Modèles prudents en apprentissage statistique supervisé

Gen Yang

Par **Gen YANG**

Modèles prudents en apprentissage statistique supervisé

Thèse présentée
pour l'obtention du grade
de Docteur de l'UTC

Soutenue le 22 mars 2016
Spécialité : Technologies de l'Information et des Systèmes

D2263

THÈSE DE DOCTORAT DE
L'UNIVERSITÉ DE TECHNOLOGIE DE
COMPIÈGNE

Spécialité : Technologies de l'Information et

des Systèmes

Laboratoire Heudiasyc (UMR CNRS 7253)

Présentée par

Gen YANG

Sujet de la thèse :

Modèles prudents en apprentissage statistique supervisé

soutenue le 22 mars 2016

devant le jury composé de :

Mme Marie-Hélène Masson	Directrice de thèse
M. Sébastien Destercke	Co-directeur de thèse
M. Alessandro Antonucci	Rapporteur
M. Didier Coquin	Rapporteur
M. Yves Grandvalet	Examinateur
M. Mathieu Serrurier	Examinateur

Remerciements

Je tiens à commencer cette thèse par remercier mes deux directeurs, Mylène Masson et Sébastien Destercke. C'est pour moi une chance immense d'avoir eu des directeurs si responsables et si attentifs, je ne les remercierai jamais assez pour leurs conseils cruciaux pour l'avancement de mes travaux et pour leur soutien infaillible lors des moments difficiles. Leur rigueur scientifique dans la conduite de la recherche et leur clairvoyance dans l'appréhension des nouveaux sujets de recherche resterons toujours des exemples et des sources d'inspiration pour moi.

Je remercie également les membres de mon jury, Messieurs Alessandro Antonucci, Didier Coquin, Yves Grandvalet et Mathieu Serrurier, pour leur bienveillance et leur diligence lors de l'examen de mon mémoire.

J'adresse aussi ma gratitude à mes collègues du Laboratoire Heudiasyc : David Savourey et sa collègue Marie Albenque du Laboratoire LIX de l'Ecole Polytechnique pour leur aide sur le sujet de génération d'arbres binaires ; Gérard Govaert et Benjamin Quost qui m'ont donné l'opportunité et m'ont aidé pour assumer des charges d'enseignements au sein de l'UTC ; Philippe Xu, Xiao Liu et Marek Kurdej pour leurs conseils qui m'ont aidé à me repérer au début de ma vie de doctorant ; et enfin mes collègues de bureau Shameem, Alberto et Linh avec qui j'ai pu partager des discussions très enrichissantes grâce à notre diversité culturelle. Enfin, je remercie tout le personnel du Laboratoire et mes collègues doctorants qui ont fait de ces

trois années de ma vie de doctorant une expérience agréable et formidable.

Enfin, j'aimerais exprimer mes chaleureux remerciements à mes très chers amis : Bob, David Aurat, Grégoire Cotté, Haïtem Korfed et Remi Takase. La vie de doctorant est parfois monotone, mais vous avez donné des couleurs et des étincelles à ces trois ans de souvenirs : les week-ends "doctorants" avec Bob, David et Grégoire ; les discussions non-intelligibles pour toute autre personne avec Haïtem ; et les nombreux voyages gastronomiques avec Lemi-chan. Je suis extrêmement heureux que vous soyez rentrés dans ma vie.

Au final, j'aimerais finir ces longs remerciements par exprimer toute mon affection à mes parents et mes grands-parents. C'est vers vous que j'ai le plus de remords, ces trois années d'études m'ont obligé à rester très souvent loin de vous. En tant que fils et petit-fils, c'est un manquement de ma part, j'espère juste que j'aurai plus de temps à l'avenir pour vous exprimer mon amour pour vous. En tout cas, je peux maintenant vous dire fièrement que j'ai rempli une partie de vos attentes exprimées via le prénom que vous m'avez donné !

Table des matières

Chapitre 1

Introduction

"*Where is the wisdom we have lost in knowledge, and where is the knowledge we have lost in information*"

– Thomas Stearns Eliot

1.1 Contexte général

Le gagnant du prix Nobel de littérature, T.S. Eliot, avait exprimé son interrogation face aux progrès des technologies d'informations telles que la radio et la télévision au siècle précédent. L'information est devenue au cours du siècle dernier de plus en plus plébiscitée, notamment avec l'essor de l'informatique et des technologies de l'Internet. Désormais, des informations de fiabilité et de qualité très disparates sont disponibles à l'usage et à l'interprétation de tous. Elles sont souvent en telle quantité qu'il devient impossible pour un cerveau humain d'y discerner la connaissance qu'elles contiennent. De plus, il s'agit le plus souvent de données brutes que nous devons nous faire face de nos jours, et il est impossible pour un simple être humain d'y discerner une quelconque information utile : les données générées par les utilisateurs du Web participatif et celles que les divers outils technologiques nous permettent désormais de recueillir en sont des cas typiques. Face à ce constat, nous ne pouvons nous empêcher de relancer le questionnement de T.S. Eliot : où est l'information que nous avons perdue dans les données ?

C'est précisément à cette question que la communauté de l'apprentissage statistique, et plus généralement de l'intelligence artificielle, s'efforce de répondre : des modèles divers sont mis en place pour identifier les motifs et les caractéristiques des données brutes et les transformer en connaissances compréhensibles par les humains. Certains modèles prédictifs accordent également une importance particulière à l'obtention d'une représentation plus fiable des données et des connaissances qu'elles induisent, afin d'assister la prise de décisions de manière prudente. C'est notamment sur ces modèles que nous allons nous concentrer. Dans ce cas, les questions de T.S. Eliot prennent tout leur sens : la prise de décision est rarement un processus péremptoire dans la pratique où il suffit de faire confiance à la prédiction donnée par le modèle, il faut par exemple tenir compte de la conséquence des prédictions en cas d'erreur, ou encore des connaissances qu'un expert extérieur pourrait apporter. Ainsi, nous nous intéresserons dans le cadre de cette thèse à la manière dont nous pouvons mettre en place des modèles prédictifs fiables et prudents qui permettent d'intégrer des connaissances d'experts et d'assister le décideur pour faire des choix "sages".

Plus spécifiquement, dans le cadre de cette thèse, nous apportons nos contributions pour relier deux sujets de recherche distincts en apprentissage statistique : le cadre des probabilités imprécises et celui de l'apprentissage sensible aux coûts. Ces deux domaines visent tous les deux à rendre les modèles d'apprentissage et les inférences plus fiables et plus prudents. Pourtant peu de travaux existants ont tenté de les relier, en raison de problèmes à la fois théorique et pratique. Nos contributions consistent dans un premier temps à clarifier ces problèmes, ensuite nous proposerons un moyen de les combiner de manière efficace et nous évaluerons les avantages et pré-requis des modèles résultant de cette combinaison.

1.2 Synopsis

On présentera dans un premier temps dans le Chapitre 2 l'état de l'art général du domaine de l'apprentissage statistique en posant le cadre et les notations formelles que nous utiliserons. Nous introduirons notamment la

notion de fonction de coûts (d'erreurs de classification) et le cadre de l'apprentissage sensible aux coûts. Dans le Chapitre 3, nous verrons comment nous pouvons étendre l'espace de prédictions habituel d'un problème multiclasse pour avoir des prédictions sous forme d'ensembles et nous introduirons le cadre des probabilités imprécises. Nous soulignerons notamment la similitude en terme d'objectif entre le cadre de l'apprentissage sensible aux coûts et celui des probabilités imprécises : les deux cherchent à rendre les prédictions plus fiables, mais nous remarquerons la difficulté pratique à les combiner. Dans le Chapitre 4, nous poserons une base théorique pour combiner ces deux cadres en proposant des propriétés et une formule générale permettant de dériver des coûts pour les prédictions sous forme d'ensembles. Dans le Chapitre 5, nous verrons une méthode pratique, les dichotomies emboîtées, pour construire des modèles prédictifs combinant les deux cadres en question. Nous y élaborerons également une série d'expériences pour caractériser le comportement et les intérêts de cette méthode. Enfin, dans le Chapitre 6, nous présenterons une autre série d'expériences avec des jeux de données ordinales pour juger de l'efficacité de la formule proposée dans le Chapitre 4 et pour caractériser les diverses approches pour produire des prédictions sous forme d'ensembles.

Chapitre 2

Préliminaires et cadre général

L'apprentissage statistique est une technique d'analyse de données multivariées qui, étant donné les variables d'entrée $\mathbf{X} = (X_1 \times \ldots \times X_m,)$ et une variable de sortie Y, a pour but d'assigner une prédiction $Y = \hat{y}$ (à valeur dans l'espace $\mathcal{Y}$) à une observation $\mathbf{X} = \mathbf{x}$ issue d'un espace $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1 \times \ldots \times \mathcal{X}_m)$.

Quand l'espace $\mathcal{Y}$ (que nous appellerons également espace de prédictions) est fini, nous appelons alors cette tâche un problème de discrimination ou de classification. Quand cet espace est continu, nous parlons de régression. Les problèmes de régression peuvent facilement être transformés en problème de classification en discrétisant la variable de sortie, et nous les référons sous le nom de classification ordinale.

Dans le cas d'une classification, l'espace de sortie est défini en fonction d'un ensemble d'étiquettes (aussi appelées l'espace des classes ou labels) $\Omega = \{\omega_1, \ldots, \omega_K\}$. En pratique, la vérité y correspondant à une observation $\mathbf{x}$ peut être de nature variée. Dans le cas où la vérité correspond à une étiquette unique ($y \in \Omega$), nous avons affaire à un problème de classification multiclasse. En revanche, si la vérité correspond à un ensemble d'étiquettes ($y \in 2^\Omega$ où 2^Ω est l'ensemble des parties de Ω), alors nous parlons d'un problème multilabel.

Cependant, pour un problème multiclasse, nous pouvons autoriser la prédiction à s'exprimer sous la forme d'un ensemble d'étiquettes ($\mathcal{Y} = 2^\Omega$) pour accroître la prudence du modèle, nous parlons alors d'un modèle prudent ou indéterminé, par opposition au modèle multiclasse usuel où seules les prédictions sous forme de singleton sont autorisées ($\mathcal{Y} = \Omega$). Le Tableau 2.1 récapitule les différences entre ces problèmes. Nous nous intéresserons essentiellement aux problèmes multiclasses dans ce chapitre, et les modèles prudents seront introduits dans le Chapitre 3.

Les algorithmes d'apprentissage peuvent aussi être classés en trois familles en fonction de la nature des données à leur disposition. Si les algorithmes sont entraînés à partir de données (dites d'apprentissage) étiquetées

$$\mathcal{D} = (\mathbf{x}_i, y_i)_{i \in [1;N]},$$

c'est-à-dire portant à la fois sur l'espace d'entrée $\boldsymbol{\mathcal{X}}$ et des étiquettes

$\mathcal{Y}$ \ Vérités	Ω	2^Ω
Ω	multiclasse	Impossible
2^Ω	multiclasse prudent	multilabel

TABLE 2.1: Tableau récapitulatif des problèmes de classification selon la taille de l'espace des vérités et des prédictions

Ω, alors nous parlons d'apprentissage supervisé. Le but ici est d'apprendre un schéma de correspondances qui permet d'associer automatiquement de nouvelles données d'entrée $\mathbf{x}'$ non étiquetées à une estimation $\hat{y}$ d'étiquette possible.

Si au contraire, nous ne disposons pas d'information sur l'espace des étiquettes (ni les éléments qui la composent, ni sa cardinalité), et que l'algorithme doit déduire lui-même une structure inhérente aux données $\mathbf{x}_i$ (non étiquetées), nous parlons alors d'apprentissage non-supervisé.

Au final, certains algorithmes combinent l'utilisation de ces deux types de données, afin d'améliorer la performance prédictive quant à la prédiction des classes y' sur des nouvelles données, nous parlons alors d'apprentissage semi-supervisé.

Dans le cadre de cette thèse, nous nous concentrons plus particulièrement sur le problème de **classification multiclasse supervisée** et à la mise en place de modèles prudents.

2.1 Cadre théorique de l'apprentissage supervisé

Nous commencerons par formaliser la tâche d'apprentissage supervisé, en s'aidant de la théorie des probabilités et de la notion de fonction de coûts. Nous verrons comment ce cadre est appliqué en pratique avec une

présentation des problématiques et des méthodes de l'état de l'art. Parmi les diverses méthodes existantes, nous choisirons d'étudier en détail les modèles probabilistes et plus particulièrement le modèle Bayésien naïf que nous utiliserons pour la suite.

2.1.1 Objectif et notations

Étant donné l'ensemble $\mathcal{D}$ des données étiquetées d'une taille finie N, l'objectif de la classification supervisée se résume à construire (apprendre) une fonction (que nous appellerons également *classifieur*)

$$f : \begin{cases} \boldsymbol{\mathcal{X}} \to \mathcal{Y} \\ \mathbf{x} \to \hat{y} = f(\mathbf{x}) \end{cases} \tag{2.1}$$

qui soit capable d'associer une prédiction $\hat{y}_i$ à l'observation $\mathbf{x}_i$. Cette prédiction se doit d'être la "meilleure" possible par rapport à la vérité y_i associée à chaque observation $\mathbf{x}_i$. Cette étape de construction de classifieur est aussi appelée étape d'apprentissage. Nous formalisons cette notion de "meilleure" dans les paragraphes suivants.

De plus, f doit pouvoir fournir de bonnes prédictions pour des instances de données futures $\mathbf{x}'$ dont les vérités associées sont supposées inconnues pour le classifieur. C'est l'étape de prédiction. Ces étapes sont illustrées sur la Figure 2.1.

Pour décrire théoriquement le fonctionnement de ces étapes, nous faisons l'hypothèse classique que les données sont issues d'une distribution jointe $\mathbb{P}[\mathbf{X}, Y]$ (Billingsley, 1979), dont nous donnons ici la fonction de masse :

$$p(\mathbf{x}, y) : \begin{cases} (\mathcal{X}_1, \ldots, \mathcal{X}_m, \Omega) \ \to [0, 1] \\ (x_1, \ldots, x_m, y) \to p(\mathbf{x}, y) \end{cases} . \tag{2.2}$$

La prédiction met également souvent en jeu le calcul de la probabilité conditionnelle, que nous pouvons obtenir par normalisation si la distribution

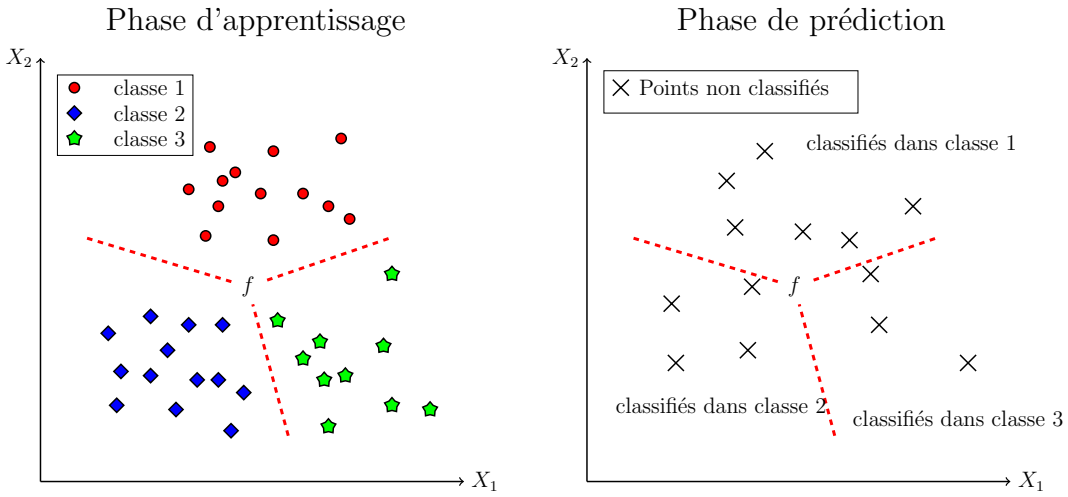


FIGURE 2.1: Illustration d'un problème d'apprentissage supervisé où nous cherchons à distinguer trois formes de points selon leur coordonnées dans (X_1, X_2). Les lignes de séparation en pointillée rouge sont censées représenter le classifieur f construit. C'est lui qui permet de décider à quelle classe les nouveaux points appartiennent.

jointe est connue pour toute y :

$$p(y|\mathbf{x}) : \begin{cases} \Omega \to [0,1] \\ y \to \dfrac{p(\mathbf{x}, y)}{\sum\limits_{y' \in \Omega} p(\mathbf{x}, y')} \end{cases} . \tag{2.3}$$

Quand il n'y aura pas de risque d'ambiguïté, nous la noterons simplement par $p(y)$.

Une fois ces probabilités estimées, le classifieur s'en sert pour inférer des prédictions. L'inférence doit donner de bonnes estimations des vérités. Pour ceci, il est nécessaire de définir une métrique pour mesurer la qualité de ces estimations.

2.1.2 Fonction de coûts

La fonction de coûts sert à quantifier l'erreur de prédiction par rapport à la vérité. Elle introduit l'idée que les erreurs de prédictions (nous parlerons aussi d'erreur de classification) ne se valent pas, et que certaines erreurs

peuvent être plus coûteuses que d'autres. L'intérêt des fonctions de coûts est donc de modéliser le coût d'erreur, *i.e.*, le coût de prendre une mauvaise décision. Pour formaliser cette notion, nous associons à chaque prédiction possible $f(\mathbf{x}) = \hat{y} (\in \mathcal{Y})$ une fonction de coût :

$$c_{\hat{y}} : \begin{cases} \Omega \to \mathbb{R}^+ \\ y \to c_{\hat{y}}(y) = c_{f(\mathbf{x})}(y) \end{cases} \tag{2.4}$$

telle que $c_{\hat{y}}(y)$ est le coût de prédire la classe $\hat{y} \in \mathcal{Y}$ quand $y \in \Omega$ est la vérité. L'Exemple 1 donne une illustration simple de cette notion du coût d'erreur.

Dans le cas le plus simple, que nous appellerons également le cas des coûts 0/1, ce coût est unitaire pour toutes les classes : le coût de prédire la bonne classe ($\hat{y} = y$) est toujours nul, et le coût de donner une mauvaise prédiction est toujours de 1. Nous utiliserons la notation

$$c_{\hat{y}}(y) = \mathbb{1}_{\hat{y} \neq y},$$

où $\mathbb{1}$ est la fonction indicatrice ($\mathbb{1}_A = 1$ si A vrai, 0 sinon).

Exemple 1. *Nous considérons ici un problème de reconnaissance d'obstacles pour véhicules intelligents qui servira d'illustration récurrente. L'ordinateur de bord du véhicule doit reconnaître si ce dernier fait face à un humain (h), une bicyclette (b) ou aucun obstacle (n) (i.e. $\Omega = \{h, b, n\}$).*

Comme l'humain et la bicyclette sont tous les deux des obstacles à éviter, une confusion entre h et b n'est donc pas très importante et entraîne un coût faible mais non nul, car les réponses optimales à chaque obstacle sont légèrement différentes en fonction de leur mobilité. Par contre, prédire h ou b quand il n'y a pas d'obstacles est déjà plus coûteux, puisque le véhicule va faire une manœuvre inutile. Enfin, prédire n quand il y a un obstacle fait encourir un grand risque car ceci peut causer un accident. Ainsi, nous voyons que les informations sur le rapport entre les classes peuvent être exprimées clairement à l'aide de cette notion de coût d'erreurs. Le Tableau 2.2 illustre un exemple de fonctions de coûts sous forme d'une matrice de coûts.

Les fonctions de coûts $c_{\hat{y}}$ s'expriment également comme des matrices $\mathcal{M}$ où pour chaque case nous avons $\mathcal{M}_{i,j} = c_{\hat{y}_i}(y_j)$.

	vérité		
$c_{\hat{y}}(y)$	$y = h$	$y = b$	$y = n$
$\hat{y} = h$	0	1	2
$\hat{y} = b$	1	0	2
$\hat{y} = n$	4	4	0

TABLE 2.2: Matrice de coûts définie selon le niveau de risque

2.1.3 Classifieur optimal

Un classifieur f optimal cherche à minimiser le coût espéré sur l'ensemble de l'espace $(\boldsymbol{\mathcal{X}}, \mathcal{Y})$ pour produire de bonnes prédictions non seulement sur les données d'apprentissage $\mathcal{D}$, mais aussi sur des données futures. Il s'agit alors de minimiser le risque moyen $\mathcal{R}_{moy}$:

$$\begin{aligned} \mathcal{R}_{moy} &= \mathbb{E}[c_{f(\mathbf{x})}(y)] \\ &= \int_{\mathbf{x}\in\boldsymbol{\mathcal{X}}} \sum_{y\in\Omega} p(\mathbf{x}, y) c_{f(\mathbf{x})}(y) d\mathbf{x}. \end{aligned} \tag{2.5}$$

Nous notons aussi que, si l'espace $\mathcal{Y}$ est continu, par exemple le cas de la régression, nous aurons une formule plus générale :

$$\mathcal{R}_{moy} = \int_{\mathbf{x}\in\boldsymbol{\mathcal{X}}} \int_{y\in\Omega} p(\mathbf{x}, y) c_{f(\mathbf{x})}(y) d\mathbf{x} dy. \tag{2.6}$$

Un classifieur f qui minimise $\mathcal{R}_{moy}$ est appelé le classifieur optimal. En utilisant la règle de chaînage qui dit que

$$p(\mathbf{x}, y) = p(y|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|y)p(y),$$

l'Équation (2.5) devient

$$\mathcal{R}_{moy} = \int_{\mathbf{x}\in\boldsymbol{\mathcal{X}}} \Big[\sum_{y\in\Omega} c_{f(\mathbf{x})}(y) p(y|\mathbf{x}) \Big] p(\mathbf{x}) d\mathbf{x}. \tag{2.7}$$

Nous remarquons que la minimisation de $\mathcal{R}_{moy}$ revient à celle du terme entre les crochets pour tout $\mathbf{x} \in \boldsymbol{\mathcal{X}}$. Ainsi, pour obtenir le classifieur optimal au sens Bayésien, il suffit de résoudre (pour tout $\mathbf{x}$) :

$$f(\mathbf{x}) = \arg\min_{\hat{y}\in\mathcal{Y}} \sum_{y\in\Omega} c_{\hat{y}}(y) p(y|\mathbf{x}). \tag{2.8}$$

Dans le cas des coûts 0/1, nous pouvons simplifier l'Équation (2.7) et nous obtenons :

$$f(\mathbf{x}) = \arg\max_{\hat{y} \in \mathcal{Y}} p(\hat{y}|\mathbf{x}). \tag{2.9}$$

C'est ce classifieur que nous appelons communément classifieur optimal de Bayes.

Cependant, même si beaucoup de classifieurs de l'état de l'art se servent de ce cadre probabiliste comme fondement théorique, tous n'utilisent pas directement les probabilités en pratique. Nous allons voir maintenant la mise en œuvre pratique de ce cadre théorique.

2.2 L'apprentissage supervisé en pratique

Dans la pratique, il est évident qu'il est impossible de réaliser la minimisation évoquée dans les Équations (2.5) et (2.7) pour tout $\mathbf{x}$, étant donné que nous n'avons qu'une quantité finie de données à notre disposition, et que les distributions des probabilités conditionnelles et jointe sont en général inconnues. Ainsi, nous ne pouvons qu'estimer une distribution approchée $\hat{p}$ à partir des données d'apprentissage $\mathcal{D}$ et calculer le risque empirique :

$$\mathcal{R}_{emp} = \frac{1}{N} \sum_{i=1}^{N} c_{f(\mathbf{x}_i)}(y_i), \tag{2.10}$$

en supposant que les données sont indépendantes et identiquement distribuées (i.i.d.) par rapport à $p(\boldsymbol{\mathcal{X}}, \Omega)$.

Dans le cas des coûts unitaires, ceci revient à calculer simplement le taux de mauvaises prédictions.

2.2.1 Divers algorithmes de l'état de l'art

Tous les algorithmes de l'état de l'art ne cherchent pas à estimer les probabilités jointe ou conditionnelles. Certains classifieurs cherchent à relier directement les entrées et les sorties, par exemple en optimisant des frontières de séparation entre les sorties, ou en utilisant une approche fonctionnelle.

Parmi ces modèles non probabilistes, nous pouvons citer à titre d'exemples :

- Les **séparateurs à vaste marge** (SVM) (Cortes et Vapnik, 1995) qui ont pour objectif de construire une séparation linéaire maximisant la marge entre les données appartenant à des classes différentes. Elles peuvent s'appliquer à un problème non linéaire en utilisant un plongement des données dans un espace de dimension supérieure (*cf.* "astuce de noyau", basée sur le Théorème de Mercer (1909)) qui rend le problème linéaire.
- Les **réseaux de neurones** qui sont un ensemble de méthodes d'apprentissage inspirées de leur homologue biologique (McCulloch et Pitts, 1943). Ils consistent à imiter artificiellement le fonctionnement des neurones avec des fonctions logiques et arithmétiques.

Dans la catégorie des modèles probabilistes, nous pouvons notamment citer :

- La **régression logistique** (Cox, 1958) qui cherche à estimer la probabilité *a posteriori* via une régression linéaire sur les entrées $\boldsymbol{\mathcal{X}}$.
- Les **modèles de mélange** (Bailey et Elkan, 1994) qui modélisent la probabilité jointe comme un mélange de lois de probabilité de même famille (par exemple un ensemble de gaussiennes, de distribution de Dirichlet...), mais de paramètres différents.
- Les **arbres de décisions** qui sont des modèles graphiques sous forme arborescente. L'arbre construit cherche à séparer, de manière récursive, l'espace des données d'apprentissage en fonction des différentes classes. Parmi les méthodes les plus connues, nous pouvons citer le C4.5 (Quinlan, 1993), ou encore CART (Breiman et collab., 1984).
- Le **classifieur Bayésien naïf** (que nous appellerons NBC) qui est un modèle Bayésien simple où tous les variables d'entrée $(X_1, \ldots, X_m)$ sont supposées indépendantes étant donné la classe, ce qui permet de simplifier le calcul de l'Équation (2.8).
- Les **réseaux Bayésiens** (Friedman et collab., 1997) qui sont des modèles probabilistes graphiques qui expriment les probabilités à l'aide d'un graphe orienté acyclique dont les nœuds sont les variables d'entrées et la classe, et dont les arcs spécifient les relations de dépendance

avec des probabilités. En même temps, l'hypothèse de Markov permet de représenter les relations d'indépendances conditionnelles des nœuds. Ces modèles ont un grand pouvoir représentatif. Par exemple, le NBC peut être représenté par un réseau Bayésien simple ; le **modèle de Markov caché** (Baum et Petrie, 1966) peut être considéré comme un réseau Bayésien dynamique (Murphy, 2002).

Il existe également des modèles qui ne rentrent pas réellement dans le cadre de l'apprentissage statistique standard (dans le sens où ils ne sont pas basées sur l'hypothèse probabiliste de l'Équation (2.2)) :

- Les modèles d'apprentissage à base de règles consistent à déduire les prédictions en se basant sur des règles de la logique formelle (Kamp, 1981).
- Le "k plus proches voisins" (k-PPV), qui consiste à déduire la prédiction associée à une nouvelle observation en se basant sur les instances de données d'apprentissage les plus proches d'elle, selon une mesure de distance à définir. Ainsi, il n'y a pas de phase d'apprentissage à proprement parler.

Dans notre travail, nous cherchons surtout à obtenir une représentation plus fiable des données et des connaissances qu'elles induisent afin de construire un modèle plus prudent (nous détaillerons ces notions dans le chapitre suivant), il nous a donc semblé plus intéressant de nous orienter vers les modèles probabilistes.

