for distance rejection : $h(x) = \underline{0}$ if $\mu_1(x) < s$, where s is a user-defined threshold

2) if x is not distance rejected, assign it to a (sub)set of selected classes of cardinality $k \in \{1, \ldots, c\}$; thus it is exclusively classified if $k = 1$ or ambiguity rejected between the selected classes if $k > 1$.

For the (sub)set of classes selection problem, we propose to use the $\Phi_{j,k}$ measure introduced in [8] for another purpose. Assuming μ to be a sorted c-tuple, $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_c$, we defined an operator based on triangular norms and the Sugeno integral which quantifies the similarity of the block of values $\{\mu_j, \cdots, \mu_k\}$:

$$\Phi_{j,k}(\mu) = \begin{cases} \dfrac{\bigvee_{i=\frac{k+j}{2}}^{k} \mu_i \top \mathcal{N}_\lambda(i,k)}{\bigvee_{i=\frac{k+j}{2}}^{j} \mu_i \top \mathcal{N}_\lambda(i,j)} & \text{if } k-j \text{ is even} \\ \dfrac{\bigvee_{i=\frac{k+j+1}{2}}^{k} \mu_i \top \mathcal{N}_\lambda(i,k)}{\bigvee_{i=\frac{k+j-1}{2}}^{j} \mu_i \top \mathcal{N}_\lambda(i,j)} & \text{if } k-j \text{ is odd} \end{cases} \tag{3}$$

where $\mathcal{N}_\lambda(i,l)$ is a gaussian kernel defined by:

$$\mathcal{N}_\lambda(i,l) = \exp \frac{-(i-l)^2}{\lambda} \quad (4)$$

The resolution parameter λ controls the area of influence: when $\lambda \to 0$, the kernel becomes a dirac centered in l, and when $\lambda \to \infty$, the kernel becomes the constant value 1. Therefore, the contribution of the intermediate values μ_{j+1}, ..., μ_{k-1} to $\Phi_{j,k}(\mu)$ is small if λ is close to zero and increases with λ. This means that increasing λ will not make two consecutive μ_i's more similar but may increase the similarity of blocks of larger size.

Since a high value of $\Phi_{1,k}(\mu(x))$ reveals that the k highest labels have similar values, then x can be associated with the corresponding classes. We propose to use an iterative scheme in order to find k, leading to the following second part for the H-step:

2) for i varying from 1 to c, set $h_i(x) = 1$ when $\Phi_{1,i}(\mu(x)) \geq t$, where t is a user-defined threshold

Note that $\Phi_{1,1}(\mu(x))$ is always greater than t, $\forall t \in [0,1]$, because $\Phi_{1,1}(\mu(x)) = \frac{\mu_1(x)}{\mu_1(x)} = 1$ for any t-norm couple. This ensure that at least one class is selected, the one which corresponds to the maximum of typicality degree, i.e. the one selected by the optimum classification rule in the sense of Chow [2]. In particular, if t is set to 1, there is no ambiguity rejection. The class-selective rejection scheme presented in Algorithm 1 can be compared to the rule proposed by Ha [5].

Algorithm 1: hardening step $H : L_{pc} \to L^c_{hc}$

Data: a sorted vector μ of typicality degrees, a membership threshold s, an ambiguity threshold t
Result: a vector h of class-selective assignements
begin
  if $\mu_1(x) < s$ **then**
    $h_i(x) \leftarrow 0 \ \forall i = 1, c$
  if $\sum_{i=1}^{c} h_i(x) > 0$ **then**
    for $i \leftarrow 1$ **to** c **do**
      if $\Phi_{1,i}(\mu(x)) \geq t$ **then**
        $h_i(x) \leftarrow 1$
      else
        $h_i(x) \leftarrow 0$
  return $h(x)$
end

4. Experiments and Results

To validate the efficiency of the proposed class-selective scheme, we present some results obtained by a resubstitution procedure on well-known real datasets from the UCI Machine Learning Repository [1] whose characteristics (number p of features, number c of classes, degree of overlap) are summarized in Table 2. The classification performance of some usual supervised classifiers with no reject options are given in Table 3 for comparison purpose : the Quadratic Bayes (QB) rule, the Nearest Neighbor ($1-NN$) rule and the Maximum Classifier (MC) based on typicality degrees in the feature space computed by (2) with $\alpha = 1$ and $d^2(x, v_i) = {}^t(x - v_i)\Sigma_i^{-1}(x - v_i)$ where the covariance matrix Σ_i and the center v_i of the class ω_i are estimated from the (learning) dataset. The same labeling L is used in the remaining experiments. We compare the performance of the proposed scheme to two class-selective rejection ones found in the literature. In [5], Ha has proposed to set the optimum cardinality of the set of selected classes by:

Table 2. Datasets used in the experiments.

data	p	c	overlap
iris	4	3	slight overlap, 2 classes
pima	8	2	medium overlap, 2 classes
vowel	10	11	slight overlap, by pairs
glass	9	6	strong overlap, up to 5 classes

Table 3. Error (E) and Correct (C) rates of some usual classifiers with no reject options.