2.2.2 Le cas du classifieur Bayésien naïf

Même si la plupart des travaux conduits dans le cadre de cette thèse ont une portée assez générale, il nous parait important d'illustrer et d'appliquer les résultats théoriques avec un classifieur pratique. Nous avons choisi de le faire avec le NBC notamment parce qu'il s'agit d'un classifieur probabiliste simple à manipuler et donnant de bonnes performances prédictives malgré sa simplicité conceptuelle. Nous détaillons d'abord quelques éléments communs aux modèles bayésiens et ensuite les spécificités de ce classifieur.

Nous avons évoqué précédemment que la distribution de probabilités

réelle p n'est pas connue et il est nécessaire d'estimer une distribution $\hat{p}$ en s'aidant uniquement des données disponibles $\mathcal{D}$. Le théorème de Bayes nous permet d'inférer la probabilité *a posteriori* à partir de celle *a priori* :

$$\hat{p}(y|\mathbf{x}) = \frac{\hat{p}(\mathbf{x}|y)\hat{p}(y)}{\hat{p}(\mathbf{x})} \tag{2.11}$$

En pratique, estimer $\hat{p}(\mathbf{x}|y)$ et $\hat{p}(\mathbf{x})$ est un problème très complexe. Ainsi, il est classique de faire des hypothèses pour simplifier cette étape d'estimation. Par exemple, les modèles bayésiens supposent que $\hat{p}$ suit une distribution *a priori* définie par une loi de probabilité connue (par exemple, loi gaussienne, uniforme,...). Il s'avère que ces modèles peuvent donner de bonnes performances prédictives, même si la validité de l'hypothèse de départ sur la distribution de p n'est pas garantie. Le cas le plus typique est celui du classifieur Bayésien naïf.

Le NBC suppose que tous les attributs $\boldsymbol{\mathcal{X}} = (X_1, \ldots, X_m)$ sont indépendants conditionnellement à la classe :

Hypothèse 1 (Indépendance conditionnelle des attributs)**.**

$$p(x_1, \ldots, x_m \mid y) = \prod_{i=1}^{m} p(x_i \mid y).$$

Ceci permet alors de simplifier le calcul de la probabilité *a posteriori* en l'appliquant à l'Équation (2.11) :

$$\hat{p}(y|\mathbf{x}) = \frac{\hat{p}(y)\prod_{i=1}^{m}\hat{p}(x_i|y)}{\sum_{y'\in\Omega}\hat{p}(y')\prod_{i=1}^{m}\hat{p}(x_i \mid y')} \tag{2.12}$$

Le NBC est connu pour avoir de bonnes performances prédictives dans le cas de coûts unitaires, même lorsque l'hypothèse d'indépendance est clairement violée (Domingos et Pazzani, 1997). Sur la Figure 2.2, nous présentons une illustration du NBC sous forme d'un réseau Bayésien simple. Des extensions du NBC ont été également proposées pour obtenir des hypothèses plus réalistes. Par exemple, le classifieur Bayésien naïf augmenté par un

arbre (*tree-augmented naive bayes*, Kohavi (1996)) propose de modéliser les dépendances entre les variables d'entrée avec une structure arborescente (cf. Figure 2.3).

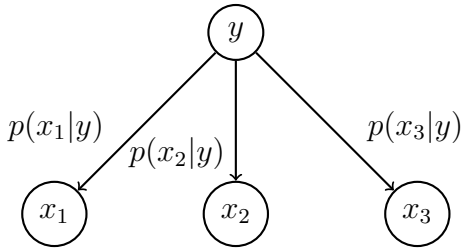


Figure 2.2: Un classifieur Bayésien naïf présenté sous forme d'un réseau Bayésien simple (un arbre avec un seul niveau de profondeur). Les seules relations causales sont celles qui relient la classe $\mathcal{Y}$ avec les variables d'entrée $\boldsymbol{\mathcal{X}} = (X_1, X_2, X_3)$ qui sont indépendantes entre elles.

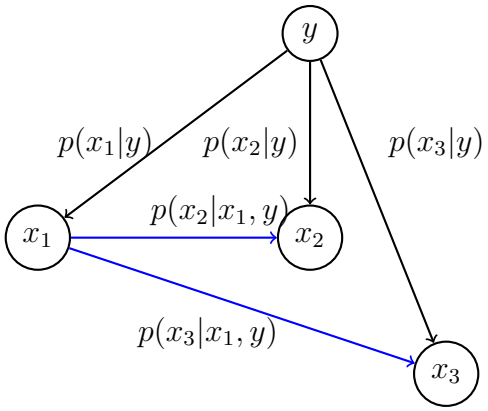


Figure 2.3: Un NBC augmenté représenté par un réseau Bayésien : il est augmenté par la partie en bleu du graphe qui spécifie une dépendance entre les variables X sous forme arborescente.

2.2.3 Le cadre de l'apprentissage et de la classification sensible aux coûts

Dans la pratique, tous les classifieurs capables de produire des probabilités *a posteriori* ne sont pas adaptés pour l'apprentissage sensible aux

coûts. Car certains classifieurs de l'état de l'art (comme le NBC), conçus initialement dans le cadre des coûts 0/1, ne sont pas des bons estimateurs de probabilités.

Si cette mauvaise qualité d'estimation n'impacte pas ou peu leur efficacité pour la prédiction avec des coûts unitaires (*i.e.*, ils estiment quand même correctement le rang de chaque classe), cette performance n'est plus garantie quand nous travaillons avec des coûts génériques (Deirdre et collab., 2008). Il est donc mal-avisé d'utiliser des classifieurs probabilistes sans garantie de qualité de leurs estimations quand il faut également intégrer les coûts d'erreur de classification (non unitaires).

En effet, dû à cette difficulté pratique, les classifieurs qui prennent en compte ces coûts d'erreur de manière générique (visant ainsi à résoudre l"Équation (2.8)) rentrent dans une catégorie particulière que nous référons sous le nom d'apprentissage sensible aux coûts (*cost-sensitive learning*).

Ce dernier est devenu un sujet important dans la recherche sur l'apprentissage statistique, notamment parce qu'il se rapproche davantage des applications réelles. Il est beaucoup appliqué aux données présentant une distribution de classes fortement disproportionnée (Japkowicz et Stephen, 2002; Maloof, 2003), par exemple en diagnostic médical, où la proportion des individus sains est souvent largement supérieure aux malades.

Dans l'état de l'art, à moins de concevoir un classifieur spécialement adapté pour l'apprentissage sensible aux coûts (Drummond et Holte, 2000; Ling et collab., 2004), deux approches principales sont proposées pour adapter les classifieurs standards et les rendre sensibles aux coûts.

La première approche consiste à effectuer un pré-traitement des données avec un ré-échantillonnage (Elkan, 2001; Zadrozny et collab., 2003) ou en donnant des poids à chaque instance de données (Ting, 1998). Ces pré-traitements permettent de remédier au problème de qualité des estimations, car l'information véhiculée par les coûts est intégrée directement dans la distribution (ré-échantillonnée) des classes. L'avantage de cette approche est qu'un classifieur standard (insensible aux coûts) peut ensuite être appliqué, mais le fait de modifier la distribution des classes est problématique en soi : la distribution modifiée n'est plus i.i.d. et des probabilités (délibé-

rément) faussées sont estimées (ce qui pose des problèmes si nous voulons combiner l'utilisation d'autres modèles, comme le modèle prudent que nous évoquerons dans le chapitre suivant).

La seconde approche consiste à faire des post-traitements sur les sorties ou les probabilités estimées d'un classifieur insensible aux coûts. Il peut s'agir d'une technique de seuillage des probabilités Domingos (1999); Chai et collab. (2004), mais si elle ne modifie pas les données d'apprentissage, elle ne résout pas automatiquement le problème de la qualité des estimations non plus. Il est nécessaire de développer des techniques spécifiques (Margineantu, 2002; Sheng et Ling, 2006) pour ceci. Un autre problème est que si la démarche de seuillage est intuitive à interpréter en classification binaire, elle l'est beaucoup moins dans le cas multiclasse.

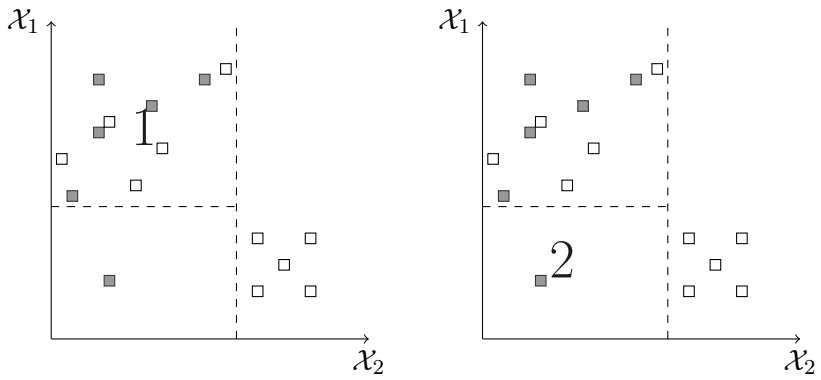
Certains font une distinction entre ces deux approches en qualifiant la première d'apprentissage (à proprement parler) sensible aux coûts, et la seconde de classification sensible aux coûts, car les coûts n'interviennent qu'à l'étape de la prédiction dans la seconde approche. Cependant, comme l'objectif final de ces approches reste identique (c.à.d. de tenir compte des coûts d'erreur de classification), nous confondrons volontairement ces deux termes dans la suite de notre travail.

Dans nos travaux, nous allons aborder ce problème d'intégration de coûts d'une manière différente. En effet, puisque la qualité des estimations de probabilités est un facteur clé, nous allons développer dans le chapitre suivant des modèles prudents qui permettent de tenir compte de la fiabilité de ces estimations.

Chapitre 3

Modèles prudents et probabilités imprécises

Dans certains contextes de l'apprentissage statistique (reconnaissance des signatures bancaires, diagnostic médical, décisions ayant de fort impact ...), il peut s'avérer préférable d'avoir des prédictions ou des modèles plus prudents et plus fiables, au lieu de se concentrer seulement sur le pouvoir prédictif (par exemple le taux de bonnes prédictions). En effet, les estimations que nous obtenons ne nous permettent pas toujours de prendre des décisions fiables. Comme l'illustre la Figure 3.1 [1], ceci peut être dû au fait que, soit plusieurs classes ont des probabilités élevées et similaires (nous parlerons alors de situation d'ambiguïté), soit il n'y a pas assez de données pour avoir des estimations fiables (nous parlerons de manque d'information).



Ambiguïté des estimations
$p(\square|1) \in [0.49, 0.51]$
$p(\blacksquare|1) \in [0.49, 0.51]$

Manque d'information
$p(\square|2) \in [0, 0.7]$
$p(\blacksquare|2) \in [0.3, 1]$

Figure 3.1: Illustration d'une situation d'ambiguïté vs. de manque d'information.

Ainsi, nous entendons par "modèles prudents", des modèles prédictifs capables de tenir compte de ces situations et d'adapter les décisions en conséquence pour accroître la fiabilité et la prudence des prédictions.

Dans ce chapitre, nous verrons d'abord comment ces modèles peuvent être mis en place avec l'option de rejet par ambiguïté (Chow, 1970). Ceci permettra notamment d'introduire l'idée des prédictions indéterminées. Enfin, nous détaillerons le cadre des probabilités imprécises, qui a un intérêt

1. Nous verrons comment obtenir des estimations de probabilités sous forme d'intervalles dans la Seciton 3.2 et 3.3.1.

non seulement dans le cadre des modèles prudents, mais aussi dans le cadre de la classification sensible aux coûts.

3.1 Modèles prudents fondés sur les probabilités standards

Les estimations de probabilités conditionnelles des classes que nous obtenons avec les classifieurs probabilistes correspondent à des mesures de ce que nous pensons connaître ou de ce que nous croyons sur ces classes. Elles contiennent donc naturellement des informations qui peuvent nous servir à quantifier l'incertitude des situations, et par conséquent rendre le modèle plus prudents.

Le problème est que ces informations ne sont pas exprimées explicitement par les prédictions produites par les classifieurs standards. Par exemple, dans le premier graphe de la Figure 3.1, si $p(\square|1)$ était à 0.49 et que $p(\blacksquare|1)$ était à 0.51, alors la classe ■ sera attribuée à la région 1 avec un classifieur probabiliste standard (sous coûts 0/1), même si les estimations de probabilités traduisent une forte similitude des deux classes, plutôt que la dominance de l'une sur l'autre.

Nous verrons ici quelques approches pour remédier à ce problème d'ambiguïté des estimations, et comment les coûts d'erreur de classification peuvent être intégrés pour aboutir enfin aux limites du cadre probabiliste standard quant à la conception d'un modèle prudent.

3.1.1 L'option de rejet

Nous avons vu que l'objectif d'un classifieur est normalement d'associer une classe à une observation d'entrée. Le principe d'une option de rejet est d'enrichir l'espace de prédictions avec l'alternative de ne pas donner de prédiction. Dans ce cas, on dit qu'il y a *rejet* ou que la prédiction est *rejetée*.

En reprenant la formalisation du classifieur optimal avec l'Équation (2.9), nous voyons que si la prédiction optimale $\hat{y}^*$ a une probabilité $p(\hat{y}^*)$ proche de $1/K$ ($K = |\Omega|$), alors le classifieur ne fait pas mieux que de jeter un dé

à K faces aléatoirement (dans le cas des coûts unitaires). Dans ce cas, il est intéressant pour le classifieur de s'abstenir de prédire, pour exprimer un doute sur la situation. [2] Ainsi, en adoptant l'option de rejet, nous détectons et évitons les prédictions qui présentent un grand risque de conduire à une erreur de classification.

Il existe plusieurs manières de définir l'option de rejet dans la littérature, nous détaillons ici la formalisation de Chow (1970) en raison de sa proximité avec le cadre probabiliste que nous étudions. Il est cependant important de noter que cette formalisation est donnée uniquement dans le cadre des coûts 0/1. Le cas où les erreurs de classification ont des coûts génériques n'est formalisé que dans le cas des problèmes binaires (Herbei et Wegkamp, 2006; Bartlett et Wegkamp, 2008).

Dans Chow (1970), l'option de rejet est choisie si la probabilité *a posteriori* correspondant à la prédiction optimale $\hat{y}^*$ est inférieure à un seuil t ($t > 1/K$) :

$$\max_{\hat{y} \in \mathcal{Y}} p(\hat{y}|\mathbf{x}) < t \Rightarrow \text{rejet de prédiction.}$$

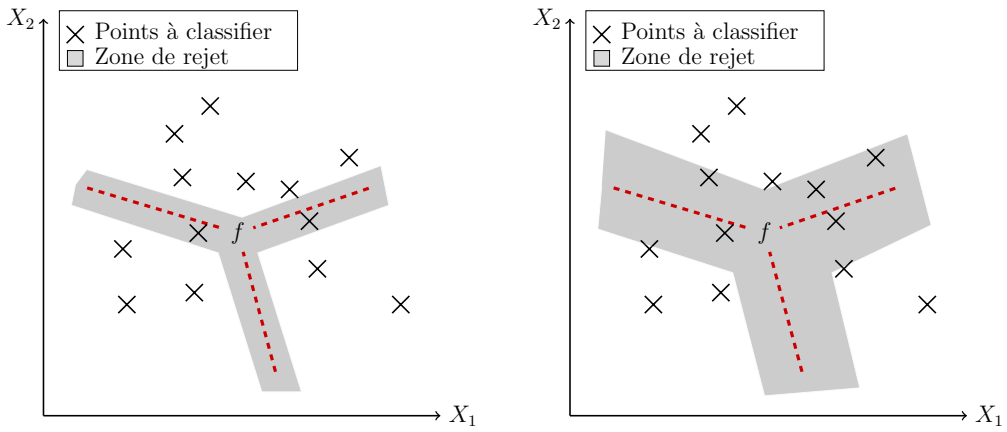
Chow (1970) démontre trois propriétés essentielles définissant le comportement de t :

- le taux d'erreur (en cas de non rejet) $E(t)$ et le taux de rejet $R(t)$ sont des fonctions respectivement décroissante et croissante de t
- t est une borne maximale pour le taux d'erreur
- t est un compromis entre le taux d'erreur et le taux de rejet

Nous illustrons le fonctionnement de ce seuil de rejet sur la Figure 3.2. Définir un seuil de rejet t revient à définir une zone de flou autour des frontières de séparation des classes : plus le seuil t est bas, plus cette zone est grande et plus il y a de rejet, et par conséquent, moins il y aura d'erreurs de classification.

Par ailleurs, Chow a démontré qu'il s'agit du meilleur compromis dans le cadre des classifieurs Bayésiens. Ainsi, il suffit au final de choisir la valeur

2. Plus spécifiquement, nous parlons ici du rejet dû à l'ambiguïté. Dans l'état de l'art, il existe également une notion de "rejet dû à la distance" (Dubuisson et Masson, 1993) qui est surtout utilisé dans le cadre de détection de nouveauté et qui se rapproche davantage de la notion de manque d'information que nous avons mentionnée.



Seuil de rejet t élevé — Seuil de rejet t faible

FIGURE 3.2: Effet du seuil de rejet

du seuil selon le taux d'erreur que l'on est prêt à accepter, comme l'illustre la Figure 3.3.

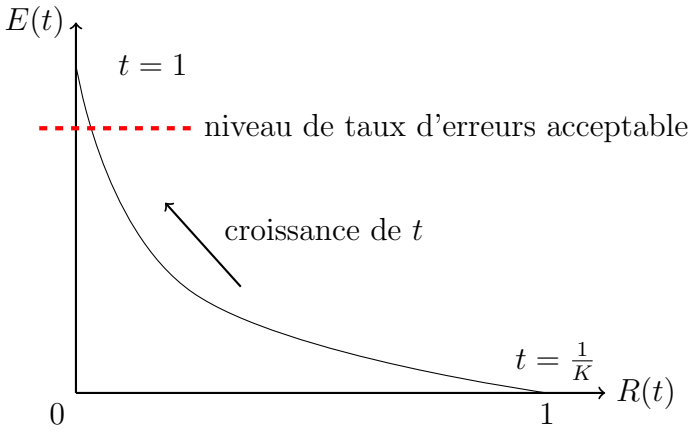


FIGURE 3.3: Courbe de compromis entre le taux d'erreur $E(t)$ et le taux de rejet $R(t)$

Cependant, cette optimalité est seulement théorique (dans le cas où les estimations sont sans erreur). Dans la pratique, il est souvent nécessaire de définir des seuils multiples (un par classe) (Fumera et collab., 2000). Il existe également d'autres approches de rejet, comme la définition d'une *classe de rejet* (Frélicot et collab., 1995) dans le problème de classification,

dont il faut tenir compte lors de l'apprentissage au même titre que les autres classes.

L'option de rejet est typiquement utile dans les applications comme le diagnostic médical où le coût de commettre une erreur sur une prédiction est supérieur au coût de chercher des alternatives pour affiner la prédiction (recueillir des nouveaux attributs, prendre un autre classifieur, demander à un expert...)

3.1.2 Extension de l'espace de prédictions $\mathcal{Y}$

L'option de rejet est une première approche aux modèles prudents où la seule alternative aux prédictions classiques est de s'abstenir de prédire. Nous allons voir maintenant comment nous pouvons obtenir des modèles plus informatifs et moins radicaux (en terme de prédiction), en étendant l'espace de prédictions à $2^{\Omega} \setminus \emptyset$ au lieu de Ω.

Sur le plan formel

Jusqu'à présent, nous avons toujours supposé que les prédictions $\hat{y}$ étaient des éléments singletons de l'espace des classes Ω. Nous référerons ce cas sous le nom de prédictions déterminées. L'option de rejet introduit la possibilité de s'abstenir de prédire, ce qui revient en réalité à prédire l'espace des classes Ω tout entier. Ainsi, il nous semble logique de considérer maintenant des sous-ensembles de Ω, $\hat{Y} \in 2^{\Omega} \setminus \emptyset$, comme des candidats potentiels pour la prédiction.

L'extension du cadre théorique est relativement directe. En considérant que nous avons maintenant $\mathcal{Y} = 2^{\Omega} \setminus \emptyset$, les Équations (2.8) et (2.9) restent valables. Il nous suffit d'étendre la définition des fonctions de coûts à

$$c_{\hat{Y}} : \begin{cases} \Omega \to \mathbb{R}^+ \\ y \to c_{\hat{Y}}(y) = c_{f(\mathbf{x})}(y) \end{cases}, \tag{3.1}$$

où $c_{\hat{Y}}(y)$ est le coût de prédire le sous-ensemble $\hat{Y}$ quand y est la classe observée. Ceci signifie que la matrice de coûts correspondante n'est plus une matrice carrée. Le nombre de lignes est désormais égal à $2^{|\Omega|} - 1$. Il

est également important de noter que l'espace des classes ne change pas, et reste Ω.

Exemple 2. *A nouveau, nous considérons l'Exemple 1 pour illustrer l'effet de cette extension de l'espace de prédictions. Maintenant, il nous est possible de prédire que l'obstacle peut être "humain ou bicyclette" ($\{h, b\}$), "bicyclette ou aucun obstacle" ($\{b, n\}$), "humain ou aucun obstacle" ($\{h, n\}$) ou même "totalement incertain" ($\{h, b, n\}$). Le Tableau 3.1 illustre la matrice de coûts étendue.*

	vérité		
$c_{\hat{Y}}(y)$	$y = h$	$y = b$	$y = n$
$\hat{Y} = \{h\}$	0	1	2
$\hat{Y} = \{b\}$	1	0	2
$\hat{Y} = \{n\}$	4	4	0
$\hat{Y} = \{h, b\}$	$c_{\{h,b\}}(h)$	$c_{\{h,b\}}(b)$	$c_{\{h,b\}}(n)$
$\hat{Y} = \{b, n\}$	$c_{\{b,n\}}(h)$	$c_{\{b,n\}}(b)$	$c_{\{b,n\}}(n)$
$\hat{Y} = \{h, n\}$	$c_{\{h,n\}}(h)$	$c_{\{h,n\}}(b)$	$c_{\{h,n\}}(n)$
$\hat{Y} = \{h, b, n\}$	$c_{\{h,b,n\}}(h)$	$c_{\{h,b,n\}}(b)$	$c_{\{h,b,n\}}(n)$

TABLE 3.1: Matrice de coûts étendue à $\mathcal{Y} = 2^{\Omega}$

Pour certaines applications, il est possible d'inférer ces coûts pour les prédictions sous forme d'ensemble, soit de par la nature même des prédictions (e.g., le coût d'utilisation d'un classifieur alternatif), soit en faisant appel à un expert.

Une fois la matrice de coûts étendue, il est possible de produire et d'évaluer les prédictions sous forme d'ensemble (que nous appellerons prédictions indéterminées) en résolvant le problème de classifieur optimal dans les Équations (2.8) et (2.9), et celui de la minimisation du risque empirique tel que décrit dans l'Équation (2.10).

Concrètement, pour résoudre les Équations (2.8) et (2.9), nous calculons le coût espéré, $\mathbb{E}[c_{\hat{Y}}]$, pour chaque prédiction $\hat{Y} \in \mathcal{Y}$:

$$\mathbb{E}[c_{\hat{Y}}] = \sum_{y \in \Omega} p(y|\mathbf{x}) c_{\hat{Y}}(y). \tag{3.2}$$

Ainsi, résoudre l'Équation (2.8) revient à chercher la prédiction (optimale) $\hat{Y}^*$ qui a le plus faible coût espéré :

$$\hat{Y}^* = \arg\min_{\hat{Y}\in\mathcal{Y}} \mathbb{E}[c_{\hat{Y}}] = \arg\min_{\hat{Y}\in\mathcal{Y}} \sum_{y\in\Omega} p(y|\mathbf{x}) c_{\hat{Y}}(y). \tag{3.3}$$

Cependant, il est important de noter que le nombre de prédictions possibles devient une fonction exponentielle du nombre de classes suite à cette extension. Le temps de calcul devient très vite prohibitif si cette méthode est utilisée directement pour la phase de prédiction quand $K = |\Omega|$ est élevé. Il reste néanmoins valide dans le cadre de l'évaluation et de la comparaison de classifieurs (capables de produire des prédictions indéterminées). Pour utiliser cette formalisation dans la pratique, les méthodes de l'état de l'art se basent souvent sur des techniques qui consistent à limiter les prédictions *plausibles* et à réduire ainsi l'espace des prédictions à un sous-ensemble de 2^Ω.

Cas pratique

Nous examinons maintenant quelques classifieurs de l'état de l'art qui sont capables de donner des prédictions indéterminées (nous les catégoriserons sous le nom de *classifieurs indéterminés*). Nous nous concentrerons sur la manière dont ils résolvent le problème de la taille exponentielle de l'espace de prédictions. Nous nous limitons ici aux problèmes avec des coûts d'erreur de classification unitaires, et aborderons les extensions aux coûts génériques au chapitre suivant.

Le rejet partiel est une méthode adaptée de la méthode de rejet qui consiste à rejeter les classes de manière sélective. Il s'agit de ne rejeter que les classes qui semblent non plausibles et de prédire l'ensemble des classes non rejetées. Comme l'illustre la Figure 3.4, nous partitionnons l'espace de rejet tel que défini sur la Figure 3.2 en plusieurs régions correspondant aux divers sous-ensembles de 2^Ω.

Les divers classifieurs de rejet partiel de l'état de l'art diffèrent sur la méthode utilisée pour la partition. Cependant la plupart consiste à ordonner

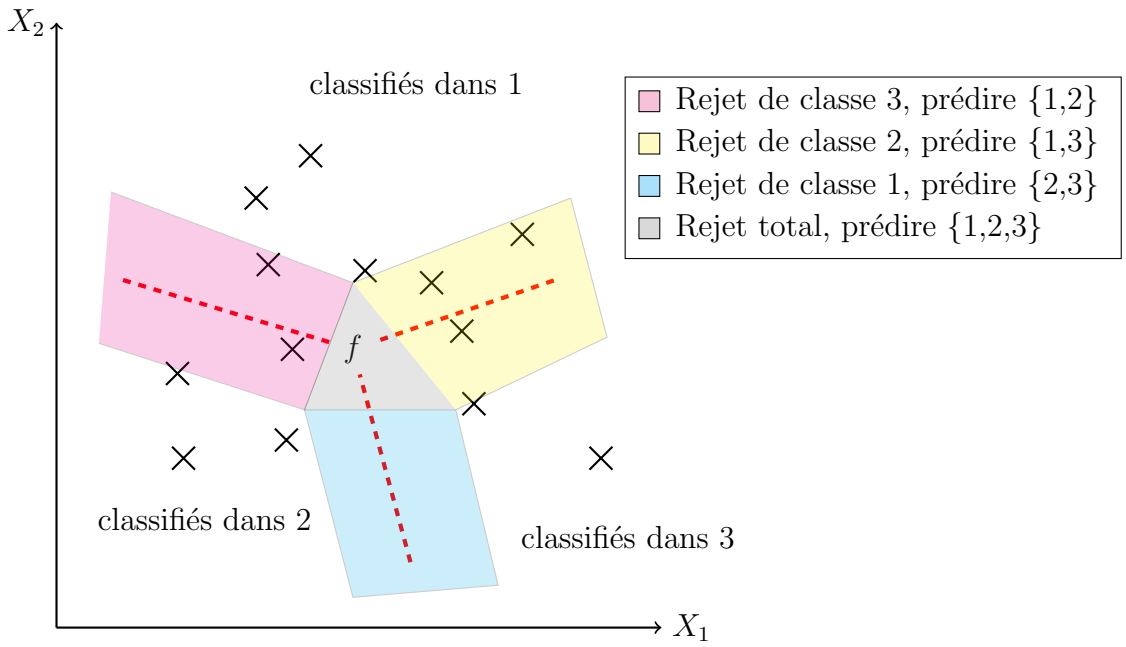


FIGURE 3.4: Exemple de rejet partiel

les classes selon leurs probabilités *a posteriori*. Ensuite diverses techniques sont utilisées pour sélectionner l'ensemble de classes les plus plausibles. Nous pouvons notamment citer la technique simpliste de "top rang n" qui consiste à toujours sélectionner les n (paramètre à fixer) premières classes dans l'ordre décroissant de leurs probabilités *a posteriori*. Elle conduit à des prédictions toujours indéterminées. Une technique plus populaire est celle dite de "risque constant" (Gupta, 1965), qui consiste à sélectionner autant de classes que nécessaires (toujours dans l'ordre décroissant de probabilités *a posteriori*) afin que la somme des probabilités *a posteriori* dépassent un seuil de risque fixé. Vovk et collab. (2005) ont développé ce qu'ils appellent la "prédiction conforme" (*conformal prediction*) en utilisant une idée similaire dans un cadre en ligne où les données arrivent sous forme d'un flux.

Nous observons que l'astuce utilisée par ces méthodes est de réduire l'espace de prédictions en ne gardant que les prédictions conformes à l'ordre décroissant des probabilités *a posteriori* établi. Nous avons ainsi $(\omega_1, \ldots, \omega_K)$

où, si $i < j \in [1; K]$ alors $p(\omega_i) > p(\omega_j)$. Cet ordre fait que la complexité de la prédiction devient linéaire par rapport au nombre de classes. Dans le cas des coûts 0/1, il semble effectivement raisonnable d'exclure les classes ayant des probabilités *a posteriori* faibles. Cependant, il est évident que cette procédure n'est plus applicable dans le cas des coûts génériques.

del Coz et collab. (2009) proposent de produire des prédictions indéterminées en appliquant la mesure F_β pour la sélection de l'ensemble de prédictions. L'idée essentielle ici est que, certes les prédictions indéterminées augmentent le pouvoir prédictif (gain de justesse de prédiction), mais elles sont aussi moins informatives que les prédictions déterminées (la question de comment discriminer les classes prédites reste en suspens). Il est donc nécessaire de balancer l'apport des prédictions indéterminées en les pénalisant en fonction du nombre de classes prédites.

Leur proposition résulte de l'utilisation de la formule (*cf.* Section 4.1.2 du Chapitre 4 pour plus de détail) :

$$F_\beta(\hat{Y}, \omega) = \frac{1 + \beta^2}{\beta^2 + |\hat{Y}|} \times \mathbb{1}_{\omega \in \hat{Y}},$$

où le paramètre β est fixé à 1 dans la plupart des cas (on parle alors de mesure F_1). Cette mesure peut être utilisée pour spécifier la fonction de coûts :

$$c_{\hat{Y}} : \begin{cases} \Omega \to \mathbb{R}^+ \\ y \to 1 - F_\beta(\hat{Y}, y). \end{cases} \tag{3.4}$$

Ainsi, cette méthode consiste aussi à choisir la prédiction $\hat{Y}$ en résolvant l'Équation (2.8).

del Coz et collab. (2009) montre que pour une prédiction $\hat{Y}_r$ composée de r classes, le coût espéré est :

$$\Delta_r = \mathbb{E}[c_{\hat{Y}_r}] = 1 - \frac{1 + \beta^2}{\beta^2 + r} \sum_{\omega \in \hat{Y}_r} p(\omega). \tag{3.5}$$

Cependant, pour dériver un algorithme efficace, la même astuce que dans le cas de rejet partiel pour la réduction de l'espace de prédictions est utilisée. Il est nécessaire d'établir un ordre de probabilités *a posteriori*

et de sélectionner les r classes ayant les probabilités *a posteriori* les plus élevées. De plus, comme la notion de mesure F_β ne tient pas compte des coûts d'erreur de classification, cette méthode est réservée au cadre des coûts unitaires.

L'algorithme mis en place en utilisant cette astuce de réduction consiste à calculer la séquence de valeurs de Δ_r (en partant de $r = 1$ et en l'incrémentant) et à retenir la prédiction $\hat{Y}_r$ minimisant cette séquence. Nous notons également qu'il n'est pas nécessaire de calculer toute la séquence Δ_r pour toutes les valeurs de r, puisque l'algorithme s'arrête dès qu'une augmentation de valeur de Δ_r est constatée.

En somme, nous voyons qu'il est difficile de mettre en place des modèles prudents capables de produire des prédictions indéterminées dans la pratique quand les coûts d'erreur de classification ne sont pas unitaires. De plus, même dans le cas de coûts unitaires, il est impossible de garantir la qualité des estimations des probabilités, puisque ces modèles ne tiennent pas compte du problème de manque d'information évoquée sur la Figure 3.1.

3.2 Probabilités imprécises

L'idée de produire des prédictions indéterminées étant d'augmenter la prudence et la fiabilité des prédictions, il est alors logique de considérer ces dernières non seulement lors de l'étape de prise de décision, mais également lors de l'apprentissage du modèle et de la représentation des estimations de probabilités. C'est l'approche que nous allons maintenant adopter en utilisant le cadre des probabilités imprécises.

3.2.1 Cadre théorique

Dans le cadre des probabilités imprécises formalisées par Walley (1991), le problème d'estimation de distribution de probabilité tel que décrit dans les Équations (2.2) et (2.3) est ré-examiné sous un angle ensembliste : l'estimation est désormais un ensemble (convexe) $\mathbf{P}$ de distributions de probabilités possibles, au lieu d'une distribution unique dans le cadre des probabilités

classiques, c'est-à-dire

$$\mathbf{P} \subseteq \mathcal{P} \tag{3.6}$$

où $\mathcal{P}$ est l'espace de toutes les distributions de probabilités sur les variables $(\mathbf{X}, Y)$.

Cet ensemble, aussi appelé *ensemble crédal* (Levi, 1983), représente l'incertitude sur la distribution de probabilité elle-même lorsqu'elle ne peut pas être identifiée avec certitude (en raison du nombre limité des données, de biais ou de bruits dans les données, ...).

Nous définissons également les ensembles crédaux locaux : $\mathbf{P}_X$ un ensemble défini pour la variable X, et $\mathbf{P}_X^y$ un ensemble de probabilités sur X conditionnelles à $Y = y$. Tout comme dans le cas des probabilités standards où nous pouvons représenter une distribution jointe $p(\mathbf{x}, y)$ par une combinaison de distributions conditionnelles (portant par exemple sur un seul variable X dans le cas du NBC à l'Equation (2.12)), ici il est possible d'inférer l'ensemble crédal portant sur les variables $(\mathbf{X}, Y)$ par des ensembles crédaux locaux $\mathbf{P}_X^y$. Il existe notamment deux manières principales pour représenter un ensemble crédal quelconque $\mathbf{P}$.

La première est d'utiliser la combinatoire. Nous appelons une *distribution extrême* de l'ensemble crédal une distribution qui ne peut pas être réécrite comme une combinaison convexe d'autres éléments de l'ensemble crédal. Nous dénotons l'ensemble des distributions extrêmes par $ext(\mathbf{P})$. Ainsi, un ensemble crédal fini peut être représenté exactement par ses distributions extrêmes. Nous parlerons alors de représentation par les sommets. Nous donnons une illustration de ce cas dans l'Exemple 3 où l'ensemble crédal est représenté sous la forme d'un polytope dans le simplexe de probabilités.

La deuxième manière est de représenter un ensemble crédal fini par un ensemble fini de contraintes linéaires (Cozman, 2000) où la j-ième contrainte peut être écrite comme :

$$\sum_i \alpha_{ij} p(x_i) < \gamma_j,$$

où $x_i \in X_i$, $\alpha_{ij} \in \mathbb{R}$ et $\gamma_j \in \mathbb{R}$.

Typiquement, les informations sur les probabilités *a posteriori* dans l'Exemple 3 peuvent aussi être vues comme des inégalités sur les bornes des probabilités. Également, les contraintes usuelles de non-négativité et de normalisation du cadre de probabilités standards sont toujours présentes. Nous parlerons dans ce cas de représentation par les contraintes.

Ces deux représentations conduisent à des méthodes différentes pour la manipulation des ensembles crédaux. La représentation par les sommets implique souvent des techniques combinatoires (énumération des sommets), alors que celle par les contraintes nécessite de résoudre des problèmes de l'optimisation sous contraintes.

Exemple 3. *Pour illustrer le cadre des probabilités imprécises, nous considérons les changements du modèle probabiliste sur l'exemple de la reconnaissance d'obstacles dans l'Exemple 1. Un classifieur utilisant le cadre de probabilités classiques produit des estimations de probabilités (conditionnelles sachant une observation* $\mathbf{x}$*) précises telles que :*

$$p(h) = 0.1 \qquad p(b) = 0.3 \qquad p(n) = 0.6$$

Nous pouvons représenter cette distribution dans un espace à trois dimension $(p(h), p(b), p(n))$ *sous forme d'un point, comme le montre la Figure 3.5.*

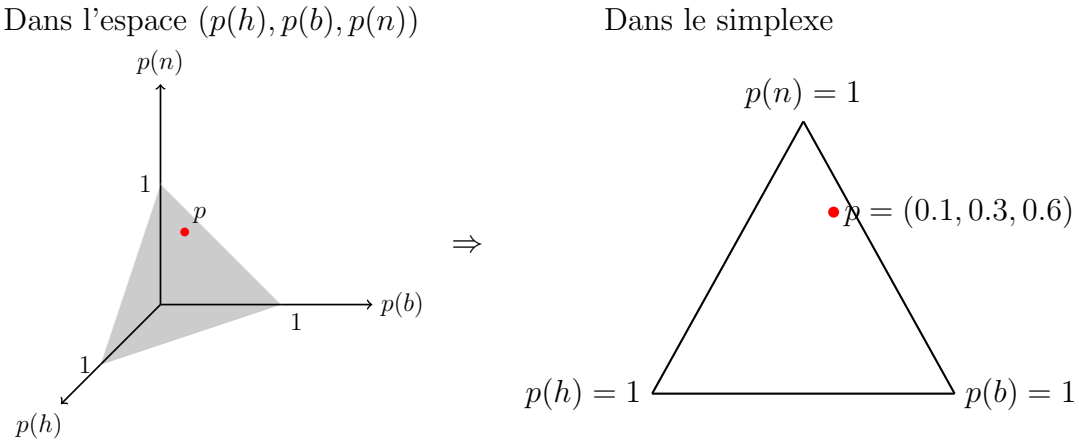


FIGURE 3.5: Une distribution de probabilités précises pour un problème à trois classes.

En effet, toutes les distributions de probabilités possibles doivent satisfaire la contrainte de normalisation

$$\sum_{y \in \Omega} p(y) = 1,$$

et la contrainte de non-négativité

$$\forall y \in \Omega, \qquad p(y) > 0$$

de ce fait, ces distributions se trouvent toutes sur le triangle unité représenté à la Figure 3.5, qui est connu sous le nom de simplexe de probabilités.

Dans le cadre des probabilités imprécises, un classifieur probabiliste imprécis produira par exemple des estimations de probabilités sous forme d'intervalles :

$$p(h) = [0; 0.2] \qquad p(b) = [0.3; 0.4] \qquad p(n) = [0.4; 0.6].$$

Ce sont les probabilités marginalisées sur les classes. La largeur des intervalles représente l'incertitude sur les estimations. Si l'intervalle est large alors cela signifie que la qualité de l'estimation est mauvaise due au manque d'informations sur la classe en question. Dans le cas contraire, le classifieur peut très bien produire des estimations "précises" (ou des intervalles très étroits) s'il y a assez d'informations/données. A noter qu'il est aisé de transformer ces intervalles de probabilités pour la représentation par les contraintes, nous aurons :

$$\begin{cases} p(h) \leq 0.2, \\ p(b) \leq 0.4, \ -p(b) \leq -0.3, \\ p(n) \leq 0.6, \ -p(n) \leq -0.4. \end{cases}$$

L'ensemble crédal $\mathbf{P}$ *est ici l'ensemble de toutes les distributions dont les probabilités précises marginalisées* $(p(h), p(b), (p(n))$ *se situent à l'intérieur des intervalles spécifiés. Nous pouvons également le représenter dans le simplexe, comme l'illustre la Figure 3.6. Ici,* $\mathbf{P}$ *est un polytope défini par son enveloppe convexe caractérisée par ses quatre sommets dans un espace à trois dimensions :*

$$ext(\mathbf{P}) = \{(0, 0.4, 0.6); (0.2, 0.3, 0.5); (0.2, 0.4; 0.4); (0.1, 0.3, 0.6)\}$$

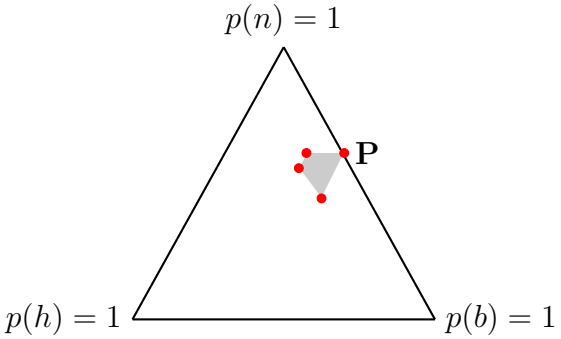


FIGURE 3.6: Un ensemble crédal déduit à partir des probabilités imprécises marginalisées.

3.2.2 Les critères de décision

Le cadre des probabilités imprécises ne change pas fondamentalement l'étape de prise de décision, qui est toujours basée sur la minimisation de coûts espérés mentionnés dans les Équations (2.8) et (3.3). L'ensemble des coûts espérés sur $\mathbf{P}$ peut être représenté par ses bornes inférieure et supérieure $\underline{\mathbb{E}}, \overline{\mathbb{E}}$. Étant donné une prédiction $\hat{y}$ et la fonction de coûts associée $c_{\hat{Y}}$, nous avons :

$$\underline{\mathbb{E}}[c_{\hat{y}}] = \min_{p \in \mathbf{P}} \mathbb{E}[c_{\hat{y}}] = \min_{p \in \mathbf{P}} \sum_{y \in \Omega} p(y|\mathbf{x}) c_{\hat{y}}(y), \tag{3.7}$$

$$\overline{\mathbb{E}}[c_{\hat{y}}] = \max_{p \in \mathbf{P}} \mathbb{E}[c_{\hat{y}}] = \max_{p \in \mathbf{P}} \sum_{y \in \Omega} p(y|\mathbf{x}) c_{\hat{y}}(y). \tag{3.8}$$

La borne supérieure du coût espéré $\overline{\mathbb{E}}$ est obtenue en remplaçant le min par le max car nous avons la dualité $\underline{\mathbb{E}}(c) = -\overline{\mathbb{E}}(-c)$. Nous utiliserons également cette notation pour représenter les bornes inférieures et supérieures des distributions de probabilités, qui seront notées respectivement $\underline{p}$ et $\overline{p}$:

$$\underline{p}(y) = \min_{p \in \mathbf{P}} p(y), \qquad \overline{p}(y) = \max_{p \in \mathbf{P}} p(y). \tag{3.9}$$

Ainsi, le changement par rapport au cas des probabilités précises est qu'il faut désormais considérer le problème de minimisation de coûts espérés énoncé dans l'Équation (3.3), dans le cadre des probabilités imprécises. Divers critères de décisions dérivés de l'Équation (3.3) sont proposés dans

la littérature (Troffaes, 2007). Certains visent à comparer ces ensembles de coûts espérés afin d'aboutir (presque toujours) à une prédiction déterminée, parmi cette catégorie de critères, nous pouvons notamment citer

- le *maximin* qui consiste tout simplement à remplacer $\mathbb{E}$ par $\underline{\mathbb{E}}$ dans l'Équation (3.3). Ceci revient à optimiser la prédiction dans le cas le plus pessimiste vis-à-vis des estimations de probabilités. Une prédiction indéterminée n'est produite qu'en cas extrême de coûts espérés identiques.
- le *maximax* qui au contraire remplace $\mathbb{E}$ par $\overline{\mathbb{E}}$ et considère l'optimisation dans le cas le plus optimiste.

Comme nous sommes intéressés par l'obtention de modèles prudents, nous écartons donc ces critères. Nous nous intéressons aux critères capables de prédire des ensembles de classes. Ces critères permet d'établir un ordre **partiel** $\succ$ sur les prédictions $\hat{y} \in \Omega$:

Définition 1 (Maximalité)**.**

$$\hat{y}_i \succ_{\mathcal{M}} \hat{y}_j \Leftrightarrow \underline{\mathbb{E}}[c_{\hat{y}_j} - c_{\hat{y}_i}] > 0. \tag{3.10}$$

L'Équation (3.10) peut être interprétée de la manière suivante : prédire $\hat{y}_i$ est préférable à $\hat{y}_j$ si le fait d'échanger $\hat{y}_i$ pour $\hat{y}_j$ entraîne toujours un coût positif quelle que soit la distribution de probabilités donnée. Nous notons également que déterminer cet ordre pour l'ensemble des classes requiert au pire $K(K-1)$ comparaisons ($K = |\Omega|$), une pour chaque paire de classes.

Définition 2 (Dominance par intervalle)**.**

$$\hat{y}_i \succ_{\mathcal{ID}} \hat{y}_j \Leftrightarrow \overline{\mathbb{E}}[c_{\hat{y}_i}] < \underline{\mathbb{E}}[c_{\hat{y}_j}]. \tag{3.11}$$

Dans ce cas, $\hat{y}_i$ est préféré à $\hat{y}_j$, si le coût le plus élevé auquel on s'expose en prédisant $\hat{y}_i$ reste inférieur au plus faible coût que nous puissions espérer en prédisant $\hat{y}_j$. Nous notons qu'ici, il se peut que deux distributions de probabilités différentes soient utilisées pour calculer séparément $\overline{\mathbb{E}}[c_{\hat{y}_i}]$ et $\underline{\mathbb{E}}[c_{\hat{y}_j}]$, contrairement à l'Équation (3.10). Ce critère est donc plus conservatif. En revanche, il ne requiert que $2K$ comparaisons dans le pire des cas.

Dans les deux cas, il est important de noter que l'ordre établi est partiel (contrairement au cas de l'Équation (3.3) où l'ordre est total), ce qui fait que les prédictions peuvent être indéterminées. Et il n'est pas nécessaire d'expliciter entièrement cet ordre partiel pour réaliser des prédictions, car il s'agit plutôt d'identifier les éléments non-dominés par cet ordre (un élément dominé est immédiatement exclus) :

$$\hat{Y} = \left\{ \hat{y}_i \in \Omega \mid \nexists \hat{y}_j : \hat{y}_j \succ \hat{y}_i \right\},$$

où $\succ$ réfère soit à $\succ_{\mathcal{M}}$, soit à $\succ_{\mathcal{ID}}$.

Nous notons aussi que, dans le cas des coûts 0/1, ces critères reviennent à comparer les intervalles de probabilités *a posteriori* :

$$\hat{y}_i \succ_{\mathcal{M}} \hat{y}_j \Leftrightarrow \min_{p \in \mathbf{P}} p(\hat{y}_i) - p(\hat{y}_j) > 0. \tag{3.12}$$

$$\hat{y}_i \succ_{\mathcal{ID}} \hat{y}_j \Leftrightarrow \min_{p \in \mathbf{P}} p(\hat{y}_i) > \max_{p' \in \mathbf{P}} p'(\hat{y}_j). \tag{3.13}$$

Exemple 4. *Etant donné les intervalles de probabilités dans l'Exemple 3 et les coûts dans l'Exemple 1, wnous pouvons maintenant calculer les coûts espérés et ainsi inférer des décisions selon les deux critères que nous venons de spécifier.*

Pour la dominance par intervalles, nous avons seulement besoin de calculer les coûts espérés des trois prédictions déterminées possibles sur les sommets de l'ensemble crédal $\mathbf{P}$ *et ensuite regarder si les intervalles obtenus sont disjoints :*

$$\begin{aligned} \mathbb{E}[c_h] &= p(b) + 2p(n) \in [1.2; 1.6] \\ \mathbb{E}[c_b] &= p(h) + 2p(n) \in [1; 1.3] \\ \mathbb{E}[c_n] &= 4p(h) + 4p(b) \in [1.6; 2.4] \end{aligned}$$

Le coût espéré de n *est nettement plus élevé que celui de* b*, donc nous avons la préférence* $b \succ_{\mathcal{ID}} n$*, par conséquent la prédiction* n *peut être exclue. Mais nous ne pouvons plus déterminer d'autres relation de préférence avec ce critère de décision, car les intervalles de coûts espérés de* h *et de* b *ne sont pas disjoints. Ainsi, la prédiction finale est l'ensemble* $\{h, b\}$ *en utilisant ce critère.*

Avec le critère de maximalité, nous pouvons obtenir $b \succ_{\mathcal{M}} n$ *en calculant* $\underline{\mathbb{E}}[c_n - c_b]$. $\underline{\mathbb{E}}$ *est obtenu en minimisant le coût espéré sur tous les sommets de l'ensemble crédal* $\mathbf{P}$ *donnés dans l'Exemple 3 :*

$$\underline{\mathbb{E}}[c_n - c_b] = \min\Big(3p(h) + 4p(b) - 2p(n)\Big) = 3 * 0.1 + 4 * 0.3 - 2 * 0.6 = 0.3.$$

De plus, nous obtenons également

$$\underline{\mathbb{E}}[c_h - c_b] = \min\Big(- p(h) + p(b)\Big) = -0.2 + 0.3 = 0.1.$$

Donc il y a aussi une préférence $b \succ_{\mathcal{M}} h$*, et la prédiction finale est seulement le singleton* $\{b\}$ *avec ce critère de décision.*

En résumé, nous voyons que l'utilisation des estimations d'une forme plus complexe (intervalles) dans le cadre des probabilités imprécises, nous a permis de simplifier la prise de décision pour produire des prédictions indéterminées par rapport aux probabilités standards : il n'est plus nécessaire d'énumérer tous les éléments de 2^{Ω}. Nous faisons alors une distinction entre les classifieurs dits imprécis, qui utilisent le cadre des probabilités imprécises dans l'apprentissage pour produire des prédictions indéterminées, et les classifieurs dits indéterminés pour désigner tout classifieur capable de produire des prédictions indéterminées.

Un avantage évident (parmi d'autres que nous verrons plus tard) des classifieurs imprécis est que les estimations sous forme d'intervalles permettent de modéliser clairement la notion de manque d'information évoquée au début du chapitre et, par conséquent, de quantifier la fiabilité des estimations en fonction des informations disponibles. Cela peut aider à remédier au problème de qualité d'estimations de probabilités qui est crucial pour l'apprentissage sensible aux coûts. La difficulté sous-jacente réside dans la complexité de la manipulation de l'ensemble crédal en pratique.

3.3 Détails pratiques sur l'estimation et la classification

Nous allons voir maintenant comment les estimations sous forme d'intervalles de probabilités conditionnelles peuvent être inférés. Nous évoquerons ensuite plusieurs approches considérées dans l'état de l'art pour l'inférence des prédictions. L'objectif n'est donc pas d'étudier en détail le fonctionnement de chacune, mais de présenter brièvement les spécificités de chaque classifieur.

3.3.1 Estimations des ensembles crédaux locaux avec le modèle imprécis de Dirichlet

Une des méthodes la plus utilisée dans l'état de l'art pour inférer des probabilités conditionnelles sous forme d'intervalles est le modèle de Dirichlet imprécis (MDI) (Walley, 1996; Bernard, 2005). Nous introduisons ici son utilisation pour inférer les ensembles crédaux locaux.

Supposons que nous disposons d'une série de tirages i.i.d. de valeurs de la variable X, que nous notons $\mathcal{T}_X$. Si nous voulons inférer la probabilité conditionnelle $\theta_x = p(X = x)$, une approche Bayésienne communément utilisée consiste à supposer une distribution de Dirichlet *a priori* pour le vecteur paramètre $\boldsymbol{\theta}$ et d'estimer $\mathbb{E}[\theta_x|x]$ l'espérance *a posteriori* du paramètre θ_x pour chaque $x \in X$. Le modèle de Dirichlet dépend d'un paramètre $s \in \mathbb{R}^+$, représentant intuitivement la force de la loi *a priori*, et d'un vecteur $\mathbf{t} = (t_x)_{x \in X}$, représentant l'apriori sur les fréquences, tel que $\sum_{x \in X} t_x = 1$. Il permet d'inférer les probabilités conditionnelles de manière simple :

$$p(X = x|\mathcal{T}_X) = \mathbb{E}[\theta_x|x] = \frac{occ_x + s \cdot t_x}{occ_X + s}, \tag{3.14}$$

où occ_x est le nombre d'instances de données où $X = x$, et occ_X est ici le nombre total de tirages que nous disposons.

Dans le cas du modèle de Dirichlet imprécis, le modèle ne dépend plus que du paramètre s et suppose que toutes les valeurs de $\mathbf{t}$ sont possibles (sous la contrainte $\sum_{x \in D_X} t_x = 1$). Cela signifie dans le cas d'une unique

variable X que $t_x \in [0;1]$ pour tout $x \in D_X$. Ceci nous permet alors de définir un ensemble de distributions possibles (un ensemble crédal) pour toute variable X,

$$p(X = x|\mathcal{T}_X) \in \left[\frac{occ_x}{occ_X + s}, \frac{occ_x + s}{occ_X + s}\right]. \tag{3.15}$$

En particulier, pour la variable de sortie Y, nous avons

$$p(Y = y) = \left[\frac{occ_y}{occ_Y + s}, \frac{occ_y + s}{occ_Y + s}\right], \tag{3.16}$$

où occ_Y est ici la taille des données d'apprentissage.

Et par conséquent, pour une classe $y \in \Omega$ donnée, les probabilités conditionnelles sont de la forme :

$$p(X = x|Y = y) = \left[\frac{occ_{x,y}}{occ_y + s}, \frac{occ_{x,y} + s}{occ_y + s}\right]. \tag{3.17}$$

Exemple 5. *En reprenant le graphique de la Figure 3.1, nous pouvons utiliser le MDI pour déterminer les probabilités dans les zones 1 et 2.*

Fixons $s = 1$, comme nous avons beaucoup d'instances de données dans la zone 1, nous obtenons des estimations similaires pour les deux classes

$$p(\square|1) \in \left[\frac{5}{10+1}; \frac{5+1}{10+1}\right] \text{ et } p(\blacksquare|1) = 1 - p(\square|1).$$

Par contre, dans la zone 2, nous constatons des différences majeures, nous avons

$$p(\square|1) \in \left[\frac{0}{1+1}; \frac{0+1}{1+1}\right] = \left[0; \frac{1}{2}\right]$$

tandis que,

$$p(\blacksquare|1) \in \left[\frac{1}{1+1}; \frac{1+1}{1+1}\right] = \left[\frac{1}{2}; 1\right].$$

Nous constatons également que, plus s est grand, plus il faut de données pour avoir des intervalles étroits, i.e., plus le modèle est prudent.

Nous notons que ces intervalles de probabilités permettent de définir les ensembles crédaux locaux $\mathbf{P}_Y$ et $\mathbf{P}_X^y$. Ainsi ce modèle est aussi appelé MDI local, par opposition à la version globale proposée par Walley (1996), qui est appliqué conjointement aux variables d'entrée et à la classe.

3.3.2 Méthodes de l'état de l'art

Nous avons vu que le cadre théorique des probabilités imprécises est en réalité une extension du cadre des probabilités standards. Ainsi, il n'est pas étonnant de constater que les approches principales pour construire un classifieur imprécis dérivent des approches que nous avons déjà vues dans le Chapitre 2. Nous rappelons ici trois extensions des approches évoquées dans la Section 2.2 et examinerons en détail celle du NBC, utilisé par la suite.