data	%	QB	1-NN	MC
iris	E	2	4.67	2
	C	98	95.33	98
pima	E	25.39	29.53	32.55
	C	74.61	70.47	67.45
vowel	E	4.58	9.77	8.08
	C	95.42	90.23	91.92
glass	E	31.10	30.64	28.5
	C	68.90	69.36	71.5

$$k_{HA} = \min\{k \in \{1, \ldots, c\} | \mu_{k+1}(x) \geq t\} \quad (5)$$

where t is a user-defined ambiguity threshold whose role is similar to the one in our scheme. Since this selection can lead to unnatural classification areas, Horiuchi has proposed in [6] to use:

$$k_{HO} = \min\{k \in \{1, \ldots, c\} | \mu_k(x) - \mu_{k+1}(x) \geq t\} \quad (6)$$

Since these two selection schemes do not allow distance rejection, we set $s = 0$ in the experiments so that re-

sults can be compared. Note moreover that there are no outliers in the considered datasets.

The results obtained by a resubstitution procedure are given in Table 4 where HA and HO stand for the Ha and the Horiuchi schemes, Φ_S, Φ_A and Φ_H stand for the proposed scheme using the different triangular norms of Table 1 (with $\gamma = 0$ for the Hamacher one). The ambiguity threshold t is (coarse) tuned so that the error rate is as much as possible equal for each rejection scheme. For the tested datasets, a very little influence of the resolution parameter λ setting was observed and we chose to report the results with $\lambda = 10$. As expected, rejecting patterns leads to decrease the error rate compared to classifiers with no reject option (Table 3). Whatever the triangular norms, the proposed class-selective rejection scheme outperforms the ones proposed by Ha and Horiuchi with respect to the correct classification rate. This efficiency is due to the fact that the ratio of membership degrees is more suited than a simple difference to ambiguity rejection: the same difference ε between two low values and two high values with Horiuchi's method will not be discriminated. On the other side, with Ha's method, the most reliable membership degree is not taken into account, which leads to unnatural decisions. We constructed the operator such that our method reap the benefits of both Ha's and Horiuchi's schemes.

Table 4. Reject (R), Error (E) and Correct (C) rates of rejection schemes.

data	%	HA	HO	Φ_S	Φ_A	Φ_H
iris	R	0.67	1.33	0.67	0.67	0.67
	E	2	1.33	1.33	1.33	1.33
	C	97.33	97.33	98	98	98
pima	R	5.60	4.30	3.52	4.30	4.04
	E	30.47	30.21	30.59	30.08	30.24
	C	63.93	65.49	65.89	65.62	65.72
vowel	R	15.15	9.49	8.89	8.79	8.79
	E	4.34	4.34	4.34	4.34	4.34
	C	80.51	86.16	86.77	86.87	86.87
glass	R	23.83	8.88	7.94	8.17	8.02
	E	22.90	22.90	22.90	22.90	22.90
	C	53.27	68.22	69.16	68.93	69.08

5. Conclusion

In this paper, a new class-selective rejection scheme is proposed. It consits of two sequential step dealing with both reject options : distance rejection for outliers and ambiguity rejection for inliers. The latter option is based on an operator which aggregates the class-degrees of typicality of the pattern to be classified. This operator measures the blockwise similarity of sorted degrees by combining them with triangular norms and the Sugeno integral. Experimental results we obtained on well-known real datasets show that the proposed scheme achieves better recognition accuracy than other similar class-selective rules. Due to lack of place, we did not discuss the choice of the triangular norms, for which we have some theoritical results according to the nature of the degrees (in $\mathcal{L}_{pc}$, $\mathcal{L}_{fc}$). We will adress this problem in a forthcoming paper. Future works will concern the definition of blockwise similarity of numbers through fuzzy residual implication. We think this could generalize the concept of ambiguity for pattern recognition problems.

References

[1] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. Dept. of Information and Computer Science, University of California, Irvine, CA, http://archive.ics.uci.edu/ml/.

[2] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[3] B. Dubuisson and M. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26(1):155–165, 1993.

[4] M. Grabisch. *Pattern Recognition - From Classical to Modern Approaches*, chapter Fuzzy pattern recognition by fuzzy integrals and fuzzy rules, pages 257–280. World Scientific, 2002.

[5] T. M. Ha. The optimum class-selective rejection rules. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):608–615, 1997.

[6] T. Horiuchi. Class-selective rejection rule to minimize the maximum distance between selected classes. *Pattern Recognition*, 31(10):1579–1588, 1998.

[7] E. P. Klement and R. Mesiar. *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*. Elsevier, 2005.

[8] H. Le Capitaine, T. Batard, C. Frélicot, and M. Berthier. Blockwise similarity in [0,1] via triangular norms and sugeno integrals – application to cluster validity. In *IEEE International Conference on Fuzzy Systems*, pages 835–840, 2007.

[9] H. J. Zimmerman and P. Zysno. Quantifying vagueness in decision models. *European Journal of Operational Research*, 22(2):148–158, 1985.

Table des figures

Liste des tableaux