Les arbres crédaux de décisions font référence aux arbres de décisions qui considèrent les probabilités imprécises lors de l'étape de la division des branches de l'arbre. Il s'agit d'intégrer les estimations sous forme d'ensembles crédaux dans le critère de division. Abellán et Moral (2005) proposent par exemple d'utiliser l'entropie maximale que nous pouvons obtenir avec une distribution de l'ensemble crédal comme critère de division de l'arbre. Une version sensible aux coûts est également proposée dans (Abellán et Masegosa, 2012).

Les réseaux crédaux sont des extensions des réseaux Bayésiens (Cozman, 2000). Il s'agit de remplacer les probabilités conditionnelles sur les arcs du graphe par des ensembles crédaux locaux $\mathbf{P}_X^{pa(X)}$ (où $pa(X)$ représente les nœuds parents du nœud X). L'inférence dans un tel réseau crédal est en général difficile (Mauá et collab., 2014). Nous étudierons plus en détail les problèmes liés dans le cas du classifieur crédal naïf.

3.3.3 Le cas du classifieur crédal naïf

Le classifieur crédal naïf (NCC) est une extension du classifieur Bayésien naïf dans le cadre des probabilités imprécises (Zaffalon, 2002). Par conséquent, il rentre aussi dans la catégorie de réseaux crédaux. Il préserve l'Hypothèse 1 d'indépendance des attributs du NBC et la règle de Bayes reste applicable, donc l'Équation (2.12) reste valable pour toute distribution

p de l'ensemble crédal. Ainsi, nous avons :

$$p(y|x_1,\dots,x_m) = \frac{p(y)\prod_{i=1}^{m} p(x_i \mid y)}{\sum_{y'\in\Omega} p(y')\prod_{i=1}^{m} p(x_i \mid y')} \tag{3.18}$$

$$= \left(1 + \frac{\sum_{y'\in\Omega, y'\neq y} p(y')\prod_{i=1}^{m} p(x_i \mid y')}{p(y)\prod_{i=1}^{m} p(x_i \mid y)}\right)^{-1}. \tag{3.19}$$

Nous pouvons passer de (3.18) à (3.19) en supposant que les $p(x_i|y)$ sont non nuls. Ainsi, pour trouver la borne inférieure (respectivement, supérieure) des probabilités *a posteriori*, le NCC résout le problème de minimisation (maximisation) suivant sur les ensembles crédaux locaux :

$$\min_{p(y)\in\mathbf{P}_\Omega} \min_{\substack{p(x_i|y)\in\mathbf{P}^y_{X_i} \\ i\in[1;m]}} \left(1 + \frac{\sum_{y'\in\Omega, y'\neq y} p(y')\prod_{i=1}^{m} p(x_i \mid y')}{p(y)\prod_{i=1}^{m} p(x_i \mid y)}\right)^{-1}, \tag{3.20}$$

Nous remarquons que, grâce à la transformation de (3.18) en (3.19), le numérateur et le dénominateur ne partagent aucun terme $p(x_i|y)$ commun. Ainsi, nous pouvons les optimiser séparément. De plus, comme $x \to \frac{1}{1-x}$ est une fonction monotone et décroissante sur $[0;1]$, par conséquent le problème de minimisation revient alors à maximiser le terme à l'intérieur de la parenthèse, c'est à dire à maximiser le numérateur et minimiser le dénominateur. Comme les $p(y)$ sont positives pour tout $y \in \Omega$, minimiser le dénominateur revient alors à minimiser le produit $\prod_{i=1}^{m} p(x_i \mid y)$, et de même, maximiser le numérateur revient à avoir $\prod_{i=1}^{m} \overline{p}(x_i \mid y')$ au numérateur. Par conséquent, (3.20) devient

$$\underline{p}(y|x_1,\dots,x_m) = \min_{p(y)\in\mathbf{P}_\Omega} \left(1 + \frac{\sum_{y'\in\Omega}^{y'\neq y} p(y')\prod_{i=1}^{m} \overline{p}(x_i \mid y')}{p(y)\prod_{i=1}^{m} \underline{p}(x_i \mid y)}\right)^{-1}. \tag{3.21}$$

De même, nous pouvons obtenir la borne supérieure :

$$\overline{p}(y|x_1,\ldots,x_m) = \max_{p(y)\in\mathbf{P}_\Omega} \left(1 + \frac{\sum_{y'\in\Omega}^{y'\neq y} p(y') \prod_{i=1}^{m} \underline{p}(x_i \mid y')}{p(y) \prod_{i=1}^{m} \overline{p}(x_i \mid y)}\right)^{-1}. \tag{3.22}$$

En utilisant le MDI local présenté précédemment pour les calculs de $p(x_i|y)$ et de $p(y)$, ces probabilités *a posteriori* permettent alors d'inférer efficacement des décisions avec le critère de dominance par intervalles et avec coûts unitaires, car les Equations (3.21) et (3.22) sont des fonctions fractionnaires linéaires [3] une fois que les probabilités conditionnelles inférieure et supérieure sont déterminées (grâce à la transformation (3.19)).

Dans le cas du critère de maximalité, où nous cherchons à trouver une relation de préférence entre y_h et y_l, il s'agit de résoudre

$$\min_{p\in\mathbf{P}} p(y_h|x_1,\ldots,x_m) - p(y_l|x_1,\ldots,x_m).$$

Nous pouvons toujours expliciter ces probabilités à l'aide de l'Équation (3.18). En factorisant le dénominateur commun et en suivant la même démarche qui nous a mené à l'Équation (3.21), nous déduisons que ceci revient à résoudre le problème de minimisation suivant sur les ensembles crédaux locaux :

$$\min_{p(y)\in\mathbf{P}_\Omega} p(y_h) \prod_{i=l}^{m} \underline{p}(x_i \mid y_h) - p(y_l) \prod_{i=1}^{m} \overline{p}(x_i \mid y_l).$$

Ainsi, l'inférence des décisions ne pose pas de problème avec le critère de maximalité en coûts unitaires non plus. Il est possible d'obtenir un algorithme prédictif efficace en temps polynomial avec un programme linéaire.

Cependant, cette simplicité n'est valable que dans le cas des coûts 0/1. En effet, dans le cas des coûts 0/1, établir une préférence entre y_h et y_l ne nécessite que de résoudre un problème d'optimisation, mettant en jeu uniquement $p(y_h)$ et $p(y_l)$, évoqué dans les Équations (3.12) et (3.13). Dans le cas des coûts génériques, pour établir la même relation de préférence, il faut faire une optimisation sur les espérances des fonctions de coûts.

3. fonction fractionnaire dont le numérateur et le dénominateur sont des fonctions linéaires des mêmes variables.

Notons $\mathcal{C} = c_{y_h} - c_{y_l}$. Nous avons alors

$$\underline{\mathbb{E}}[\mathcal{C}] = \min_{p(y)} \min_{p(x_i|y')} \frac{\sum_{y\in\Omega} \mathcal{C}(y)p(y) \prod_{i=1}^{m} p(x_i \mid y)}{\sum_{y\in\Omega} p(y) \prod_{i=1}^{m} p(x_i \mid y)} \tag{3.23}$$

La simplification utilisée dans le cas des coûts unitaires n'est plus valable : il n'est plus possible d'éliminer les termes partagés par le dénominateur et le numérateur avec la transformation utilisée dans l'Équation (3.19), à cause du facteur $\mathcal{C}(y)$, et donc de simplifier le calcul sur les produits $\prod_{i=1}^{m} p(x_i|y)$. Ainsi, $\underline{\mathbb{E}}[\mathcal{C}]$ ne peut être transformée en une fonction fractionnelle linéaire. Par conséquent, il est beaucoup plus difficile de résoudre ce problème sans approximation ou hypothèse supplémentaire. Nous verrons dans les chapitres ultérieurs comment ce problème peut être résolu.

Conclusion

Nous avons étudié comment produire des prédictions indéterminées sans se préoccuper de la manière dont nous pouvons définir ou obtenir des fonctions de coûts associées dans la pratique. En effet, si les coûts pour les prédictions déterminées peuvent être obtenus relativement simplement, soit par des informations des experts, soit en analysant la structure de Ω (par exemple dans la classification ordinale, ou hiérarchique), il est en revanche plus difficile de déterminer des coûts pour les prédictions indéterminées. En se référant au Tableau 3.1, la problématique que nous posons ici est donc de savoir comment remplir la partie manquante du tableau de manière raisonnable. C'est à cette problématique nous allons tenter de donner des éléments de réponse dans le chapitre suivant.

Chapitre 4

Intégration des coûts génériques aux modèles prudents

Dans le chapitre précédent, nous avons détaillé comment produire des prédictions indéterminées pour les problèmes où être prudent et fiable est un critère tout aussi important que la justesse de prédiction. Cependant, la question de comment définir les coûts d'erreur de classification de ces prédictions indéterminées n'est pas encore abordée. En effet, même si l'utilisation du cadre des probabilités imprécises semble pouvoir s'abstraire de ces coûts lors de l'apprentissage, ces derniers restent primordiaux pour comparer les classifieurs (qu'ils produisent ou non des prédictions indéterminées). Une telle comparaison est essentielle pour aborder certaines problématiques, comme la définition d'un "meilleur" modèle et comment aboutir à ce modèle.

Dans le cadre des probabilités imprécises, une méthodologie raisonnable et convaincante pour permettre cette comparaison dans le cas des coûts unitaires a été élaborée par Zaffalon et collab. (2012), en se basant sur un cadre de pari. Cependant la situation est beaucoup plus complexe lorsque les coûts ne sont pas unitaires, et bien que des solutions spécifiques aient été proposés (Ha, 1997; Abellán et Masegosa, 2012), nous n'avons trouvé aucun travail dans la littérature proposant des lignes directrices génériques pour produire et comparer des prédictions indéterminées sensibles aux coûts.

Tel est l'objectif de ce chapitre, dans lequel nous adoptons un point de vue axiomatique du problème. Nous allons d'abord présenter les propositions de la littérature qui établissent des mesures d'évaluation de classifieurs permettant de comparer simultanément les classifieurs déterminés et indéterminés dans le cadre des coûts unitaires. Nous proposerons ensuite des propriétés que nous pensons être nécessaires ou souhaitables pour définir de manière générique les coûts des prédictions indéterminées. Ceci nous permettra d'examiner des propositions existantes en fonction de ces propriétés. Enfin, nous proposerons une formule générale qui permet de dériver des coûts de prédictions indéterminées de manière raisonnable à la lumière des propriétés définies.

4.1 Propositions existantes pour la comparaison des classifieurs

Nous nous intéressons ici uniquement au problème de comparaison et d'évaluation des classifieurs. Concrètement, étant donné deux classifieurs potentiellement indéterminés, il s'agit de trouver une mesure pour évaluer et comparer leur pouvoir prédictif de manière équitable. Nous avons vu que le risque empirique énoncé dans l'Équation (2.10) est une mesure communément utilisée, quand les prédictions sont toujours déterminées, pour obtenir une mesure unique d'évaluation des classifieurs. Cette mesure reste valable pour les prédictions indéterminées en étendant l'espace de prédictions à 2^{Ω}. La difficulté est dans la définition des coûts des prédictions indéterminées : comme l'espace des prédictions est différent, calculer simplement le taux de bonnes ou de mauvaises prédictions n'est plus adapté puisque cela avantage toujours injustement l'un des deux classifieurs.

Dans la suite de cette partie, nous verrons deux propositions existantes pour définir ces coûts de prédictions indéterminées. Elles partent toutes les deux de la situation où les coûts des prédictions déterminées sont unitaires, et se basent sur l'idée qu'il faut pénaliser l'indétermination et trouver un compromis entre l'informativité (le niveau d'indétermination) et la justesse des prédictions.

4.1.1 Justesse affaiblie selon l'utilité

Dans le cadre des coûts unitaires, c'est à dire quand $c_{\hat{y}}(y) = 0$ si $\hat{y} = y$ et $c_{\hat{y}}(y) = 1$ sinon, une première idée pour adapter les coûts aux prédictions indéterminées est d'utiliser ce qu'on appelle la justesse affaiblie telle que

$$c_{\hat{Y}}(y) = \begin{cases} 1 - 1/|\hat{Y}| \text{ si } y \in \hat{Y}, \\ 1 \text{ sinon} \end{cases}$$

qui consiste à pénaliser les prédictions indéterminées en fonction du nombre de classes prédites. Cette mesure est empruntée à la littérature de la classification multilabel (Tsoumakas et Vlahavas, 2007). Tandis que son utilisation

dans le cadre multilabel est justifiée, elle l'est beaucoup moins dans le cadre de la classification indéterminée (nous démontrerons ceci dans le cadre plus général des coûts génériques dans la Section 4.2, Proposition 2). En effet, elle a été fortement critiquée par Zaffalon et collab. (2012) sur le fait que la justesse affaiblie considère que prédire un ensemble $\hat{Y}$ revient à choisir aléatoirement une prédiction parmi $\hat{Y}$ du point de vue des coûts d'erreurs. Autrement dit, elle ne donne aucune valeur à la prudence de prédiction, ce qui est contradictoire à l'idée de construire un classifieur prudent et fiable.

Pour tenir compte de ce facteur de prudence, Zaffalon et collab. (2012) proposent de rajouter une fonction d'utilité g à la formule de la justesse affaiblie :

$$c_{\hat{Y}}(y) = \begin{cases} 1 - g(1/|\hat{Y}|) \text{ si } y \in \hat{Y}, \\ 1 \text{ sinon} \end{cases} \tag{4.1}$$

où g est une fonction sur $[0; 1]$ telle que $g(1/|\hat{Y}|) > 1/|\hat{Y}|$, $g(0) = 0$ et $g(1) = 1$.

Ils interprètent g comme une fonction concave modélisant l'aversion au risque (*i.e.*, l'utilité), ou la préférence à la prudence du décideur. En particulier, ils proposent d'utiliser des fonctions quadratiques pour g. Les valeurs de $g(0)$ et $g(1)$ étant fixées naturellement ($g(0) = 0, g(1) = 1$), il suffit alors de spécifier un point supplémentaire pour définir g. Par exemple, en spécifiant que $g(0.5) = 0.65$, nous obtenons la fonction d'utilité suivante (que nous référerons sous le nom de u_{65}) :

$$g(x) = -0.6x^2 + 1.6x. \tag{4.2}$$

Le Tableau 4.1 illustre la matrice de coûts obtenue dans le cas où les coûts sont unitaires dans l'Exemple 1. Les propriétés de base que nous allons proposer dans le paragraphe suivant se baseront sur une observation similaire concernant les valeurs des prédictions indéterminées. Cependant, il y aura quelques différences avec le cas des coûts unitaires, notamment parce que les coûts de différentes erreurs de classification ne seront pas identiques.

4.1.2 Mesure F_β

Une autre proposition est d'utiliser la mesure F_β bien connue dans le domaine de la recherche d'information, qui consiste à calculer la moyenne

$c_{\hat{y}}(y)$	vérité $y = h$	$y = b$	$y = n$
$\hat{Y} = \{h\}$	0	1	1
$\hat{Y} = \{b\}$	1	0	1
$\hat{Y} = \{n\}$	1	1	0
$\hat{Y} = \{h, b\}$	0.35	0.35	1
$\hat{Y} = \{b, n\}$	1	0.35	0.35
$\hat{Y} = \{h, n\}$	0.35	1	0.35
$\hat{Y} = \{h, b, n\}$	0.54	0.54	0.54

TABLE 4.1: Matrice de coûts définie en utilisant la justesse affaiblie selon l'utilité

harmonique (pondérée avec le coefficient β) entre la *précision* P et le *rappel* R :

$$F_\beta = \frac{(1 + \beta^2)PR}{\beta^2 P + R}. \tag{4.3}$$

Cette approche a été utilisé par del Coz et collab. (2009) pour évaluer les prédictions indéterminées et a été démontrée dans (Zaffalon et collab., 2012) comme étant assimilable à une autre instance spécifique de la fonction d'utilité g. La *précision* mesure combien de classes prédites dans $\hat{Y}$ sont pertinentes, et le *rappel* mesure combien de classes pertinentes sont prédites (dans le cadre de classification multiclasse, il est soit égal à 0 soit à 1). Par conséquent, pour une prédiction indéterminée $\hat{Y}$ et la vérité $y \in \Omega$, nous pouvons construire une table de contingence. Pour chaque valeur possible de la classe $\omega \in \Omega$, nous avons quatre possibilités :

$\omega \in \Omega$	$\omega = y$	$\omega \neq y$
$\omega \in \hat{Y}$	VP	FP
$\omega \notin \hat{Y}$	FN	VN

TABLE 4.2: Table de contingence

Le Tableau 4.2 exprime quatre situations : les éléments "vrais positifs" $\{\omega \in \hat{Y} : \omega = y\}$) (nous utilisons VP pour désigner le cardinal de l'ensemble) ; les "faux positifs" $\{\omega \in \hat{Y} : \omega \neq y\}$) (similairement, FP désigne le cardinal). Nous définissons également les "faux négatifs" et les "vrais néga-

tifs", où FN, VN désignent respectivement leurs cardinaux. Ainsi, $VP+FP$ donne le cardinal de l'ensemble prédit $\hat{Y}$ et $VP + FN$ donne le nombre de vérités.

Dans un problème multiclasse, il y a toujours uniquement une seule vérité y ($VP + FN = 1$), nous avons alors la précision et le rappel de $\hat{Y}$:

$$P(\hat{Y}, y) = \frac{VP}{VP + FP} = \frac{\mathbb{1}_{\hat{Y}}(y)}{|\hat{Y}|}, \qquad R(\hat{Y}, y) = \frac{VP}{VP + FN} = VP = \mathbb{1}_{\hat{Y}}(y),$$

où $\mathbb{1}_{\hat{Y}}$ est la fonction indicatrice de $\hat{Y}$ et $|\hat{Y}|$ est le cardinal of $\hat{Y}$.

Il est important de noter que F_β mesure la récompense d'une prédiction qui est la notion duale de coût. Par conséquent, le coût de la prédiction indéterminée est donné par $1 - F_\beta$ dans le cadre des coûts unitaires. Le Tableau 4.3 illustre la matrice de coûts dérivée avec la mesure F_β ($\beta = 1$) correspondant au cas où les coûts sont unitaires dans l'Exemple 1.

	vérité		
$c_{\hat{y}}(y)$	$y = h$	$y = b$	$y = n$
$\hat{Y} = \{h\}$	0	1	1
$\hat{Y} = \{b\}$	1	0	1
$\hat{Y} = \{n\}$	1	1	0
$\hat{Y} = \{h, b\}$	1/3	1/3	1
$\hat{Y} = \{b, n\}$	1	1/3	1/3
$\hat{Y} = \{h, n\}$	1/3	1	1/3
$\hat{Y} = \{h, b, n\}$	0.5	0.5	0.5

Table 4.3: Matrice de coûts définie par la mesure F_1 dans le cas des coûts unitaires

Comme nous pouvons le remarquer, cette formule est valable uniquement avec des fonctions de coûts unitaires. Dans l'Exemple 1, si la vérité est h, le fait de prédire n ou b ne devrait pas être équivalent dans la pratique, mais ici, ils conduiront aux mêmes scores de précision et de rappel.

Au final, nous avons vu que, même dans le cas simple des coûts unitaires, nous pouvons définir les coûts de prédictions indéterminées de manières très

différentes. Ainsi, il nous semble important de dégager des propriétés générales qui serviront de lignes directrices pour évaluer les prédictions indéterminées selon l'aversion au risque et le niveau de prudence du décideur.

4.2 Propriétés générales pour les coûts des prédictions indéterminées

Dans ce paragraphe, nous explorons les propriétés qui peuvent être considérées comme souhaitables lors de la définition des coûts des prédictions indéterminées. Nous allons commencer avec des propriétés qui rendent les prédictions indéterminées possibles, et nous allons ensuite suggérer certaines propriétés de bon sens. Enfin, nous allons discuter de certaines propriétés qui peuvent être souhaitables dans certains contextes, et indésirables dans d'autres. Cela nous permettra ensuite d'examiner certains travaux existants à la lumière de ces propriétés, avant de proposer une formulation générique pour définir les coûts des prédictions indéterminées.

4.2.1 Rendre les prédictions indéterminées possibles

La démarche et l'idée de pénalisation de l'indétermination vues dans le paragraphe 4.1 pour la justesse affaiblie nous donnent certaines idées transposables pour définir les coûts des prédictions indéterminées de manière générique. Nous commençons par définir ce que nous appellerons le *coût affaibli*.

Définition 3. *Etant donné les coûts $c_{\hat{y}}$ des prédictions déterminées $\hat{y} \in \Omega$, le coût affaibli d'une prédiction indéterminée $\hat{Y}$ est défini comme la moyenne (notée $\bar{c}$) des coûts des éléments singletons $\hat{y} \in \hat{Y}$ qui le composent :*

$$\bar{c}_{\hat{Y}}(y) = \frac{\sum_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)}{|\hat{Y}|}. \tag{4.4}$$

Nous notons que le coût affaibli se réduit à la justesse affaiblie quand les coûts sont unitaires.

Exemple 6. *Nous illustrons ici le coût affaibli dans l'exemple de la reconnaissance d'obstacles. Comme les coûts des prédictions déterminées sont donnés dans le Tableau 3.1 de l'Exemple 2, nous pouvons maintenant déduire ceux pour les prédictions indéterminées :*

	vérité		
$c_{\hat{Y}}(y)$	$y = h$	$y = b$	$y = n$
$\hat{Y} = \{h\}$	0	1	2
$\hat{Y} = \{b\}$	1	0	2
$\hat{Y} = \{n\}$	4	4	0
$\hat{Y} = \{h, b\}$	$1/2$	$1/2$	2
$\hat{Y} = \{b, n\}$	$5/2$	2	1
$\hat{Y} = \{h, n\}$	2	$5/2$	1
$\hat{Y} = \{h, b, n\}$	$5/3$	$5/3$	$4/3$

Table 4.4: Matrice de coûts étendue à $\mathcal{Y} = 2^{\Omega} \setminus \emptyset$ selon la définition du coût affaibli

Nous dirons qu'une prédiction indéterminée est *possible* si elle satisfait la propriété suivante :

Propriété 1 (Possibilité d'une prédiction indéterminée $\hat{Y}$)**.** *Une prédiction indéterminée $\hat{Y}$ est dite possible s'il existe une distribution de probabilités p telle que :*

$$\mathbb{E}[c_{\hat{Y}}] < \min_{\hat{y} \in \hat{Y}} \mathbb{E}[c_{\hat{y}}]$$

Cette propriété est liée à la définition de la prédiction optimale énoncée dans l'Équation (3.3). Elle traduit l'idée que, pour une matrice de coûts donnée, pour qu'une prédiction indéterminée $\hat{Y}$ soit possible, il faut qu'elle soit meilleure (moins coûteuse) que tous les éléments $\hat{y} \in \hat{Y}$ qui le composent pour une distribution donnée.

Définition 4. *Nous définissons également des relations d'ordre $<$ et $>$ pour comparer les fonctions de coûts. Soit deux prédictions $\hat{Y}_1$ et $\hat{Y}_2$, nous avons*

$$c_{\hat{Y}_1} < c_{\hat{Y}_2} \Leftrightarrow \forall y \in \Omega,\ c_{\hat{Y}_1}(y) < c_{\hat{Y}_2}(y) \tag{4.5}$$

$$c_{\hat{Y}_1} > c_{\hat{Y}_2} \Leftrightarrow \forall y \in \Omega,\ c_{\hat{Y}_1}(y) > c_{\hat{Y}_2}(y) \tag{4.6}$$

Maintenant, si nous voulons produire des prédictions indéterminées, une première exigence évidente est qu'au moins l'une d'elles soit possible, ce qui conduit à la propriété suivante :

Propriété 2 (Possibilité de prédictions indéterminées)**.** *Les coûts sont considérés comme rendant les prédictions indéterminées possibles si au moins une prédiction indéterminée $\hat{Y} \in 2^{\Omega} \setminus \emptyset$ est possible.*

Ceci est une propriété essentielle, car si elle n'est pas remplie, alors parler de classification indéterminée n'a pas de sens. A la lumière de cette propriété, nous pouvons facilement montrer que le coût affaibli ou n'importe quelle fonction de coût c telle que $c_{\hat{Y}} > \bar{c}_{\hat{Y}}$ ne rend pas les prédictions indéterminées possibles.

Proposition 1. *Le coût affaibli $\bar{c}_{\hat{Y}}(y)$ est tel que pour tout $\hat{Y} \in 2^{\Omega} \setminus \emptyset$*

$$\mathbb{E}[\bar{c}_{\hat{Y}}] \geq \min_{\hat{y} \in \hat{Y}} \mathbb{E}[c_{\hat{y}}]$$

Démonstration. Nous avons

$$\begin{aligned}\mathbb{E}[\bar{c}_{\hat{Y}}] &= \sum_{y \in \Omega} p(y) \bar{c}_{\hat{Y}}(y) = \frac{1}{|\hat{Y}|} \sum_{y \in \Omega} p(y) \sum_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y) \\ &= \frac{1}{|\hat{Y}|} \sum_{\hat{y} \in \hat{Y}} \sum_{y \in \Omega} p(y) c_{\hat{y}}(y) = \frac{1}{|\hat{Y}|} \sum_{\hat{y} \in \hat{Y}} \mathbb{E}[c_{\hat{y}}].\end{aligned}$$

Ainsi $\mathbb{E}[\bar{c}_{\hat{Y}}]$ est la moyenne des $\mathbb{E}[c_{\hat{y}}]$ pour tout $\hat{y} \in \hat{Y}$, par définition, elle ne peut être inférieure à $\min_{\hat{y} \in \hat{Y}} \mathbb{E}[c_{\hat{y}}]$. □

Cette proposition montre que le coût affaibli ne peut jamais remplir la Propriété 1 quel que soit $\hat{Y}$, ce qui prouve bien que ni $\bar{c}$ ni c telle que $c_{\hat{Y}} > \bar{c}_{\hat{Y}}$ ne peuvent satisfaire la Propriété 2. Ainsi, $c_{\hat{Y}} \not\geq \bar{c}_{\hat{Y}}$ est une condition nécessaire pour que les prédictions indéterminées soient possibles. Cela montre également que le coût affaibli, tout comme sa version en coûts unitaires la justesse affaiblie, ne constituent pas des choix raisonnables pour produire et évaluer des prédictions indéterminées.

Une propriété plus forte est d'exiger toutes les prédictions indéterminées à satisfaire cette condition nécessaire :

Propriété 3 (permissivité de l'indétermination)**.** *Les coûts sont dits permissifs vis-à-vis de l'indétermination si, pour tout* $\hat{Y} \in 2^{\Omega} \setminus \emptyset$*, nous avons*

$$c_{\hat{Y}} \not\geq \bar{c}_{\hat{Y}}$$

A moins qu'il y ait des prédictions indéterminées que nous voulons absolument éviter ou fortement pénaliser lors de l'évaluation, cette propriété nous parait raisonnable. Nous considérerons maintenant des propriétés qui ne sont pas nécessaires pour rendre les prédictions indéterminées possibles, mais qui nous paraissent comme de bon sens et généralement souhaitables.

4.2.2 Propriétés de bon sens

Une première propriété de bon sens, en rapport avec les critiques faites à la justesse affaiblie précédemment dans le cas des coûts unitaires, est que la prudence devrait être récompensée dans une certaine mesure. Autrement dit, une prédiction indéterminée $\hat{Y}$ devrait être récompensée si elle contient la valeur réelle y, *i.e.*, si $y \in \hat{Y}$.

Propriété 4 (Récompense de la prudence légitime)**.** *Les coûts sont dits récompensant la prudence légitime si, pour tout* $\hat{Y}$*, nous avons*

$$y \in \hat{Y} \Rightarrow c_{\hat{Y}}(y) < \bar{c}_{\hat{Y}}(y)$$

Cette propriété complète celle de la permissivité de l'indétermination (Propriété 3), car elle spécifie au moins un sous-ensemble de valeurs pour lesquelles l'inégalité $c_{\hat{Y}} \not\geq \bar{c}_{\hat{Y}}$ doit être satisfaite.

Jusqu'à présent, nous avons exploré des propriétés pour rendre les prédictions indéterminées possibles, mais une autre exigence logique est que les prédictions déterminées devraient également rester possibles. En particulier, cela signifie qu'aucune des prédictions indéterminées ne doit avoir un coût toujours plus bas que le minimum des coûts de ses éléments.

Propriété 5 (Non dominance des prédictions indéterminées)**.** *Les prédictions indéterminées sont dites non-dominantes si, pour tout* $\hat{Y} \in 2^{\Omega} \setminus \emptyset$*, nous avons*

$$\forall y \in \Omega : \; c_{\hat{Y}}(y) \geq \min_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)$$

Concrètement, cette propriété permet de garantir le fait que prédire uniquement la classe soit préférable à prédire un ensemble contenant la vraie classe. En effet, la relation $c_{\hat{Y}}(y) \geq \min_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)$ devrait être vérifiée pour tout $y \in \hat{Y}$. Car si la classe observée y est incluse dans la prédiction indéterminée $\hat{Y}$, il est logique de considérer que le coût de la prédiction déterminée $\hat{y} = y$ soit inférieur à celui de prédire $\hat{Y}$.

En résumé, les Propriétés 3 et 5 définissent déjà certaines contraintes que $c_{\hat{Y}}$ devrait suivre. Une autre est que, si les vecteurs de coûts formés par les éléments d'une prédiction indéterminée sont les mêmes à une permutation près, pour deux classes observées y et y', alors $c_{\hat{Y}}(y)$ et $c_{\hat{Y}}(y')$ doit être identiques, pour des raisons de symétrie. Avant d'énoncer la propriété en question, nous introduisons quelques notations. Si $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_n\}$ est une prédiction indéterminée et y une classe, nous noterons $C_{(\hat{Y})}(y) = (c_{\hat{y}_{(1)}}(y), \ldots, c_{\hat{y}_{(n)}}(y))$ le vecteur ordonné tel que $c_{\hat{y}_{(i)}}(y) \leq c_{\hat{y}_{(i+1)}}(y)$

Propriété 6 (Invariance par permutation)**.** *Les coûts de la prédiction indéterminée $\hat{Y}$ sont dits invariants par permutation si, pour deux classes différentes y et y', nous avons :*

$$C_{(\hat{Y})}(y) = C_{(\hat{Y})}(y') \Rightarrow c_{\hat{Y}}(y) = c_{\hat{Y}}(y')$$

Par exemple, dans l'Exemple 6, nous avons $C_{\{h,b,n\}}(h) = C_{\{h,b,n\}}(b) = (0, 1, 4)$, ainsi si nous voulons respecter la propriété de l'invariance par permutation, nous devons avoir $c_{\{h,b,n\}}(h) = c_{\{h,b,n\}}(b)$ dans le Tableau 4.4. Nous ne voyons pas de raisons particulières pour ne pas respecter cette propriété d'invariance, nous la considérons donc comme une propriété généralement souhaitable.

4.2.3 Propriétés dépendant des contextes

Nous allons maintenant étudier des propriétés dont la nécessité peut dépendre du contexte de l'utilisation des prédictions indéterminées. En particulier, nous allons distinguer deux cadres différents :

– **Le cadre du filtrage** : dans ce cas, nous voulons filtrer certaines classes (éventuellement avec des méthodes de calcul de faible coût)

avant d'appliquer une procédure plus coûteuse. Dans ce cas, il semble essentiel que la prédiction indéterminée contienne la vraie classe, sinon la procédure coûteuse est appliquée pour rien. Dans ce cas, il s'agit donc essentiellement "d'éliminer" les mauvaises classes ;

– **Le cadre de la décision** : dans ce cas, l'indétermination vise à donner des prédictions prudentes pour le décideur, qui peut alors faire une décision selon l'information fournie si elle/il veut. Dans ce cas, une prédiction indéterminée est avant tout une information précieuse en elle-même, car elle peut indiquer au décideur une situation ambiguë. Une telle information peut donc être souhaitable même si la véritable classe n'est pas en son sein.

Nous illustrons dans les paragraphes suivants et avec l'Exemple 7 que ces cadres sont étroitement liés à la façon dont les prédictions sont utilisées. Décrivons maintenant quelques propriétés.

Comportement en cas d'erreurs

Quand des erreurs sont commises avec les prédictions indéterminées, nous pouvons avoir deux comportements : soit pénaliser ces erreurs (par rapport à une prédiction déterminée), soit chercher quand même la prudence même en cas d'erreurs. Cela peut se traduire par les deux propriétés suivantes :

Propriété 7 (Aversion aux erreurs)**.** *Les coûts sont dits opposés aux erreurs si, pour tout $\hat{Y}$, nous avons*

$$y \notin \hat{Y} \Rightarrow c_{\hat{Y}}(y) \geq \bar{c}_{\hat{Y}}(y)$$

Propriété 8 (Enclin à la prudence)**.** *Les coûts sont dits enclins à la prudence si, pour tout $\hat{Y}$, nous avons*

$$y \notin \hat{Y} \Rightarrow c_{\hat{Y}}(y) \leq \bar{c}_{\hat{Y}}(y)$$

De toute évidence, ces deux propriétés complètent la Propriété 4 de manières opposées, en spécifiant le comportement souhaité pour les prédictions indéterminées en cas d'erreur. Dans le cadre du filtrage, faire une erreur en

donnant une prédiction indéterminée est clairement pénalisant, ainsi la Propriété 7 est plus adaptée que la Propriété 8 au cadre du filtrage.

Dans le cadre de la décision nous pensons que le choix n'est pas ainsi évident. Bien sûr, le décideur peut vouloir de pénaliser le fait de commettre une erreur en étant indéterminé plus qu'une prédiction déterminé. Mais il est aussi possible que le décideur préfère être plus prudent quand il y a une manque d'information, même si la prédiction est erronée. En fait, en exprimant clairement l'ambiguïté de la situation avec une prédiction indéterminée, des inspections supplémentaires pourraient conduire à une bonne prédiction, tandis qu'une prédiction déterminée va dissuader le décideur à mener des inspections supplémentaires. Par exemple, dans notre exemple de reconnaissance d'obstacles, une indétermination va toujours déclencher un message d'alarme au conducteur, tandis que si le système décide péremptoirement qu'il n'y a pas d'obstacle, aucune alarme ne sera donnée. Ainsi, nous pensons que dans ce cadre, toutes les deux propriétés peuvent être justifiées selon le contexte pratique.

Exemple 7. *Dans l'exemple de la reconnaissance d'obstacles, le choix entre le cadre du filtrage ou de la décision dépend de la manière dont la prédiction donnée par le véhicule intelligent est utilisée.*

Dans le cadre du filtrage, nous pouvons imaginer que le véhicule dispose de plusieurs algorithmes de reconnaissance, dont un à faible complexité mais d'un pouvoir prédictif correct dans tous les cas, et une panoplie d'algorithmes spécialisés de très bonne performance ***uniquement*** *dans des cas précis : par exemple, un algorithme pour reconnaître les obstacles immobiles (arbres, panneaux,. . .), et un autre pour pister les objets mouvant (piétons, bicyclettes, autres véhicules, . . .).*

Nous voulons utiliser l'algorithme générique en premier lieu pour résoudre les cas faciles, mais nous l'autorisons à produire des prédictions indéterminées dans les cas plus complexes, pour notamment savoir quel algorithme spécialisé ***adapté*** *à appliquer dans un second temps. Ce cadre d'utilisation de prédictions indéterminées est typique des problèmes de préclassification (Ha, 1997), et il est plus intéressant de respecter la Pro-*

priété 7.

Cependant, dans le cadre d'une assistance à la conduite, la décision et les actions finales sont réalisées par le conducteur humain. Dans ce cas, respecter la Propriété 8 permet de donner des informations aussi fiables que possibles au conducteur. Par exemple, donner une alarme (même fausse) : "Attention, situation incertaine devant" va accroître la vigilance du conducteur, alors que donner une fausse prédiction "Attention bicyclette" peut être déroutant voire provoquer la confusion chez le conducteur, pire encore, si une fausse prédiction "Aucun obstacle" est faite, le système ne donnera aucun avertissement au conducteur.

Variabilité des coûts indéterminés selon leurs éléments

La Propriété 6 nous dit que les vecteurs de coûts identiques devraient donner des coûts identiques pour les prédictions indéterminées correspondantes. Pourtant, il ne précise pas le comportement lorsque les vecteurs sont différents. Les deux propriétés suivantes remédient à cette situation.

Propriété 9 (Invariabilité des coûts indéterminés)**.** *Pour tout $\hat{Y}$ et toute paire de classes $y, y' \in \hat{Y}$, $c_{\hat{Y}}$ est dite invariable si nous avons*

$$c_{\hat{Y}}(y) = c_{\hat{Y}}(y')$$

Propriété 10 (Variabilité des coûts indéterminés)**.** *$c_{\hat{Y}}$ est dite variable selon ses éléments y et y' s'il existe une prédiction $\hat{Y}$ et une paire de classes $y, y' \in \hat{Y}$ telles que*

$$C_{(\hat{Y})}(y) \neq C_{(\hat{Y})}(y') \Rightarrow c_{\hat{Y}}(y) \neq c_{\hat{Y}}(y')$$

Les Propriétés 9 et 10 sont surtout utiles pour l'évaluation des classifieurs. Il faut donc choisir en fonction de l'objectif du classifieur. La première propriété est plus adaptée pour évaluer un classifieur utilisé dans le cadre du filtrage, puisque le but ici est tout simplement de reconnaître les classes non pertinentes. Par exemple, dans l'Exemple 7, le coût d'une prédiction indéterminée correspond au coût de l'utilisation de l'algorithme spécialisé adapté, qui est indépendant de la vérité.

En revanche, la seconde propriété correspond mieux au cadre de la décision, si les coûts d'une prédiction indéterminé varient en fonction des vérités, alors ils devraient nous permettre de recommander parfois différentes classes optimales, même avec une distribution de probabilités identique p. L'Exemple 8 donne une illistration des Propriétés 9 et 10.

Exemple 8. *Pour simplifier les choses, supposons dans cet exemple que seules les classes h et n sont présentes dans pour la reconnaissance d'obstacles. La matrice de coûts donnée dans le Tableau 4.5 satisfait les Propriétés 3, 4 et 9.*

	vérité	
$c_{\hat{Y}}(y)$	$y = h$	$y = n$
$\hat{Y} = \{h\}$	0	2
$\hat{Y} = \{n\}$	4	0
$\hat{Y} = \{h, n\}$	0.5	0.5

TABLE 4.5: Matrice de coûts avec seulement les classes h et n et vérifiant la Propriété 9

Nous pouvons alors représenter aisément la règle de décision comme une fonction de $p(n|\mathbf{x})$. La règle de décision associée au Tableau 4.5 est représentée dans la partie supérieure de la Figure 4.1.

Ensuite, nous considérons maintenant l'opinion d'un expert qui juge que le fait de prédire $\{h, n\}$ quand la vérité est $\{n\}$ est une sorte de fausse alarme et est par conséquent plus coûteux que quand $\{h\}$ est la vérité. Supposons que l'expert juge que ce premier est trois fois plus coûteux, et posons que le coût espéré de prédire $\{h, n\}$ est de 0.5 quand $p(h) = p(n)$ (pour rester consistant avec le Tableau 4.5), ceci donne alors la matrice du Tableau 4.6, qui vérifie maintenant les Propriétés 3, 4 et 10.

Dans ce cas, la règle de décision est légèrement modifiée et est représentée dans la partie inférieure de la Figure 4.1. Nous constatons que la frontière entre $\{h, n\}$ et $\{h\}$ a décalé vers la gauche à cause de la diminution du coût $c_{\{h,n\}}(h)$: la décision $\{h, n\}$ est plus favorisé pour les valeurs faibles de $p(n|\mathbf{x})$ par rapport à ce que nous avions avec la matrice de coûts

$c_{\hat{Y}}(y)$	vérité $y = h$	$y = n$
$\hat{Y} = \{h\}$	0	2
$\hat{Y} = \{n\}$	4	0
$\hat{Y} = \{h, n\}$	0.25	0.75

Table 4.6: Matrice de coûts avec seulement les classes h et n et vérifiant la Propriété 10

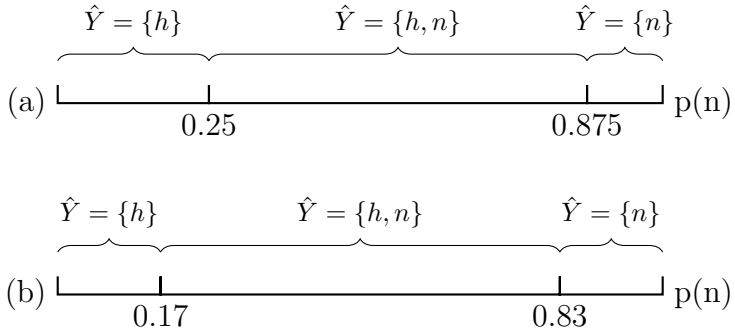


Figure 4.1: Représentation graphique de la régle de décision comme étant une fonction de $p(n|\mathbf{x})$; (a) : décision avec la matrice de coûts du Tableau 4.5 ; (b) : décision avec la matrice de coûts du Tableau 4.6.

du Tableau 4.5. D'un autre côté, comme le coût $c_{\{h,n\}}(n)$ a été augmenté, la frontière entre $\{h, n\}$ et $\{n\}$ a aussi décalé vers la gauche, ce qui pénalise la prédiction $\{h, n\}$ pour les grandes valeurs de $p(n|\mathbf{x})$. Nous voyons que les changements de coûts impactent bien sur la prise de décision selon la manière voulue.

Borne supérieure

En complément à la Propriété 5, il peut être souhaitable de borner le coût des prédictions indéterminées par une valeur supérieure :

Propriété 11 (Borne supérieure)**.** *Une prédiction indéterminée est dite bornée supérieurement si pour tout $\hat{Y} \in 2^{\Omega} \setminus \emptyset$, nous avons*

$$\forall y \in \Omega, \qquad c_{\hat{Y}}(y) \leq \max_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)$$

Que cette propriété soit aussi souhaitable que la Propriété 5 n'est pas toujours évident. En effet, dans le cadre du filtrage, le coût de la procédure supplémentaire peut aller au-delà du coût maximal des éléments de $\hat{Y}$, ainsi, en cas d'erreur nous pouvons vouloir donner une pénalité très forte pour éviter d'utiliser la procédure supplémentaire pour rien. Remplir cette propriété, cependant, prend tout son sens dans le cadre de la décision, puisque quel que soit la décision finale du décideur, le coût encouru ne peut pas être plus grande que le pire des choix, quelle que soit la vérité.

Remarque 1. *Il convient de noter que si l'on remplit les deux Propriétés 5 et 11, alors il implique que, si tous les éléments $\hat{y}$ d'une prédiction indéterminée $\hat{Y}$ sont tels que $c_{\hat{y}}(y) = c^{ste}$ (c^{ste} une constante) pour une classe y quelconque, alors nous avons $c_{\hat{Y}}(y) = c^{ste}$.*

Par exemple, dans l'Exemple 6, puisque nous avons $c_{\{h\}}(n) = c_{\{b\}}(n) = 2$, si nous respectons les Propriétés 5 et 11, alors nous avons $c_{\{h,b\}}(n) = 2$.

4.3 Revue des travaux similaires

Dans cette section, nous allons ré-examiner des propositions existantes à la lumière des propriétés précédemment définies.

Justesse affaiblie selon l'utilité La proposition de Zaffalon et collab. (2012) est très intéressante, car elle satisfait absolument toutes nos propriétés proposées (les propriétés dépendant des contextes ne sont pas mutuellement exclusives en coûts unitaires), en raison de la symétrie inhérente des coûts considérés et de la restriction aux coûts unitaires.

Même si ce n'est pas nécessaire (en raison des fortes bases théoriques déjà présentes dans la proposition initiale), ceci montre que leur proposition est tout à fait justifiée dans le cas des coûts unitaires.

Règle de rejet sélectif optimal Ha (1997) a élaboré un classifieur indéterminé sur la base du rejet partiel. Son but est de trouver le compromis optimal entre le rejet et l'erreur de classification en utilisant une structure

de perte spécifique (matrice de coût) défini comme suit :

$$c_{\hat{Y}}(y) = L_{\hat{Y}}(y) + L_{ip}(\hat{Y})$$

où $L_{\hat{Y}}(y) = 0$ si la vraie classe y est incluse dans la prédiction $\hat{Y}$, et sinon $L_{\hat{Y}}(y) = L_{err}(y)$ qui reflète la perte encourue en manquant la vraie classe y. $L_{ip}(\hat{Y}) = C_{ip}(|\hat{Y}| - 1)$ où C_{ip} est un paramètre constant représentant le coût d'être indéterminé, avec la condition que $C_{ip} < 1/2 C(y)$ pour tout y. La matrice de coût obtenue quand $\Omega = \{a, b, c\}$ est donnée par le Tableau 4.7.

	vérité		
$c_{\hat{Y}}(y)$	$y = a$	$y = b$	$y = c$
$\hat{Y} = a$	0	$L_{err}(b)$	$L_{err}(c)$
$\hat{Y} = b$	$L_{err}(a)$	0	$L_{err}(c)$
$\hat{Y} = c$	$L_{err}(a)$	$L_{err}(b)$	0
$\hat{Y} = \{a, b\}$	C_{ip}	C_{ip}	$L_{err}(c) + C_{ip}$
$\hat{Y} = \{a, c\}$	C_{ip}	$L_{err}(b) + C_{ip}$	C_{ip}
$\hat{Y} = \{b, c\}$	$L_{err}(a) + C_{ip}$	C_{ip}	C_{ip}
$\hat{Y} = \{a, b, c\}$	$2C_{ip}$	$2C_{ip}$	$2C_{ip}$

Table 4.7: Matrice de coûts construite selon la définition de Ha

La condition $C_{ip} < 1/2 C(y)$ assure que les Propriétés 3 et 4 sont satisfaites. En outre, cette proposition satisfait également les Propriétés 5 et 6.

En ce qui concerne les propriétés dépendant du contexte, cette proposition satisfait les Propriétés 9 et 7, et ne satisfait pas la Propriété 11. À la lumière de ces propriétés, il est clair que cette proposition est plus adaptée au cadre du filtrage qu'à celui de la décision, ce qui est précisément le cadre dans lequel Ha (1997) définit son travail, puisque Ha a pour l'objectif de construire un classifieur adapté pour la pré-classification en reconnaissance d'images (similaire à ce que nous évoquons dans l'Exemple 7).

Matrice de coût définie en utilisant différentes hypothèses Abellán et Masegosa (2012) proposent une autre mesure pour les classifieurs imprécis. Leur principal objectif est d'élaborer une mesure de comparaison pour les classifieurs imprécis (l'indétermination des prédictions provient dans

leur cas de l'utilisation de probabilités imprécises). La matrice de coûts qu'ils proposent, une fois une transformation linéaire appliquée pour avoir $c_{\hat{y}}(\hat{y}) = 0$ (afin de faciliter la comparaison avec les approches précédentes), sont les suivantes

– si $y \in \hat{Y}$, alors

$$c_{\hat{Y}}(y) = \log |\hat{Y}|$$

– si $y \notin \hat{Y}$, alors

$$c_{\hat{Y}}(y) = \log |\Omega| \left(\frac{\max_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)}{|\Omega| - 1} + 1 \right)$$

Cela satisfait les Propriétés 5, 11 et 6, mais il n'y a aucune garantie que les Propriétés 3 et 4 soient satisfaites, ce qui est un inconvénient potentiel. La proposition modifie également les coûts initiaux (les valeurs $c_{\hat{y}}(y)$ sont modifiées), mais avec une transformation linéaire, donc ce n'est qu'un inconvénient mineur. Elle satisfait les Propriétés 7 et 9, ce qui semble indiquer qu'elle est plus adaptée au cadre du filtrage, mais il est pas tout à fait clair que Abellán et Masegosa (2012) considère une telle application, comme ils n'utilisent les coûts définis que pour comparer les classifieurs imprécis sensibles aux coûts.

4.4 Formule pour dériver les coûts des prédictions indéterminées

Dans cette section, nous proposons une manière générale pour produire des coûts pour les prédictions indéterminées à partir des coûts déterminés $c_{\hat{y}}$, en se basant sur la notion de justesse affaiblie selon l'utilité introduite dans la Section 4.1, que nous appellerons " coût affaibli généralisé ".

4.4.1 Formulation de base

Nous rappelons ici la forme de base du coût affaibli avec des fonctions de coûts génériques précédemment énoncée dans l'Équation (4.4) :

$$\bar{c}_{\hat{Y}}(y) = \frac{\sum_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)}{|\hat{Y}|}.$$

Comme il s'agit seulement d'une moyenne arithmétique simple, un moyen naturel d'étendre cette formule est d'utiliser une moyenne généralisée, qui est

$$\bar{c}^p_{\hat{Y}}(y) = \left(\frac{1}{|\hat{Y}|} \sum_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)^p \right)^{\frac{1}{p}} \tag{4.7}$$

avec $p \in]-\infty, \infty[$.

Nous obtenons $\bar{c}_{\hat{Y}}(y)$ lorsque $p = 1$. Une caractéristique intéressante des moyennes généralisées est que si $q < p$, alors $\bar{c}^q_{\hat{Y}}(y) \leq \bar{c}^p_{\hat{Y}}(y)$, avec les deux étant égales si et seulement si $c_{\hat{y}}(y) = c_{\hat{y}'}(y)$ pour tout $\hat{y}, \hat{y}' \in \hat{Y}$. Par conséquent, si nous voulons un coût inférieur à $\bar{c}_{\hat{Y}}$, nous avons juste à choisir $p < 1$, et dans le cas contraire $p > 1$ pour un coût supérieur.

Cependant, nous ne pouvons pas définir $\bar{c}^p_{\hat{Y}}(y)$ pour $p < 0$ si nous avons $c_{\hat{y}}(y) = 0$ pour certaines valeurs de $\hat{y}$ et y, en raison de la division par zéro. Cette question numérique n'est pas un inconvénient important car elle peut être contournée. Comme la formule est principalement utilisée pour comparer les classifieurs, alors nous pouvons simplement ajouter une translation ϵ à la matrice de coût, de sorte qu'il n'y ait plus de 0 dans la matrice. Comme la translation affecte de manière uniforme tous les classifieurs, il n'y aura pas d'impact sur la comparaison.

En outre, il n'est pas rare d'avoir une matrice de coût qui ne comporte aucun élément nul : dans un cadre où les coûts représentent des coûts monétaires, le choix de la bonne classe a également un coût (le coût de base de la fabrication, le coût d'achat minimal,...). Dans le cas où la matrice de coût ne peut pas être modifiée arbitrairement (par exemple les coûts sont définis par un expert), il est possible d'éviter ce problème de division par zéro en considérant la matrice duale (appelée matrice de récompenses). Ainsi, une case à 0 dans une matrice de coûts sera celle avec la récompense maximale (par exemple 1, si la matrice est normalisée) dans la matrice de récompenses correspondante. Cette transformation peut être utile, car il semble plus intuitif de supposer qu'il n'y ait pas d'élément nul pour une matrice de récompense que pour une matrice de coût : chaque classe peut donner une sorte d'informations (*i.e.* récompense), même si elle n'est pas la vérité.

4.4.2 Variantes et propriétés de la formule

Quand $p = 0$, nous choisissons de nous référer, par convention, à la moyenne géométrique :

$$\bar{c}^{0}_{\hat{Y}}(y) = \Big(\prod_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y) \Big)^{\frac{1}{|\hat{Y}|}} \tag{4.8}$$

Ainsi, en choisissant une valeur réelle $r \in [0; 1]$, nous pouvons définir deux variantes du coût affaibli généralisé en fonction des Propriétés 7et 8 :

Définition 5. *Variante "enclin à la prudence", où pour toute classe y nous avons*

$$c_{\hat{Y}}(y) = \bar{c}^{1-r}_{\hat{Y}}(y)$$

Définition 6. *Variante "aversion aux erreurs",où pour toute classe $y \in \hat{Y}$ nous avons*

$$c_{\hat{Y}}(y) = \bar{c}^{1-r}_{\hat{Y}}(y)$$

et pour toute classe $y \notin \hat{Y}$ nous avons

$$c_{\hat{Y}}(y) = \bar{c}^{1+r}_{\hat{Y}}(y)$$

Quelle que soit la variante choisie, la formule satisfait les Propriétés 3, 4 et 6. Comme nous avons par définition de la moyenne généralisée $\bar{c}^{-\infty}_{\hat{Y}} = \min$ et $\bar{c}^{+\infty}_{\hat{Y}} = \max$, les Propriétés 11 et 5 sont aussi naturellement satisfaites pour toute valeur r. En ce qui concerne les propriétés dépendant des contextes, notre formule satisfait la Propriété 10 (contrairement aux autres travaux que nous avons vu précédemment), et chaque variante satisfait soit la Propriété 8 soit la Propriété 7. Dû au fait que la formule satisfait la Propriété 10, elle est plus adaptée au cadre de la décision.

L'avantage de définir un paramètre r est qu'il devrait être capable (ceci est à vérifier dans dans le cas de la régression ordinale dans le Chapitre 6) de calibrer le degré de prudence que nous aimerions avoir. En effet, plus r est élevé, plus la prudence est récompensée (et plus les erreurs sont pénalisées si nous prenons la variante "aversion aux erreurs"), et quand nous prenons $r = 0$, aucune récompense n'est créditée à la prudence.

Exemple 9. *Nous donnons ici un exemple des matrices de coûts que nous pouvons obtenir avec le coût affaibli généralisé. En fixant $r = 0.5$, nous pouvons obtenir les matrices pour les deux variantes :*

	vérité		
$c_{\hat{Y}}(y)$	$y = h$	$y = b$	$y = n$
$\hat{Y} = \{h\}$	0	1	2
$\hat{Y} = \{b\}$	1	0	2
$\hat{Y} = \{n\}$	4	4	0
$\hat{Y} = \{h, b\}$	0.25	0.25	2
$\hat{Y} = \{b, n\}$	2.25	1	0.5
$\hat{Y} = \{h, n\}$	1	2.25	0.5
$\hat{Y} = \{h, b, n\}$	1	1	0.89

Table 4.8: Matrice de coûts étendue à $\mathcal{Y} = 2^{\Omega} \setminus \emptyset$ selon la variante "enclin à la prudence" du coût affaibli généralisé

	vérité		
$c_{\hat{Y}}(y)$	$y = h$	$y = b$	$y = n$
$\hat{Y} = \{h\}$	0	1	2
$\hat{Y} = \{b\}$	1	0	2
$\hat{Y} = \{n\}$	4	4	0
$\hat{Y} = \{h, b\}$	0.25	0.25	2
$\hat{Y} = \{b, n\}$	2.73	1	0.5
$\hat{Y} = \{h, n\}$	1	2.73	0.5
$\hat{Y} = \{h, b, n\}$	1	1	0.89

Table 4.9: Matrice de coûts étendue à $\mathcal{Y} = 2^{\Omega} \setminus \emptyset$ selon la variante "aversion aux erreurs" du coût affaibli généralisé

Nous constatons que la différence des deux variantes se trouve pour $c_{\{b,n\}}(h)$ et $c_{\{h,n\}}(b)$. Nous avons $\bar{c}_{\{b,n\}}(h) = \bar{c}_{\{h,n\}}(b) = 2.5$, ainsi, les valeurs prises par les deux variances sont bien conformes aux Propriétés 8 et 7.

Conclusion

Dans ce chapitre nous avons pu définir une ligne directrice pour établir des matrices de coûts étendues aux prédictions indéterminées, en proposant des propriétés que ces dernières devraient ou pourraient satisfaire. Nous avons vu que la définition de ces coûts doivent se faire en fonction de leur usage. Dans notre cas, ils donnent une première méthode pour définir des classifieurs indéterminés (avec des probabilités précis dans l'espace de prédictions 2^{Ω}), mais ils permettent surtout de comparer des classifieurs indéterminés et déterminés de manière équitable dans le cadre de la classification sensible aux coûts. Nous avons également élaboré une formule qui permet de dériver ces coûts de manière systématique. Il convient alors de tester et de caractériser les comportements de cette formule, notamment en fonction du paramètre de niveau de prudence r que nous avons défini. Nous effectuerons ces tests dans le Chapitre 6, dédié aux problèmes ordinaux.

Chapitre 5

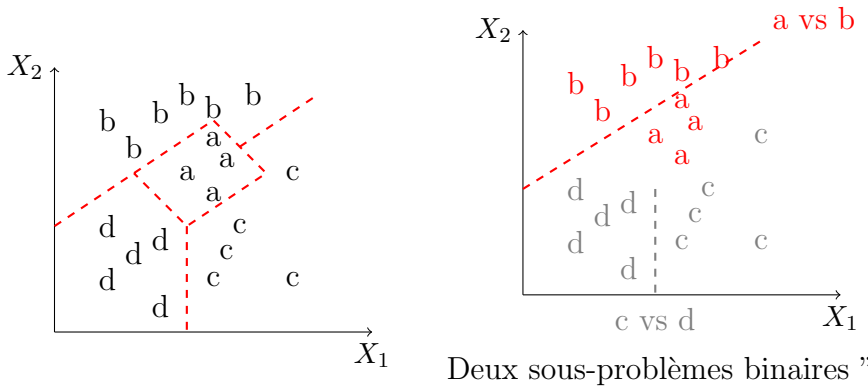
Les dichotomies emboîtées imprécises

Nous avons vu l'intérêt des probabilités imprécises pour construire des modèles prudents dans le Chapitre 3 et nous avons également détaillé comment définir des coûts d'erreur de classification génériques et les utiliser dans le cadre d'une classification sensible aux coûts pour mesurer la qualité des prédictions dans le Chapitre 4. Cependant, peu de classifieurs proposés dans la littérature utilisent ces deux cadres conjointement, en raison notamment du temps de calcul. Dans ce chapitre, nous allons proposer une méthode pour combiner ces deux cadres d'une manière efficace à l'aide d'une technique de décomposition d'un problème multiclasse en plusieurs problèmes binaires, les dichotomies emboîtées.

Dans un premier temps, nous allons donner une présentation générale des techniques de décomposition en problèmes binaires. Il ne s'agit pas de faire un état de l'art comparatif de toutes les approches existantes, mais plutôt de considérer les raisons théoriques qui font que les dichotomies emboîtées s'avèrent être une approche particulièrement intéressante dans notre cas. Nous verrons ensuite que l'utilisation des dichotomies emboîtées a des pré-requis, notamment en ce qui concerne le choix de la structure des dichotomies. Au final, nous caractériserons les comportements et les propriétés de cette approche à l'aide d'expériences.

5.1 Présentation générale

Les techniques de décomposition (ou de réduction) sont des approches pour résoudre des problèmes de classification multiclasses (et multilabels). Leur idée principale est de décomposer le problème initial complexe en un ensemble de sous-problèmes plus simples et plus faciles à résoudre. Une grande variété des éléments du problème initial sont décomposables : les variables d'entrée, l'espace de ces variables, les instances de données ou encore la classe. Rokach (2006) donne une revue complète de ces différents types de décompositions possibles. Nous nous intéressons plus particulièrement à celle que Rokach appelle *l'agrégation des concepts*. Il s'agit d'agréger les différentes valeurs possibles de la classe initiale en des sous-ensembles de l'espace de sortie, afin de décomposer le problème initial de classifica-



Problème multiclasse

Deux sous-problèmes binaires "a vs b" (en rouge) et "c vs d" (en gris)

FIGURE 5.1: Passage d'un problème multiclasse en problèmes binaires

tion multiclasse en une série de problèmes binaires. Un classifieur (que nous appellerons *classifieur de base*) est construit pour chaque problème binaire. Nous appellerons ce type de techniques décomposition binaire par la suite.

5.1.1 Cadre théorique

Soit un problème de classification multiclasse où l'ensemble des étiquettes est Ω, une décomposition binaire consiste à former ℓ paires d'événements $\{A_i, B_i\}$ $(i \in [1, \ell])$ où $A_i \cap B_i = \emptyset$ et $A_i, B_i \subseteq \Omega$. Ainsi, comme l'illustre l'Exemple 10, il s'agit de résoudre ℓ problèmes binaires où nous estimons si la vérité y appartienne à A_i ou B_i pour tout $i = 1, \ldots, \ell$, au lieu d'estimer directement le modèle joint sur Ω.

Exemple 10. *Sur la Figure 5.1, nous montrons que, même dans un cas simple, les frontières de décisions d'un problème multiclasse (côté gauche) peuvent être difficile à exprimer. Elles deviennent beaucoup plus simples (linaires) quand nous ne considérons que des sous-problèmes binaires (côté droit) où uniquement deux classes sont examinées à chaque fois. Une des spécificités des techniques de décomposition est qu'il est nécessaire de recombiner les résultats fournis par les sous-problèmes (par exemple les frontières "a vs b" et "c vs d") à la fin pour obtenir le résultat du modèle global.*

Dans une approche probabiliste, nous devons estimer les probabilités $\hat{p}(A_i \mid \{A_i, B_i\}) = \alpha_i$ et $\hat{p}(B_i \mid \{A_i, B_i\}) = 1 - \alpha_i$, avec ce que nous appellerons le classifieur (binaire) de base. Ces probabilités conditionnelles permettent de dériver des contraintes sur la distribution de la probabilité jointe du modèle global (multiclasse) :

$$\begin{cases} \sum_{y \in A_i} \hat{p}(y) = \alpha_i \sum_{y \in A_i \cup B_i} \hat{p}(y) & (i = 1, \ldots, l) \\ \sum_{y \in \Omega} \hat{p}(y) = 1 \end{cases} \tag{5.1}$$

Un problème fréquent avec un tel ensemble générique d'estimations des probabilités conditionnelles est que les contraintes dans l'Équation (5.1) sont la plupart du temps incompatibles (Hastie et collab., 2001; Wu et collab., 2004), dans le sens où il n'existe aucune solution faisable à l'Équation (5.1). La levée de cette incompatibilité n'est pas évidente et il n'y a pas de solution optimale unique, même si nous autorisons ces probabilités à devenir des intervalles dans le cadre des probabilités imprécises (Destercke et Quost, 2011). Une stratégie habituelle consiste à trouver une probabilité jointe en minimisant une distance donnée (Hastie et Tibshirani, 1998; Wu et collab., 2004) sur les estimations $\hat{p}(y|\{A_i, B_i\})$.

5.1.2 Cadre pratique

Il existe une multitude de stratégies possibles pour former les ensembles $\{A_i, B_i\}$ (Aly, 2005), nous donnons ici les stratégies les plus couramment utilisées :

- la stratégie "un contre tous" consiste à former K problèmes binaires tels que, sachant $\Omega = \{\omega_1, \ldots, \omega_K\}$, nous avons $A_i = \omega_i$ et $B_i = \Omega \setminus A_i$ pour tout $i \in [1, K]$. Pour la prédiction, c'est le classifieur de base produisant le $\hat{p}(A_i|\{A_i, B_i\})$ maximal qui l'emporte. Malgré sa simplicité, cette stratégie peut donner de bonnes performances prédictives (Rifkin et Klautau, 2004).
- la stratégie "un contre un" (Hastie et Tibshirani, 1998) consiste à n'examiner qu'une paire de classes pour chaque sous-problème binaire.

Ainsi $\frac{K(K-1)}{2}$ problèmes sont formés. La phase de prédiction consiste à comptabiliser les prédictions des classifieurs de base comme des votes, la classe accumulant le plus de votes est la prédiction finale. Allwein et collab. (2001); Hsu et Lin (2002) suggèrent que cette stratégie est en général meilleure que celle de "un contre tous".

– la stratégie de "codes correcteurs d'erreurs" consiste appliquer la notion homonyme provenant des sciences de la télécommunication à la classification (Dietterich et Bakiri, 1995). Il s'agit de former N sous-problèmes binaires et de représenter chaque classe ω par un code correcteur de N bits (le i-ième bit est à 1 si $\omega \in A_i$, -1 si $\omega \in B_i$, 0 sinon). Pour la prédiction, un code similaire est construit pour chaque instance de données à prédire. Il suffit alors de comparer les codes prédits de chaque classe avec leurs codes définis initialement et de sélectionner la classe dont la distance (de Hamming) entre le code prédit et le code défini soit minimal.
– la stratégie de "classification hiérarchique" consiste à utiliser une division hiérarchique de l'espace Ω (Kumar et collab., 2002; Chen et collab., 2004; Eibe et Stefan, 2004). Il s'agit souvent d'arranger les classes dans une structure arborescente en mettant l'ensemble Ω en tant que racine et en effectuant un partitionnement itératif. Les dichotomies emboîtées (Fox, 1997) font partie de cette famille de techniques de décomposition binaires.

En général, les techniques de décomposition binaires partagent quelques avantages par rapport aux approches multiclasses directes :

– **Gain de pouvoir prédictif** : l'utilisation d'un ensemble de modèles peut aider à réduire la variance des modèles (Dietterich et Bakiri, 1995). De plus, l'utilisation des sous-ensembles $\{A_i, B_i\}$ permet parfois d'exploiter des informations spécifiques dont un modèle unique ne peut tenir compte (par exemple, une variable d'entrée peut être un facteur déterminant pour discriminer entre deux classes spécifiques, mais avoir un faible pouvoir discriminant en général).
– **Gain en temps de calcul** : s'il n'y a pas de dépendance entre les sous-problèmes, et comme un ensemble de classifieur simple est utilisé,

il est possible d'effectuer des parallélisations pour accélérer le temps de calcul. De plus, chaque classifieur de base ne travaille souvent que sur une portion des données, ce qui réduit également le temps de calcul et est particulièrement intéressant si la quantité de données est importante.

- **Simplification conceptuelle** : les modèles binaires sont en général plus facile à appréhender et à manipuler que les modèles multiclasses. Dans notre cas, nous verrons que le problème de minimisation du coût espéré dans le cadre du NCC évoqué dans l'Equation (3.23) peut être résolue aisément dans le cas binaire, alors qu'il est insoluble en temps polynomial dans le cas général. De plus, dans le cas de la classification hiérarchique, les regroupements des classes peuvent être significatifs en soi et rendre les problèmes plus interprétables.
- **Modularité** : la décomposition en sous-problèmes suggèrent qu'il est possible d'utiliser des classifieurs de base différents ou même un même classifieur de base mais avec des paramètres différents pour chaque sous-problème. Cette modularité peut être exploitée pour accroître davantage la performance prédictive.

Dans notre cas, nous avons choisi de nous concentrer sur la stratégie de décomposition hiérarchique, et plus particulièrement sur celle de dichotomies emboîtées, car elle ne souffre pas du problème d'incompatibilité mentionné dans la Section 5.1.1. En effet, les contraintes induites sont toujours compatibles en raison de la structure arborescente de la décomposition. Ainsi, nous allons utiliser la notation p au lieu de $\hat{p}$ par la suite. En outre, nous allons voir dans le prochain paragraphe que les dichotomies emboîtées permettent de calculer les coûts espérés d'une manière directe même quand nous nous plaçons dans le cadre des probabilités imprécises, ce qui les rend très adaptées pour la construction de classifieurs indéterminés sensibles aux coûts.

5.2 Dichotomies emboîtées et probabilités imprécises

Le principe de dichotomies emboîtées est de former une structure d'arbre binaire avec les classes, qui détermine les sous-problèmes binaires à résoudre. La technique consiste à partitionner de manière récursive un nœud de l'arbre $C \subseteq \Omega$ en deux sous-ensembles A et B (une dichotomie), jusqu'à ce que chaque nœud feuille corresponde à un singleton de la classe ($|C| = 1$). Le nœud racine est l'ensemble des classes Ω.

Par conséquent, chaque nœud C est associé à un problème de classification binaire où nous devons décider si la classe appartient à l'ensemble A ou B. Si un classifieur de base probabiliste standard (manipulant des probabilités précises) est utilisé, nous obtenons alors les probabilités conditionnelles $p(A|C)$ et $p(B|C) = 1 - p(A|C)$. Ceci est ce que nous appelons les dichotomies emboîtées précises.

En revanche, si les probabilités conditionnelles calculées sont des intervalles, alors nous sommes dans le cadre des probabilités imprécises décrit dans la Section 3.2. Dans ce cas, chaque nœud C est associé à un intervalle de probabilités $p(A|C) \in [\underline{p}(A \mid C); \overline{p}(A \mid C)]$, et par dualité nous avons $\underline{p}(B \mid C) = 1 - \overline{p}(A \mid C)$ et $\overline{p}(B \mid C) = 1 - \underline{p}(A \mid C)$. Nous pouvons considérer les dichotomies emboîtées imprécises comme une généralisation du cas précis, car nous obtenons des dichotomies emboîtées précises quand $\underline{p}(A \mid C) = \overline{p}(A \mid C)$ pour tout nœud C.

Nous montrons sur la Figure 5.2 une illustration d'un arbre de dichotomies imprécises avec les contraintes de probabilités conditionnelles en utilisant l'exemple de la reconnaissance d'obstacles.

Il est également intéressant de remarquer que les modèles locaux sont complètement indépendants. Ce qui signifie que, une fois la structure de l'arbre déterminée, le calcul des probabilités conditionnelles par des classifieurs de base peut être fait indépendamment et simultanément, à la fois pour l'apprentissage et le test. Même si nous ne faisons pas des copies de l'ensemble de données, nous pouvons quand même paralléliser le calcul pour les nœuds de l'arbre de la même profondeur étant donné qu'ils travaillent

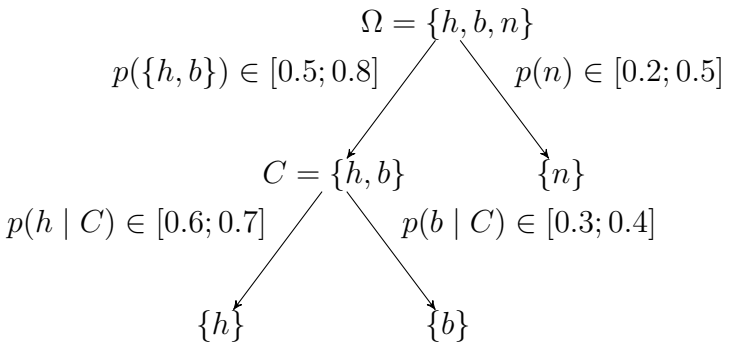


Figure 5.2: Arbre de dichotomies emboîtées utilisant les probabilités imprécises

sur des parties disjointes de l'ensemble des données. Cette propriété peut réduire considérablement le temps de calcul. Ainsi, les dichotomies emboîtées possèdent tous les avantages mentionnés précédemment concernant les techniques de décomposition.

5.2.1 Prise de décision avec les dichotomies emboîtées

L'inférence avec les dichotomies emboîtées suit la même méthodologie que celle décrite dans la Section 3.2.2. Il s'agit d'établir un ordre de préférence sur les classes en calculant les coûts espérés. Le calcul des coûts espérés est simplifié grâce à la structure de l'arbre de dichotomies.

Supposons que nous ayons un partitionnement $\{A, B\}$ d'un nœud C, et étant donné une fonction à valeurs réelles c définie sur $\{A, B\}$, le coût espéré du nœud C est défini comme :

$$\mathbb{E}_C[c] = \mathbb{E}_{p(\cdot|C)}[c] = p(A \mid C)c(A) + p(B \mid C)c(B), \tag{5.2}$$

L'Équation(5.2) reste vraie pour les dichotomies emboîtées imprécises (De Cooman et Hermans, 2008), dans le sens où les coût espéré inférieur $\underline{\mathbb{E}}_C(c)$ peut simplement être calculé par :

$$\underline{\mathbb{E}}_C[c] = \min \left(\begin{array}{l} \underline{p}(A \mid C)c(A) + \overline{p}(B \mid C)c(B); \\ \\ \overline{p}(A \mid C)c(A) + \underline{p}(B \mid C)c(B) \end{array} \right). \tag{5.3}$$

Le coût espéré supérieur du modèle global $\overline{\mathbb{E}}_\Omega$ est obtenu en remplaçant min par max dans l'Équation (5.3) puisque nous avons la dualité $\underline{\mathbb{E}}[c] = -\overline{\mathbb{E}}[-c]$. Nous notons que l'Équation (5.3) consiste à appliquer l'Équation (5.2) deux fois, ce qui signifie que les calculs avec des dichotomies emboîtées imprécises sont seulement deux fois plus coûteux que ceux avec les précises.

Nous notons également que, puisqu'il s'agit de problèmes binaires, le problème de minimisation du coût espéré qui était difficile à résoudre avec le NCC dans l'Equation (3.23) est maintenant facile à résoudre en énumérant les sommets de l'ensemble crédal, comme il n'y a que deux sommets possibles : $(\underline{p}(A \mid C), \overline{p}(B \mid C))$ et $(\overline{p}(A \mid C), \underline{p}(B \mid C))$.

Les probabilités conditionnelles $p(A|C)$ et $p(B|C)$ (et leurs bornes supérieures / inférieures) sont estimées par le classifieur de base. Si A, B sont des singletons (nœuds feuilles), alors $c(A), c(B)$ sont connus par la définition de leurs fonctions de coûts associées. Si A, B sont des nœuds internes, alors $c(A), c(B)$ correspondent aux coûts espérés des nœuds A, B qui peuvent être calculés de manière récursive. Par conséquent, le coût espéré du modèle global peut être obtenu facilement par récurrence rétrograde à partir des feuilles à la racine, même dans le cadre de probabilités imprécises (De Cooman et Hermans, 2008; Walley, 1991).

Si nous utilisons le critère de dominance par intervalle pour la prise de décision, alors la fonction c au niveau des feuilles correspond à la fonction de coût $c_{\hat{y}}$ d'une prédiction potentielle $\hat{y} \in \Omega$. Si nous utilisons la maximalité à la place, alors la fonction c au niveau des feuilles correspond à la différence de deux fonctions de coûts $c_{\hat{y}_1} - c_{\hat{y}_2}$ $(\hat{y}_1, \hat{y}_2 \in \Omega)$.

Nous pouvons déduire un algorithme récursif pour calculer $\mathbb{E}_\Omega$. Pour des raisons de clarté, nous écrivons uniquement l'algorithme dans le cas précis. L'Algorithme 1 est ainsi déduit à partir de l'Équation (5.2). Dans le cas imprécis, nous remplaçons simplement chaque déclaration *retourner* par la

formule correspondante dans l'Équation (5.3).

Algorithme 1 : Fonction $\mathbb{E}$ (calculant le coût espéré de c)

Entrées : C=nœud en cours, initialisé à Ω
$A, B \leftarrow$ nœuds descendants de C;
si *A et B sont singletons* **alors**
  retourner $p(A|C)c(A) + p(B|C)c(B)$
sinon si *A est singleton* **alors**
  `/* récursion sur le nœud` B`: calcul de` $\mathbb{E}(B)$ `*/`
  retourner $p(A|C)c(A) + p(B|C)\mathbb{E}(B)$
sinon si *B est singleton* **alors**
  `/* récursion sur le nœud` A`: calcul de` $\mathbb{E}(A)$ `*/`
  retourner $p(A|C)\mathbb{E}(A) + p(B|C)c(B)$
sinon aucun nœud descendant n'est singleton
  `/* récursion sur` A `et` B`: calcul de` $\mathbb{E}(A), \mathbb{E}(B)$ `*/`
  retourner $p(A|C)\mathbb{E}(A) + p(B|C)\mathbb{E}(B)$
fin

Exemple 11. *Nous montrons dans cet exemple ce que nous pouvons dire au sujet de la relation de préférence entre les classes b et h avec les probabilités conditionnelles imprécises indiquées sur la Figure 5.2. Avec le critère de maximalité, nous calculons le coût espéré* $\underline{\mathbb{E}}_\Omega[c_b - c_h]$ *(voir Figure 5.3) :*

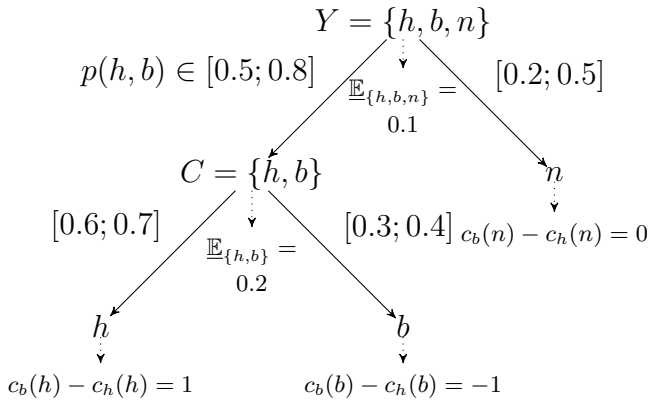


Figure 5.3: Exemple de calcul du coût espéré $\underline{\mathbb{E}}[c_b - c_h]$ avec les dichotomies emboîtées dans le cadre des probabilités imprécises

En utilisant l'Équation(5.3) *récursivement, nous calculons d'abord :*

$$\underline{\mathbb{E}}_{\{h,b\}}[c_b - c_h] = \min\Big(0.6 - 0.4; 0.7 - 0.3\Big) = 0.2$$

et ensuite,

$$\begin{aligned} & \underline{\mathbb{E}}_{\{h,b,n\}}[c_b - c_h] \\ = \ & \min\Big(0.2 \cdot 0.8 + 0 \cdot 0.2; 0.2 \cdot 0.5 + 0 \cdot 0.5\Big) \\ = \ & 0.1 > 0 \end{aligned}$$

Ainsi, nous pouvons déduire que la classe "humain" (h) est préférée à "bicyclette" (b) selon le critère de maximalité spécifié par la Définition (3.10)*.*

Nous pouvons aussi calculer

$$\underline{\mathbb{E}}_{\{p,b,n\}}[c_b] = \min\Big(0.6 \cdot 0.8 + 2 \cdot 0.2; 0.6 \cdot 0.5 + 2 \cdot 0.5\Big) = 0.88,$$

$$\overline{\mathbb{E}}_{\{p,b,n\}}[c_h] = \max\Big(0.4 \cdot 0.8 + 2 \cdot 0.2; 0.4 \cdot 0.5 + 2 \cdot 0.5\Big) = 1.2.$$

Nous n'avons pas $\overline{\mathbb{E}}_{\{p,b,n\}}[c_h] < \underline{\mathbb{E}}_{\{p,b,n\}}[c_b]$ *donc la même relation de préférence ne peut pas être trouvé avec le critère de dominance par intervalle spécifié par la Définition* (3.11)*.*

Comme l'Exemple 11 le montre, le calcul des coûts espérés inférieurs (supérieures) avec les dichotomies emboîtées imprécises est aussi simple qu'avec celui des dichotomies précises : les estimations des bornes inférieures et supérieures sont multiplicatives tout au long d'une branche. Cet avantage de l'utilisation dichotomies emboîtées le démarque d'autres classifieurs probabilistes imprécis, puisqu'elles peuvent gérer les coûts d'erreurs de classification unitaires et génériques avec le même ordre de complexité.

Remarque 2. *Quand il s'agit d'appliquer la dominance par intervalle avec des coûts unitaires, il est facile de calculer les probabilités inférieures et supérieures des différentes classes. Par exemple, sur la Figure 5.2, afin d'estimer les bornes de la probabilité a posteriori du problème multiclasse initial* $p(y = h)$*, nous avons besoin de calculer :*

$$\begin{aligned}\underline{p}(y=\{h\}) &= \underline{p}(y\in\{h,b\}\mid\Omega)\times\underline{p}(y=h\mid y\in\{h,b\})\\ &= 0.5\times 0.6=0.3\end{aligned}$$

$$\begin{aligned}\overline{p}(y=\{h\}) &= \overline{p}(y\in\{h,b\}\mid\Omega)\times\overline{p}(y=h\mid y\in\{h,b\})\\ &= 0.8\times 0.7=0.56\end{aligned}$$

Nous pouvons voir que pour calculer les bornes de la probabilité d'une classe donnée, nous avons juste besoin de multiplier les probabilités conditionnelles de la branche qui relie la classe donnée (nœud feuille) à la racine (Ω).

5.2.2 Déterminer la structure de l'arbre de dichotomies

Un point crucial des dichotomies emboîtées est le choix de la structure de dichotomie. En effet, il existe de nombreuses structures possibles d'arbres de dichotomies. Différentes structures conduisent à des problèmes de classification binaire différents, pour lesquels les estimations de probabilités conditionnelles pourront avoir des qualités variables, influençant ainsi directement la qualité du modèle global.

Par conséquent, un mauvais choix de la structure de l'arbre peut entraîner des estimations biaisées ou pauvres, dont l'accumulation peut conduire à de mauvaises prédictions. En autorisant les probabilités conditionnelles à être des intervalles, les inférences seront plus robustes et fiables puisque la quantité d'informations disponibles pour chaque modèle local est prise en considération. Nous pensons donc que les dichotomies emboîtées imprécises seront moins influencées par un mauvais choix de la structure. Nous évaluerons ceci avec les expériences dans le paragraphe suivant.

Différentes méthodes ont été proposées dans la littérature pour choisir la structure de dichotomies : quand nous avons des connaissances *a priori* ou des avis d'experts sur la structure de la classe, par exemple dans le cas de la classification ordinale (Huhn et Hullermeier, 2008), la structure de l'arbre de dichotomies peut être directement dérivée de cette information. Lorsque ce genre d'information est indisponible (ou non-existante), deux approches

ont été proposées pour faire face à ce problème : une solution naturelle est d'utiliser un ensemble de structures de dichotomies, tandis qu'une autre approche consiste à appliquer des traitements statistiques ou des procédés de fouille de données sur les données d'apprentissage, afin de reconstituer des relations entre les classes.

Forêts d'arbres de dichotomies

Une manière de résoudre le problème de sélection d'un arbre de dichotomies optimal est d'utiliser un ensemble Λ d'arbres de dichotomies possibles généré aléatoirement et uniformément (Eibe et Stefan, 2004). Dans ce cas, le processus de décision spécifié dans la Section 5.2.1 doit être adapté au fait que nous avons maintenant un ensemble de classifieurs, et que les résultats des différents classifieurs doivent être agrégés. Il y a principalement deux façons d'effectuer cette agrégation. La première, commune à toutes les méthodes ensemblistes de l'apprentissage, est d'utiliser des techniques de vote sur les prédictions produites par les différents classifieurs. Comme nous utilisons des classifieurs probabilistes, il y a une seconde approche qui est d'agréger les probabilités *a posteriori* estimées par les classifieurs. Puis les prédictions peuvent être déduites des estimations de probabilités agrégées.

Les techniques de vote sont largement utilisées dans la littérature pour les méthodes d'apprentissage basées sur l'utilisation d'un ensemble de classifieurs (Rokach, 2010), il est donc logique de prendre cette approche en considération. L'approche de base pour l'agrégation des prédictions est d'utiliser un système de votes à la majorité. Il a pour avantage de ne pas modifier les critères de décisions spécifiés dans la Section 3.2.2.

Définition 7 (Vote par majorité)**.** *Soit* $(\hat{Y}^{\lambda})_{\lambda\in\Lambda}$ *les prédictions obtenues par* $|\Lambda|$ *classifieurs issus de l'ensemble des arbres de dichotomies pour une observation* $\mathbf{x}$ *quelconque, nous définissons* $\hat{Y}$ *la prédiction finale issue d'un vote par majorité comme suit*

$$\hat{Y} = \left\{ y \in \Omega : \sum_{\lambda\in\Lambda} \mathbb{1}_{\hat{Y}^{\lambda}}(y) > \frac{|\Lambda|}{2} \right\}. \tag{5.4}$$

$\hat{Y}$ contient l'ensemble des classes qui sont prédites par plus de la moitié des classifieurs. Cela signifie que chaque votant (classifieur) peut choisir de voter pour plus d'une classe. Par conséquent, cette démarche autorise les prédictions imprécises qui se produisent quand il y a plus d'une classe pour laquelle la majorité des votants sont d'accord. Cependant, cette approche fait l'hypothèse implicite que la plupart des votants sont compétents pour prédire correctement, sinon les résultats pourraient être affectés par des classifieurs de piètres performances. Dans notre cas, nous savons que certaines structures peuvent s'avérer moins raisonnables que d'autres (par exemple, il est logique de regrouper "humain" et "bicyclette" ensemble pour le problème de reconnaissance d'obstacles, et les autres structures sont moins intuitives), donc il sera intéressant de voir si cette technique donne de bonnes performances dans les expériences.

D'autre part, nous pouvons agréger les estimations de probabilités à la place (approche choisie par Eibe et Stefan (2004)), ou plus précisément, les coûts espérés dans notre cas. Chaque classifieur dérivé d'un arbre de dichotomies estime une distribution de probabilités p^λ et est donc associé à un coût espéré $\mathbb{E}^\lambda = \mathbb{E}[\ldots|p^\lambda]$. Nous pouvons regrouper ces coûts espérés en un seul afin d'inférer une prédiction. Pour cela, nous pouvons par exemple utiliser la moyenne arithmétique des coûts espérés comme agrégat, ce qui équivaut à calculer la somme des coûts espérés et examiner son signe. Les critères de décision peuvent être exprimés comme suit :

Définition 8 (Agrégation des coûts espérés)**.** *Soit deux prédictions potentielles $\hat{y}_1, \hat{y}_2$ et un ensemble de classifieurs définis par les arbres de dichotomies Λ, le critère de maximalité peut être appliqué à l'agrégation des coûts espérés*

$$\hat{y}_1 \succ_{\mathcal{M}} \hat{y}_2 \Leftrightarrow \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \underline{\mathbb{E}}^\lambda[\hat{y}_2 - \hat{y}_1] > 0. \tag{5.5}$$

De même, le critère de dominance par intervalle devient

$$\hat{y}_1 \succ_{\mathcal{ID}} \hat{y}_2 \Leftrightarrow \sum_{\lambda \in \Lambda} \overline{\mathbb{E}}^\lambda[\hat{y}_1] < \sum_{\lambda \in \Lambda} \underline{\mathbb{E}}^\lambda[\hat{y}_2]. \tag{5.6}$$

Contrairement au vote à la majorité, cette approche prend la force des estimations (*i.e.*, les valeurs de coûts espérés) en considération. Par exemple,

si un classifieur donne une estimation du coût espéré très élevée pour une préférence (donc, en un sens, nous pouvons dire qu'il est sûr de son estimation), puis un autre classifieur indique un léger désaccord (un coût espéré légèrement négatif) sur la même préférence, alors le second n'annulera pas le premier lors de l'agrégation. Cette particularité peut être bénéfique ou néfaste suivant la capacité du classifieur de base pour donner des estimations de probabilités correctes.

Nous pouvons voir que les deux techniques d'agrégation ont leurs avantages et sont liées à des contextes spécifiques. Nous donnerons une comparaison de leur performance dans les expériences. Bien sûr, il est également possible d'aborder une approche plus générale en choisissant par exemple une agrégation du type $\sum_{\lambda\in\Lambda}\mu_\lambda\underline{\mathbb{E}}^\lambda[\hat{y}_2-\hat{y}_1]$ où les coefficients $(\mu_\lambda)_{\lambda\in\Lambda}$ sont génériques et représentent la force de chaque modèle. Mais le choix de ces coefficients est un sujet de recherche en soi (Corani et Zaffalon, 2008), nous allons donc nous restreindre aux approches que nous avons détaillées. Ceci n'est pas particulièrement gênant, puisque notre objectif premier n'est pas d'avoir un classifieur entièrement optimisé, mais plutôt d'étudier les intérêts des dichotomies emboîtées imprécises.

Choix d'une structure unique de dichotomies

Nous voyons qu'une forêt de dichotomies emboîtées peut être utilisée pour éviter le problème de trouver la structure optimale de l'arbre de dichotomies, mais il ne permet pas de savoir comment notre approche se comporte quand une structure unique d'arbre de dichotomies est utilisée. En effet, une structure unique d'arbre de dichotomies présente des avantages que la forêt de dichotomies ne possède pas : d'abord une structure unique est beaucoup plus facile à interpréter et à analyser. En outre, cette structure pourrait également être fournie ou dérivée de certaines connaissances d'experts dans la pratique (par exemple, la classe peut avoir une structure naturelle en génétique, biologie ou reconnaissance d'images). Ainsi, il est intéressant d'évaluer la performance et les comportements d'un arbre de dichotomies unique par rapport à une approche ensembliste.

Si nous ne disposons pas de connaissances *a priori*, dans ce cas il est possible de construire l'arbre de dichotomies en utilisant des traitements statistiques sur les données d'apprentissage. Lorena et de Carvalho (2010) proposent de dériver les structures d'arbres binaires avec des mesures de séparabilité sur les classes. L'idée de base est de regrouper les classes en fonction de leur *similarité statistique* ou *distance*, afin de construire des problèmes binaires dont les sous-ensembles de classes sont bien séparés entre eux. Les classes peuvent être regroupées en utilisant des techniques telles que le regroupement hiérarchique ou les *k-means*. Cependant, Lorena et de Carvalho (2010) et des tests effectués de notre côté avec différentes techniques de regroupement révèlent qu'il n'y a pas un traitement statistique qui mènent toujours à la structure optimale. Les choix de mesures de séparabilité et de techniques de regroupement varient selon les jeux de données. Par conséquent, cette approche est plus appropriée dans les applications pratiques, où un jeu de données spécifique doit être étudié.

Une autre approche plus heuristique est de reprendre la forêt d'arbres de dichotomies et d'effectuer une validation croisée sur les données d'apprentissage afin de choisir la structure ayant le pouvoir prédictif le plus élevé. Dans le cas où aucune information *a priori* n'est disponible, cette approche pourrait être préférée. L'inconvénient est que, quand le nombre de classes est grand, le nombre de structures possibles pour l'arbre de dichotomies croît de manière exponentielle. Eibe et Stefan (2004) évalue le nombre d'arbres de dichotomies possibles à $(2K-3)!!$ (factorielle impaire, *e.g.*, $1 \times 3 \times 5 \ldots$) pour un problème à K classes. Même si nous mettons les modèles binaires en mémoire (pour éviter les répétitions de calcul), il y a quand même $\frac{3^K-(2^{K+1}-1)}{2}$ sous-problèmes binaires possibles. Cette approche n'est donc pas adaptée pour l'étude d'un jeu de données pratique avec un nombre relativement important de classes, puisqu'il est peu probable que la bonne structure soit dans la forêt.

5.3 Expériences

Nous avons montré qu'un classifieur construit en combinant les dichotomies emboîtées avec les probabilités imprécises présente certains avantages théoriques et exigences spécifiques. Il est maintenant intéressant d'évaluer les impacts pratiques de ces facteurs. A travers les expériences, nous nous concentrons en particulier sur deux points : le premier est d'étudier si les dichotomies emboîtées imprécises augmentent véritablement la fiabilité et la prudence des prédictions par rapport à des approches standards. Le second est d'évaluer l'intérêt des dichotomies emboîtées imprécises par rapport aux autres classifieurs utilisant les probabilités imprécises.

Il est important de noter que nous cherchons à évaluer l'effet de l'introduction des dichotomies emboîtées uniquement. Ainsi, pour isoler et identifier ses influences, nous n'allons pas intégrer de coûts génériques dans ces expériences, puisque l'introduction de coûts génériques va modifier à la fois la phase de décision et le critère d'évaluation. Ces coûts peuvent donc aussi influencer les résultats, ce qui risque de fausser nos interprétations.

5.3.1 Cadre expérimental

Nous allons conduire les expériences sur 14 jeux de données du dépôt de données UCI (Bache et Lichman, 2013), dont les détails sont donnés dans le Tableau 5.1. Nous les montrons dans l'ordre croissant du nombre de classes.

Nous notons que ce sont des jeux de données d'apprentissage à usage général : il n'y a pas de structure prédéfinie pour les classes (pas d'information *a priori*, ni de connaissances d'experts) et de coûts d'erreur de classification spécifiques. Il est donc normal d'utiliser des coûts unitaires pour l'apprentissage et l'évaluation.

Classifieur de base et discrétisation

Dans nos expériences, nous voulons un classifieur de base probabiliste qui peut être étendu pour gérer les probabilités imprécises. Ceci nous permettra de comparer la version précise et l'imprécise, afin que nous puissions

Nom	Attributs (C)ontinus/(D)iscrets	Nombre d'instances de données	Nombre de classes
balance-scale	D	625	3
wine	C	178	3
iris	C	150	3
car	D	1728	4
lymph	D	148	4
grub-damage	C	155	4
nursery	D	12960	5
page-blocks	C	5473	5
glass	C	214	6
zoo	D	101	7
segment	C	2310	7
ecoli	C	336	8
pendigits	C	10992	10
soybean	D	562	15

Table 5.1: Les jeux de données utilisés pour les expériences sur les dichotomies emboîtées

évaluer l'impact d'être imprécis dans les dichotomies emboîtées. Pour cette raison, nous utilisons le classifieur crédal naïf (NCC) mentionné dans la Section 3.3.3 comme classifieur de base. Et nous mettons le paramètre de l'imprécision s du NCC à 1 pour toutes ces expériences.

Comme le NCC ne peut pas gérer nativement les variables continues, ces dernières dans les jeux de données ont été discrétisées. Nous avons choisi de discrétiser toutes les variables continues en divisant leur domaine en 5 intervalles de largeur égale. Nous n'utilisons pas de méthode de discrétisation supervisée telle que proposée par Fayyad et Irani (1993), puisque les classes impliquées changent entre le problème multiclasse initial et chaque sous-problème binaire. Pour rester le plus équitable possible, nous avons donc utilisé la même discrétisation pour toutes les approches dans les expériences.

Critère d'évaluation

Comme nous travaillons avec des coûts unitaires, les travaux de Zaffalon et collab. (2012) sur la justesse affaiblie selon l'utilité que nous avons introduit dans la Section 4.1.1 semble être un bon choix. Il nous permet de comparer à la fois des classifieurs déterminés et indéterminés en équilibrant la pénalisation de l'imprécision et la récompense de la prudence.

Nous avons choisi d'utiliser la fonction d'utilité u_{65} donnée par l'Equation (4.2) comme critère d'évaluation. Nous rappelons ici son expression :

$$g(x) = -0.6x^2 + 1.6x.$$

Elle consiste à récompenser légèrement plus la prudence par rapport à la justesse affaiblie, comme l'illustre la Figure 5.4.

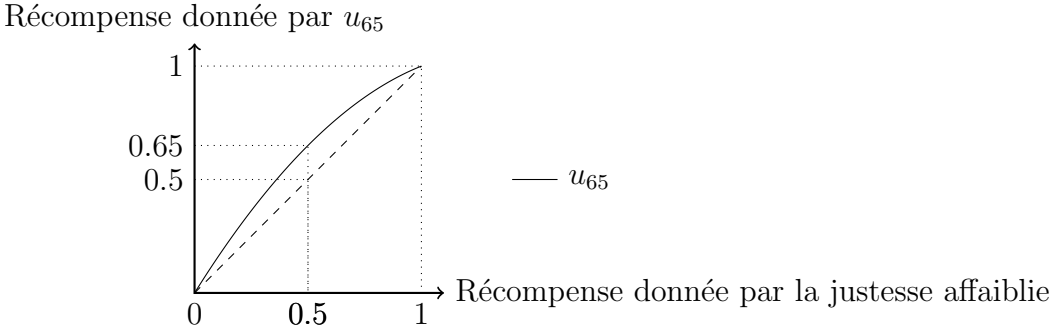


FIGURE 5.4: Fonction d'utilité quadratique u_{65} dérivée de la justesse affaiblie

Ceci nous permet de calculer le score de performance pour chaque jeu de données d'évaluation $\mathcal{D}$

$$u_{65}(\mathcal{D}) = \frac{1}{N} \sum_{(\mathbf{x},y)\in\mathcal{D}} g\left(\frac{\mathbb{1}_{y\in f(\mathbf{x})}}{|f(\mathbf{x})|}\right),$$

où N est le nombre d'instances de données, et $\mathbb{1}_{y\in f(\mathbf{x})}$ est la fonction indicatrice qui vaut 1 si la vérité y est incluse dans la prédiction $f(\mathbf{x})$ faite par le classifieur et 0 sinon.

Choix de la structure et du nombre d'arbres de dichotomies

Nous allons évaluer et comparer les performances des approches ensemblistes et "structure unique" (*cf.* Section 5.2.2). Les deux approches ensemblistes (par vote ou par agrégation des coûts espérés) seront examinées.

Cependant, comme nous utilisons un ensemble de jeux de données génériques, il n'est pas adapté de choisir l'approche utilisant des mesures de séparabilité pour déterminer une structure unique de dichotomies. Nous choisissons donc d'utiliser l'approche heuristique : nous sélectionnons l'arbre de dichotomies avec la meilleure performance avec une validation croisée (10 répétitions) parmi la forêt d'arbres.

Il faut également fixer le nombre d'arbres à utiliser. Eibe et Stefan (2004) ont suggéré que $|\Lambda = 20|$ est suffisant pour obtenir des résultats stables dans la pratique. Nous avons tendance à confirmer ce résultat à l'aide de nos expériences préliminaires. Nous avons effectué un test signé des rangs de Wilcoxon unilatéral sur le score u_{65} des forêts avec 20 et 50 arbres sur les 14 jeux de données. L'augmentation de la performance n'est pas statistiquement significative avec une p-valeur de $0,05$. Cependant, si nous limitons notre analyse aux résultats sur les ensembles de données avec un nombre de classes élevé (égal ou supérieur à 5), on observe un gain significatif de performances lorsque nous travaillons avec 50 au lieu de 20 arbres. Par conséquent, cela semble être une question dépendant des jeux de données. Dans la pratique, il serait sage de paramétrer le nombre d'arbres en fonction du nombre de classes. Nous allons utiliser $\lambda = 50$ pour nos expériences pour avoir une meilleure garantie de convergence des résultats. Des investigations supplémentaires seraient nécessaires pour établir une règle empirique sur la corrélation entre la convergence et le nombre d'arbres nécessaires.

5.3.2 Comparaison de performances prédictives

Dans les expériences dont les résultats sont donnés dans le Tableau 5.2, nous comparons cinq méthodes :

– **ND+NBC** : dichotomies emboîtées avec le classifieur de Bayes naïf comme classifieur de base, la structure de l'arbre de dichotomie est

déterminée avec l'approche heuristique détaillée dans la Section 5.2.2.
- **NCC** : le classifieur crédal naïf utilisé directement comme une référence.
- **DC** : méthode développée par del Coz et collab. (2009) (détaillée dans la Section 3.1.2) qui permet de produire des prédictions indéterminées à partir des estimations de probabilités précises. Nous utilisons la méthode ND+NBC pour fournir ces estimations précises.
- **ND+NCC** : méthode identique à ND+NBC, mais avec le NCC comme classifieur de base, ainsi les prédictions peuvent être imprécises.
- $\mathbf{F_v}$: ensemble de dichotomies emboîtées avec le NCC où la technique de vote à la majorité est utilisée pour l'agrégation des prédictions.
- $\mathbf{F_m}$: même chose que ci-dessus mais avec la moyenne des coûts espérés comme agrégat.

Cela nous permet d'effectuer trois types de comparaisons :
- le classifieur déterminé (ND+NBC) contre les classifieurs indéterminés
- les dichotomies emboîtées imprécises (ND+NCC, F_v et F_m) contre la version multiclasse imprécise sans dichotomies (NCC)
- les prédictions indéterminées provenant des dichotomies emboîtées imprécises contre celles provenant des dichotomies emboîtées précises (DC)

Pour faciliter la comparaison, nous donnons aussi le rang de chaque score (s'il y a des éléments *ex æquo*, alors le rang moyen leur est attribué). Afin de rendre les résultats plus lisibles, nous surlignons également le meilleur score pour chaque jeu de données.

Afin de vérifier statistiquement les différences entre les méthodes, nous suivons l'approche suggérée par Demšar (2006) et nous appliquons le test de Friedman (Friedman, 1937, 1940) sur les rangs obtenus pour chaque jeu de données. Nous trouvons une valeur de $2,08$ pour la statistique de khi-deux avec cinq degrés de liberté, de sorte que la p-valeur est de $0,84$ et nous ne pouvons donc pas rejeter l'hypothèse nulle. Cela signifie que toutes les méthodes ont des performances comparables en terme de justesse de prédictions et les différences ne sont pas statistiquement significatives.

	score u_{65} sous forme "pourcentage (rang)"					
	ND+NBC	NCC	DC	ND+NCC	F_v	F_m
balance	90.72 (5)	90.78 (3)	84.39 (6)	**90.88(1)**	90.77 (4)	90.82 (2)
wine	95.51 (6)	97.07 (2.5)	95.84 (5)	96.13 (4)	97.07 (2.5)	**97.16(1)**
iris	93.33 (5)	93.27 (6)	93.5 (3)	**93.7(1)**	93.5 (3)	93.5 (3)
car	88.37 (2)	86.16 (4)	86.35 (3)	**89.03(1)**	85.85 (5)	85.46 (6)
lymph	82.64 (2)	70.07 (6)	**84.36(1)**	82 (3)	73.46 (5)	75.48 (4)
grub-d.	47.52 (6)	52.48 (2)	49.86 (5)	50.44 (4)	52.2 (3)	**52.9(1)**
nursery	91.54 (2)	90.46 (4)	88.17 (6)	**91.73(1)**	90.34 (5)	90.57 (3)
page-b.	91.67 (3)	91.17 (6)	**92.03(1)**	91.83 (2)	91.56 (5)	91.61 (4)
glass	53.74 (2)	51.38 (6)	**58.7(1)**	51.81 (5)	53.38 (3)	52.37 (4)
zoo	91.73 (2)	83.79 (5)	**92.38(1)**	85.22 (3)	84.68 (4)	81.6 (6)
segment	86.58 (5)	**89.92(1)**	86.84 (3.5)	86.4 (6)	86.84 (3.5)	88.34 (2)
ecoli	80.95 (4)	79.47 (5)	**82.22(1)**	78.41 (6)	81.24 (2.5)	81.24 (2.5)
pendigits	81.41 (6)	**85.81(1)**	82.13 (4)	81.42 (5)	82.73 (3)	84.48 (2)
soybean	87.37 (5)	**90.26(1)**	87.62 (4)	82.99 (6)	89.27 (2)	88.01 (3)
rang moyen	3.93	3.75	3.18	3.43	3.61	3.11

Table 5.2: Comparaison des scores u_{65} des différentes méthodes

Cependant, malgré le fait que les rangs se compensent en moyenne, nous pouvons remarquer que nos approches ont généralement un comportement très différent par rapport à la méthode de del Coz et collab. (2009) : la différence des rangs (et de la performance) sur un jeu de données est souvent très important.

Il est aussi intéressant de noter que l'approche avec une structure unique de l'arbre de dichotomies semble très performante quand il y a peu de classes (égale ou inférieure à 5), et devient moins efficace lorsque les jeux de données ont plus de 5 classes. Sachant que le nombre d'arbres de dichotomies possibles est de 105 pour 5 classes et 945 pour 6 classes (cf. Section 5.2.2), cela est donc clairement dû au nombre trop faible (50) d'arbres utilisées pour déterminer la structure "optimale" lorsque le nombre de classes est grand. En revanche, cela montre qu'utiliser un arbre unique de dichotomies peut donner des très bons résultats si une structure *optimale* de dichotomies peut être établie, ce qui suggère que les dichotomies emboîtées sont bien adaptées dans le cas où les connaissances d'experts ou les informations *a priori* sont présentes et peuvent être traduites en un arbre de dichotomies.

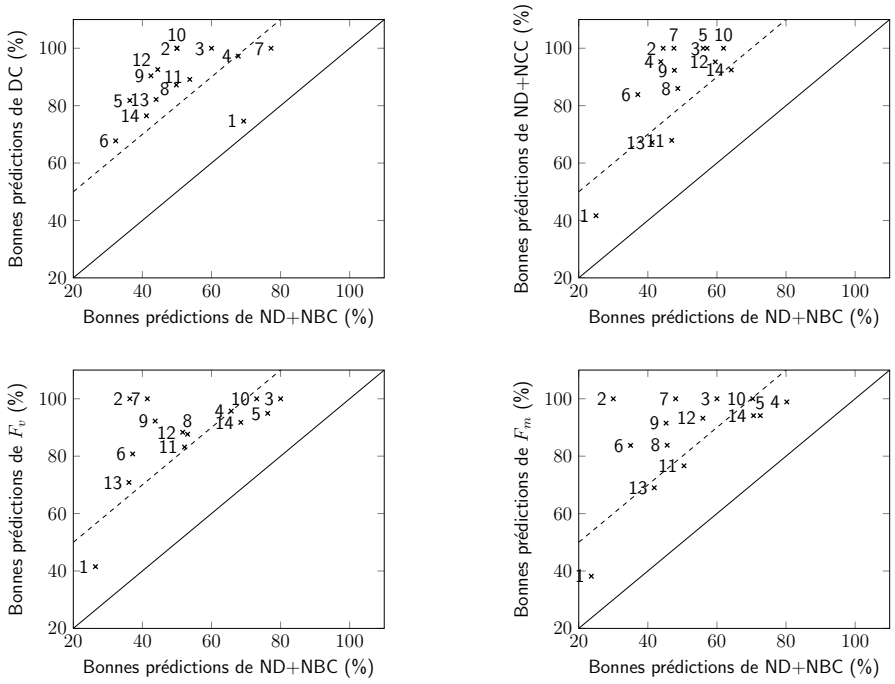
D'un autre côté, nous observons que les approches ensemblistes ont des performances relativement indépendantes du nombre de classes, ceci est dû au fait que les biais issus de chaque arbre peuvent se compenser les uns avec les autres, rendant ainsi possible de n'utiliser qu'un ensemble limité d'arbres de dichotomies pour avoir de bonnes performances.

Au final, nous notons également que les performances de nos approches imprécises ont encore beaucoup de possibilités d'amélioration : dans le cas de la forêt, il est par exemple possible d'améliorer la procédure d'agrégation (similairement à ce qui est fait dans d'autres travaux basant sur l'utilisation d'un ensemble de modèles (Cesa-Bianchi et collab., 1997; Corani et Zaffalon, 2008)) ; dans le cas d'une structure unique, nous pouvons améliorer séparément le classifieur de base de chaque problème binaire, par exemple en optimisant la discrétisation ou les paramètres du classifieur, ou la structure de dichotomies (en utilisant des mesures de séparabilité adaptées).

5.3.3 Gain de justesse sur les instances "difficiles" à prédire

L'objectif principal des classifieurs indéterminés est de faire des prédictions indéterminées incluant la vraie classe sur les cas (et idéalement seulement sur ceux-là) où le classifieur déterminé échoue. Pour montrer que tel est bien la tendance ici, la Figure 5.5 illustre, pour chaque classifieur indéterminé (DC, ND+NCC, F_v et F_v) et sur les instances de données où les prédictions indéterminées sont faites par chaque classifieur, les pourcentages de fois où la vraie classe est incluse dans les prédictions, à la fois pour le classifieur indéterminé et la version déterminée correspondante (ND+NBC).

Alors qu'il est trivial que le classifieur indéterminé ait plus de bonnes prédictions que la version déterminée, il est cependant important de noter que le pourcentage de bonnes prédictions pour le classifieur déterminé, sur ces instances de données où des prédictions indéterminées sont faites par leurs versions indéterminées, est généralement beaucoup plus faible que ce que celui sur la totalité des instances de données (voir Tableau 5.2). Les cas les plus illustratifs sont les points autour du jeu de données *wine* (point



1 = balance-scale	2 = wine	3 = iris	4 = car	5 = lymph
6 = grub-damage	7 = nursery	8 = page-blocks	9 = glass	10 = zoo
11 = segment	12 = ecoli	13 = pendigits	14 = soybean	

Figure 5.5: Pourcentage de bonnes prédictions de "ND+NBC" contre celui des différents classifieurs indéterminés sur les instances de données où des prédictions indéterminées sont faites par ces-derniers.

numéro 2 dans la Figure 5.5), où le score des deux classifieurs déterminé et indéterminé sont au-dessus 90% sur la totalité des instances de données (score de bonnes prédiction est supérieur au score u_{65}), mais il chute pour le classifieur déterminé à environ 50% sur les instances de données où des prédictions indéterminées sont faites par les classifieurs indéterminés.

Cette baisse de la performance pour le classifieur déterminé, lorsque nous nous restreignons au cas où les prédictions indéterminées sont faites, indique que ce sont en effet les cas "difficiles à classer" pour le classifieur déterminé. Ainsi, l'utilisation des estimations indéterminées est judicieuse dans ce cas.

Nous pouvons également noter que tous les quatre classifieurs indéter-

minés étudiés partagent cette propriété. Toutefois, si nous considérons le nombre de jeux de données où le gain de bonnes prédictions est à plus de 30% (points au-dessus de la ligne pointillée sur la Figure 5.5), alors il semble que DC (11 jeux de données au-dessus de la ligne pointillée) et ND+NCC (10 jeux de données au-dessus) réussissent un peu mieux à trouver ces cas "difficile à classer" que les deux méthodes ensemblistes (8 pour "F_v" et 7 pour "F_m") .

En outre, il est intéressant de noter que tandis que les positions relatives des points sont similaires pour les trois approches de dichotomies emboîtées imprécises, elles sont très différentes de celles de la méthode développée par del Coz et collab. (2009). Cela signifie que les diverses approches choisissent d'être indéterminées sur des instances de données différentes, et les instances de données "difficile à classer" ne sont pas nécessairement les mêmes pour les diverses approches, *i.e.* la difficulté à classer ne semble pas être inhérente aux jeux de données , mais plutôt aux classifieurs.

5.3.4 Comparaison de niveaux d'indétermination

Similairement au Tableau 5.2, nous illustrons maintenant les pourcentages de prédictions indéterminées faites par les divers classifieurs indéterminés dans le Tableau 5.3.

Dans le Tableau 5.3, nous pouvons voir que nos méthodes sont toujours plus déterminées que le NCC. Couplé avec le Tableau 5.2, nous pouvons maintenant dire que nos approches surpassent le NCC à la fois en termes de justesse et de niveau de détermination. Ainsi, l'utilisation d'arbres de dichotomies apporte des améliorations nettes par rapport à l'approche directe avec les probabilités imprécises.

Nous pouvons encore une fois remarquer que le comportement de nos approches avec les dichotomies emboîtées imprécises est très différent de l'approche de del Coz et collab. (2009). En effet, même si le pouvoir prédictif des deux approches est très similaire, le niveau d'indétermination est très différent pour presque tous les jeux de données. Elles semblent même être antagonistes : à chaque fois qu'une approche a un niveau d'indétermination

	Pourcentages de prédictions indéterminées (rang)				
	NCC	DC	ND+NCC	F_v	F_m
balance-scale	10,4 (4)	30,24 (5)	**7, 68(1)**	8,48 (2)	8,8 (3)
wine	6,18 (4,5)	**2, 25(1)**	5,06 (2)	6,18 (4,5)	5,62 (3)
iris	4 (4)	3,33 (2)	4,67 (5)	3,33 (2)	3,33 (2)
car	5,38 (4)	33,91 (5)	**3, 7(1)**	4,05 (2)	5,27 (3)
lymph	50 (5)	**7, 43(1)**	16,89 (2)	39,86 (4)	34,46 (3)
grub-damage	52,9 (5)	**20(1)**	40 (2)	50,32 (3)	51,61 (4)
nursery	1,07 (3)	27,48 (5)	1,1 (4)	**0, 54(1)**	0,59 (2)
page-blocks	1,86 (3)	5,28 (5)	2,74 (4)	1,48 (2)	**1, 24(1)**
glass	61,21 (5)	34,11 (2)	**30, 37(1)**	48,13 (3)	49,53 (4)
zoo	23,76 (3)	**1, 98(1)**	20,79 (2)	25,74 (4)	26,73 (5)
segment	**3, 07(1)**	7,58 (5)	3,51 (2)	4,89 (4)	4,63 (3)
ecoli	22,62 (5)	**8, 04(1)**	12,5 (2)	17,86 (4)	17,56 (3)
pendigits	1,16 (2)	8,01 (5)	1,61 (4)	**0, 66(1)**	1,17 (3)
soybean	8,72 (2)	**3, 02(1)**	18,86 (5)	12,99 (3)	15,12 (4)
rang moyen	3.61	2.86	2.64	2.82	3.07

Table 5.3: Pourcentages de prédictions indéterminées faites par les classifieurs indéterminés

bas, l'autre a un niveau beaucoup plus élevé. Cependant, les différences sont compensées en moyenne, de sorte qu'il n'y a pas de différence statistiquement significative à la fin.

Le Tableau 5.3 montre également que le niveau d'indétermination peut aller de très faible (autour de 1% pour *pendigits*) à très élevé (autour de 50 % pour *grub-damage*), ce qui montre que les probabilités imprécises et l'approche indéterminée avec les probabilités précises sont tous les deux capables d'adapter le niveau d'indétermination selon le jeu de données. En outre, il n'y a pas de corrélation apparente entre la performance prédictive et le niveau d'indétermination : un niveau plus élevé d'indétermination ne signifie pas forcément un score u_{65} plus élevé, ce qui soutient le fait que u_{65} reste un critère équitable pour comparer classifieurs déterminés et indéterminés.

Enfin, ces expériences nous donnent également quelques idées sur les différentes façons de construire des arbres de dichotomies. Nous pouvons voir que l'utilisation d'un arbre de dichotomie unique (ND+NCC) permet d'obtenir plus de prédictions déterminées tout en gardant un taux élevé

de bonnes prédictions, puisque le niveau d'indétermination moyen est inférieur à celui des forêts et que les performances sont similaires. Ceci suggère que l'utilisation d'un seul arbre de dichotomies induit par les connaissances d'experts ou par la structure de classe ne diminuera pas nécessairement les performances par rapport à des approches ensemblistes plus lourdes en termes de calcul. D'autre part, l'utilisation des forêts d'arbres de dichotomies permet de produire des prédictions plus prudentes, et peut être très efficace pour traiter les problèmes où il n'y a pas d'informations *a priori* disponibles sur la structure des classes.

Maintenant que nous avons validé l'intérêt de combiner les dichotomies emboîtées avec les probabilités imprécises, il nous est désormais possible d'effectuer des expériences pour évaluer l'intérêt de combiner le cadre des probabilités imprécises avec celui de classification sensibles aux coûts. En effet, comme nous l'avons vu, les dichotomies emboîtées peuvent gérer les problèmes avec des coûts d'erreurs de classification génériques avec la même complexité calculatoire qu'avec des coûts unitaires.

Chapitre 6

Le cas des données ordinales

Nous avons vu dans les chapitres précédents, de manière indépendante, les avantages et les spécificités des prédictions indéterminées (plus particulièrement du cadre des probabilités imprécises) et de l'utilisation des coûts génériques d'erreur de classification. Nous avons également vu que les dichotomies emboîtées permettent d'avoir un classifieur reliant ces deux cadres et qui conserve à la fois une bonne performance et des avantages spécifiques (calculabilité, flexibilité). Cependant il reste deux points à examiner : premièrement il convient de caractériser, à l'aide des expériences, le comportement et les spécificités de la formule de coût affaibli généralisé que nous avons proposée dans le Chapitre 4. Deuxièmement, une fois la formule validée et caractérisée, il sera intéressant de l'appliquer pour l'évaluation des classifieurs. Nous comparerons dans un premier temps l'approche d'inférence des prédictions indéterminées par probabilités précises avec l'approche par probabilités imprécises. Dans un second temps, nous évaluerons le classifieur construit à l'aide des dichotomies emboîtées imprécises avec des coûts génériques.

6.1 Présentation du cadre expérimental

Le coût affaibli généralisé que nous avons défini sert surtout à évaluer les problèmes dans lesquels une structure de coût peut être dérivée de l'aversion au risque de l'utilisateur, mais il est difficile de trouver des jeux de données qui soient fournis naturellement avec des coûts d'erreurs prédéterminées. Nous utilisons donc des jeux de données ordinales pour nos expériences, où le coût est induit par la structure de l'espace des classes. Pour ces données, l'ensemble Ω fini d'étiquettes possibles est naturellement ordonné. Par exemple, les avis sur des films peuvent être exprimés en utilisant les labels suivants : *Très-Mauvais*, *Mauvais*, *Moyen*, *Bon*, *Très-Bon* qui sont ordonnés du pire avis au meilleur. Cela nous donnera un moyen facile d'établir une métrique sur les classes.

Nom	Nb. instances	Nb. attributs	Nb. classes
ERA	1000	5	9
ESL	488	5	9
LEV	1000	5	5

TABLE 6.1: Détails sur les jeux de données ordinales

6.1.1 Données utilisées et traitements associés

Les expériences seront menées sur trois jeux de données ordinales du dépôt de données UCI (Bache et Lichman, 2013), dont les détails sont donnés dans le Tableau 6.1. Comme notre objectif n'est pas de valider numériquement des modèles, mais d'étudier les comportements de notre formule, l'utilisation de seulement 3 jeux de données n'est pas problématique. Par ailleurs, puisque nous allons continuer à nous servir du NCC qui ne peut traiter les variables continues de façon native, une discrétisation à fréquence égale en 5 intervalles est utilisée. Nous n'utilisons pas de techniques supervisées pour la discrétisation pour la même raison que dans la Section 5.3.

Pour rendre les résultats plus fiables, nous utilisons également une validation croisée à 5 blocs que nous utiliserons parfois de deux manières : soit avec 4/5 des données pour l'apprentissage et 1/5 pour le test, soit avec 1/5 pour l'apprentissage et 4/5 pour le test. Ceci nous permet d'étudier le comportement des classifieurs indéterminés quand le nombre d'instances de données pour l'apprentissage varie. Sauf mention spéciale, nous utiliserons le premier cadre avec 4/5 des données pour l'apprentissage.

6.1.2 Norme ℓ_1 comme mesure de coûts et spécificités des données ordinales

Comme les classes dans ces jeux de données ordinales sont ordonnées, nous pouvons utiliser la distance en norme ℓ_1 entre les classes comme fonction de coûts. Soit l'espace des classes ordonnées $\Omega = \{\omega_1, \ldots, \omega_K\}$, alors pour ω_i et ω_j deux classes de l'espace Ω, le coût $c_{\omega_i}(\omega_j)$ de prédire la classe ω_i quand ω_j est vraie et le coût $c_{\omega_j}(\omega_i)$ de prédire ω_j quand ω_i est vraie

sont tous les deux définis par :

$$c_{\omega_i}(\omega_j) = c_{\omega_j}(\omega_i) = |i - j|.$$

Bien sûr, d'autres normes ℓ_k avec $k \neq 1$ pourraient également être utilisées pour des raisons spécifiques selon les contextes. Dans nos expériences, nous choisissons d'utiliser ℓ_1, vu que dans le cas des probabilités standards, elle conduit à prendre comme prédiction la médiane de la distribution, qui a l'avantage de ne dépendre que de l'ordre des valeurs $\omega_1, \ldots, \omega_K$. Nous donnons une illustration de la matrice de coûts produite avec le Tableau 6.2 avec l'exemple de l'évaluation des films. La partie des coûts pour les prédictions indéterminées peut ainsi être déduite en fonction de la formule de coût affaibli généralisé choisie.

Vérités / Prédictions	Très-Mauvais	Mauvais	Moyen	Bon	Très-Bon
Très-Mauvais	0	1	2	3	4
Mauvais	1	0	1	2	3
Moyen	2	1	0	1	2
Bon	3	2	1	0	1
Très-Bon	4	3	2	1	0

Table 6.2: Matrice de coûts pour les prédictions déterminées de l'évaluation des films

Grâce à cet ordre spécifique des classes, nous pouvons également proposer un moyen de restreindre l'espace de prédictions à un sous-ensemble de 2^Ω dans le cas des classifieurs indéterminés utilisant le cadre des probabilités standards (*cf.* Section 3.1). En effet, étant donné les étiquettes {*Mauvais*, *Moyen*, *Bon*}, le fait de regrouper *Mauvais* et *Bon* et de laisser *Moyen* de côté semble contre-intuitif, et il semble naturel de n'autoriser à prédire que des classes contiguës quand une prédiction indéterminée doit être faite. Cependant cette hypothèse, naturelle en apparence, n'est pas toujours vraie en pratique. En effet, pour un film controversé, il se peut qu'il y ait deux grandes tendances qui se contredisent (par exemple *Très Mauvais* et *Très bon*). Par conséquent, la restriction aux classes contiguës peut entraîner une

certaine perte d'information, qui peut être précieuse surtout lorsque nous essayons de construire un classificateur fiable.

6.1.3 Construction de deux classifieurs de référence

Comme discuté dans le Chapitre 3, il y a peu de classifieurs existants permettant de gérer à la fois les prédictions indéterminées et la sensibilité aux coûts. De plus, pour bien isoler les effets des facteurs que nous cherchons à évaluer (le coût affaibli généralisé, le cadre des probabilités utilisé, ...) et minimiser l'influence d'autres facteurs (différence de classifieur de base, différence de paramètres, ...), il est préférable d'utiliser des classifieurs similaires (par exemple, utilisant un même classifieur de base). Ceci limite encore nos choix de référentiels. Ainsi, pour mener correctement nos expériences, nous construisons nos propres référentiels.

Définition d'un classifieur indéterminé sensible aux coûts avec probabilités précises

Nous voulons définir un classifieur sensible aux coûts capable de produire des prédictions indéterminées à partir des probabilités précises. Nous avons vu dans la Section 3.1.2 que ceci peut être fait facilement (d'un point de vue théorique) en étendant l'espace de prédictions à 2^{Ω}. Il suffit pour cela de résoudre le problème de minimisation du coût espéré énoncé dans l'Équation (3.3), par une énumération exhaustive des prédictions indéterminées.

Cette approche est inefficace en pratique à cause du temps de calcul qui augmente exponentiellement en fonction du nombre de classes, mais comme nous avons seulement 9 classes au maximum pour les jeux de données choisis, une procédure de minimisation sur la totalité de l'espace de prédictions 2^{Ω} reste envisageable dans le cadre de nos expériences. Ainsi, nous l'utiliserons (avec le NBC comme classifieur de base) comme une méthode de référence.

L'avantage d'utiliser une telle approche est que, puisque del Coz et collab. (2009) et Ha (1997) ont obtenu des classifieurs avec un très bon pouvoir

prédictif en suivant une approche similaire (mais simplifiée, puisque l'espace des prédictions est réduit), il est alors raisonnable de considérer que cette approche (certes lourde en terme de temps de calcul) comme étant un "bon" classifieur du point de vue du pouvoir prédictif.

Définition d'un classifieur imprécis sensible aux coûts en se basant sur le NCC

Nous avons vu à la fin de la Section 3.3.3 que ce qui nous empêche d'utiliser le NCC avec des coûts génériques est que le problème de minimisation du coût espéré à l'Équation (3.23) a un temps de calcul prohibitif. Pour simplifier ce problème, nous proposons d'adopter une approximation : au lieu de minimiser directement le coût espéré, nous allons estimer d'abord les bornes des probabilités *a posteriori* $p(y|\mathbf{x})$ en résolvant les Équation (3.21) et (3.22) (*i.e.*, comme si c'est un problème avec coûts unitaires), et ensuite nous résolvons le problème de minimisation des coûts espérés suivant :

$$\underline{\mathbb{E}}[\mathcal{C}] = \min_{p(y|\mathbf{x})\in\{\underline{p}(y|x),\overline{p}(y|x)\}} \sum_{y\in\Omega} \mathcal{C}(y)p(y|\mathbf{x}).$$

En faisant ceci, on esquive le problème de calcul évoqué dans l'Équation (3.23). Une telle approche est standard dans la littérature pour approximer des ensembles crédaux en utilisant des intervalles de probabilités (Antonucci et Cuzzolin, 2010).

Concrètement, ceci revient à supposer que tous les ensembles crédaux sont d'une forme particulière (cf. Figure 6.1) et à remplacer l'ensemble crédal d'origine par un sur-ensemble de cette forme. La conséquence de cette approximation est que, comme nous considérons un sur-ensemble de l'ensemble crédal initial, les prédictions faites par ce modèle approché sont plus indéterminées que celles du modèle initial.

L'intérêt de faire cette approximation est d'obtenir un modèle simple basé sur le NCC, dont l'unique source influant le niveau d'indétermination des prédictions est le choix du paramètre s. Notre intention ici est d'obtenir un classifieur imprécis sensible aux coûts qui soit basique et qui pourra

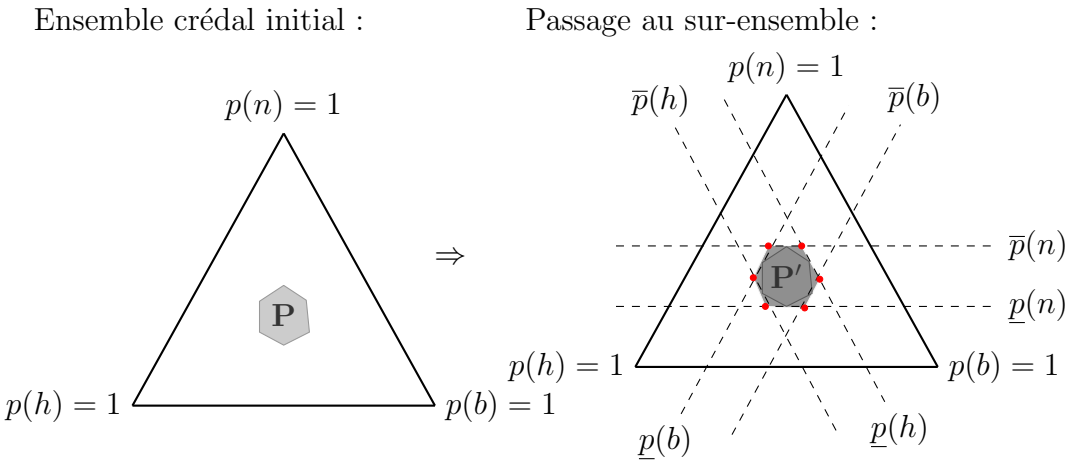


FIGURE 6.1: Passage d'un ensemble crédal initial à un sur-ensemble. La forme de l'ensemble crédal est toujours un hexagone, mais celle du sur-ensemble a tous ses côtés parallèles à un côté du triangle simplexe

servir de point de comparaison avec les autres approches. Par la suite, nous continuerons à désigner par "NCC", ce classifieur construit.

6.2 Expérience 1 : caractérisation de la formule du coût affaibli généralisé

Le coût affaibli généralisé permet de déterminer le coût des prédictions indéterminées lors de l'évaluation des classifieurs. L'objectif du Chapitre 4 était de proposer un cadre pour définir ces coûts de manière justifiée afin de permettre des comparaisons équitables entre les classifieurs (déterminés ou non) en fonction de l'aversion à la prudence. Ainsi, il est important d'assurer que la formule du coût affaibli généralisé proposée permet de calibrer et de refléter réellement cette aversion. Pour ceci, nous allons étudier la corrélation entre le paramètre r du coût affaibli généralisé et le paramètre s du NCC (construit tel que c'est décrit dans le paragraphe précédent). Comme nous savons que s permet de paramétrer le niveau d'imprécision du NCC, une corrélation de ces deux paramètres nous permet de déduire le comportement du paramètre r du coût affaibli généralisé.

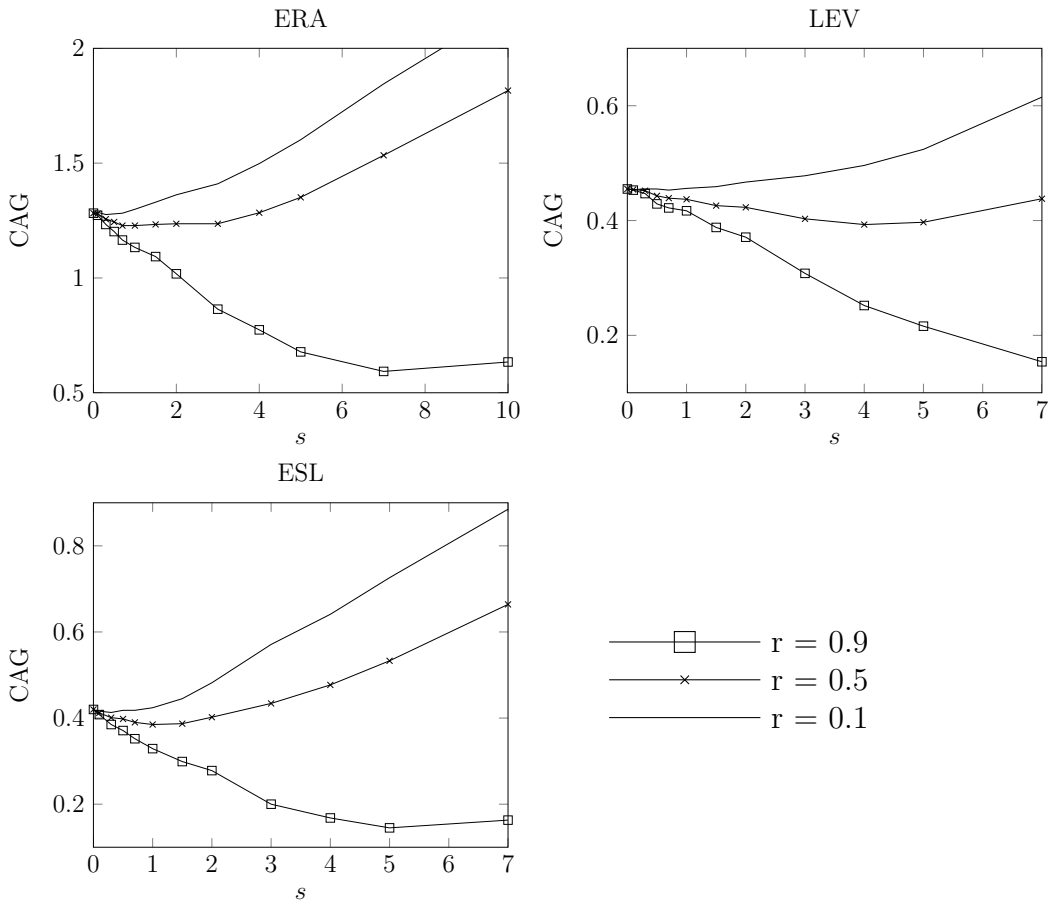


Figure 6.2: Le coût affaibli généralisé (CAG sur l'axe y) de ERA, LEV et ESL avec différentes combinaisons de r et s (axe x)

Sur la Figure 6.2, nous montrons pour $r \in \{0.1, 0.5, 0.9\}$, le coût affaibli généralisé défini selon la Définition 5 et obtenu par le classifieur NCC avec s allant de 0,001 à 10. Nous pouvons remarquer que l'évolution du coût affaibli généralisé suit le même comportement général. En effet, les courbes sont toutes approximativement convexes : d'abord le coût diminue quand on augmente s, car la prudence ajoutée est bénéfique et permet d'obtenir des prédictions plus justes pour les instances "difficile à classer", puis, après une valeur donnée de s, le coût commence à ré-augmenter, car trop d'indéterminations (inutiles) sont ajoutées.

Ceci est particulièrement visible pour les courbes avec $r = 0,5$ et $r =$

$0,9$, où nous pouvons voir clairement le point d'inflexion (sauf pour LEV avec $r = 0,9$ où ceci est au-delà de $s = 10$). Pour les autres cas, le point d'inflexion est toujours présent, mais moins visible dans les graphiques (trop proche de $s = 0$). En outre, plus r est proche de 0, plus le s optimal est proche de 0 et moins il y a de prédictions indéterminées. C'est le comportement initialement voulu, car nous voulons une aversion à la prudence faible quand r est faible. En revanche, lorsque r est proche de 1, le s optimal est atteint beaucoup plus tard, ce qui signifie que le niveau d'imprécision optimal est plus grand. Ceci suggère que r peut en effet définir et calibrer l'aversion à la prudence. Le même constat peut être fait si nous définissons le coût affaibli généralisé selon la Définition 6, la seule différence est que la valeur de s optimal sera atteinte plus tôt, comme il définit une attitude moins encline à la prudence.

Nous pouvons également noter que la plage de valeurs de r qui rendent les prédictions indéterminées intéressantes est $r \in]0; 1[$. Lorsque $r = 0$, nous n'accordons pas de récompense à la prudence, les courbes du coût affaibli généralisé seront strictement croissantes. De même pour $r = 1$, le coût est nul chaque fois que la vérité est incluse dans la prédiction, la courbe est alors strictement décroissante.

Cependant, pour un même niveau de r, le s optimal n'est pas toujours le même pour les différents jeux de données. Il est donc impossible de définir la relation entre les deux paramètres. Des techniques telles que la validation croisée pourraient être utilisées si nous voulons évaluer le niveau d'indétermination optimal d'un jeu de données selon r.

6.3 Expérience 2 : probabilités précises vs. imprécises

Ici, nous comparons les deux approches introduites dans le Chapitre 3 pour produire des prédictions indéterminées : soit avec le cadre des probabilités précises/standards (que nous appellerons également "PP") ou avec celui des probabilités imprécises (appelé aussi "IP" par la suite). Pour cette

comparaison, nous allons utiliser les deux classifieurs de référence que nous avons définis dans la Section 6.1.3. Les défauts liés à ces deux classifieurs (le temps de calcul du premier et le manque de garantie de performance prédictive du second) ne sont pas importants ici, car notre objectif n'est pas d'optimiser la performance, mais d'étudier en quoi ces deux approches diffèrent et comment nous pouvons les utiliser en pratique.

Nous tenons à souligner que, même si ces deux approches utilisent le même algorithme de base pour estimer les probabilités (le NCC est dérivé du NBC), ils diffèrent sur le processus de prise de décisions. L'approche PP résout un problème de minimisation du coût affaibli généralisé sur l'espace de toutes les prédictions possibles (*i.e.* 2^{Ω}) pour trouver la prédiction optimale. L'approche IP n'a pas besoin de résoudre un tel problème, au lieu de ceci, elle identifie les classes non-dominées sur l'espace des classes Ω. Si nous voulons aussi optimiser les prédictions de l'approche IP, il faudrait régler et trouver le paramètre s optimal comme indiqué dans la section précédente. Ceci est nécessaire si nous voulons comparer IP et PP équitablement. Mais ce problème ne peut être résolu de manière exacte comme le nombre de valeurs possibles pour s est infini. Pour remédier à ce problème, une approximation heuristique peut être obtenue en sélectionnant la valeur de s donnant le coût le plus faible sur une large plage de valeurs possibles avec une validation croisée sur les données d'apprentissage.

6.3.1 Comparaison du pouvoir prédictif

Nous évaluons ici brièvement le pouvoir prédictif des approches IP et PP avec le coût affaibli généralisé. Pour ceci, nous utilisons une validation croisée de deux manières : soit avec 4/5 de données pour l'apprentissage (le reste pour le test), soit avec seulement 1/5 pour l'apprentissage. Ces deux cadres nous permettent d'évaluer l'efficacité des deux classificateurs lorsque la quantité de données disponibles varie. Nous insistons ici sur le fait que l'objectif ici n'est pas de savoir quelle approche est "meilleure", mais d'assurer la comparabilité de ces deux approches et l'impact de l'approximation que nous avons fait pour rendre NCC sensible aux coûts.

Pour rendre ces deux approches comparables, nous optimisons le paramètre s de l'approche IP dans cette expérience. Ainsi, une validation croisée (à 10 blocs) sur les données d'apprentissage est utilisée pour déterminer la valeur optimale pour s sur une large plage de valeurs pré-sélectionnées variant de 0,001 à 10. Nous fixons aussi le paramètre r du coût affaibli généralisé à 0,5 pour toutes les expériences suivantes.

Données	IP		PP	
	coût	σ	coût	σ
ERA	1.26	0.08	1.23	0.06
LEV	0.41	0.04	0.39	0.04
ESL	0.39	0.04	0.35	0.04

TABLE 6.3: coût affaibli généralisé et l'écart-type (σ) quand 4/5 des données sont utilisées pour l'apprentissage

data sets	IP		PP	
	coût	σ	coût	σ
ERA	1.32	0.05	1.32	0.05
LEV	0.45	0.03	0.42	0.01
ESL	0.45	0.03	0.42	0.02

TABLE 6.4: coût affaibli généralisé et l'écart-type (σ) quand 1/5 des données sont utilisées pour l'apprentissage

Les Tableaux 6.3 et 6.4 nous montrent le coût affaibli généralisé obtenu avec l'écart-type pour les deux approches. Nous pouvons voir que les deux approches ont une performance similaire dans les deux cadres. Cependant, nous remarquons que l'approche PP semble être légèrement meilleure quand il y a moins de données (dans le cadre où 1/5 des données sont utilisées pour l'apprentissage). Ceci peut être expliqué par le fait que les prédictions faites par cette approche sont des solutions exactes du problème de minimisation des coûts, tandis que celles données par l'approche IP sont des approximations dépendant des valeurs de s choisies. En effet, le classifieur dérivé du NCC que nous avons construit pour l'approche IP était loin d'être optimal de par sa construction et son pouvoir prédictif n'est pas du tout garanti. Ainsi, cette comparaison nous permet de savoir que, malgré les approximations faites pour rendre le NCC sensible aux coûts, son pouvoir prédictif

reste acceptable pour qu'on puisse le considérer comme un point de comparaison ultérieurement.

Toutefois, pour valider si les différences sont statistiquement significatives, nous aurions besoin de tester sur plus de jeux de données et sur un échantillon plus condensé pour les valeurs de s. Ce n'est pas notre objectif, mais allons donner quelques arguments dans les paragraphes suivants qui nous ont conduits à penser qu'une telle différence statistiquement significative est peu probable.

6.3.2 Comparaison du niveau d'indétermination

Dans le paragraphe précédent, nous avons vu que les deux approches ont des performances comparables. Ici, nous nous concentrons sur ce qui les différencie. Tout d'abord, nous évaluons l'indétermination moyenne (taille moyenne des prédictions) des deux approches en faisant varier la proportion des données utilisés pour l'apprentissage. Pour cette expérience, notre intention est de décrire le comportement en fonction du niveau d'indétermination pour chaque approche séparément, il n'y a plus de raison d'optimiser le paramètre s. Nous gardons donc $s = 2$ quel que soit le nombre d'instances des données d'apprentissage, de sorte qu'il reste logique de comparer les niveaux d'indétermination lorsque la quantité de données change.

La Figure 6.3 montre que l'approche PP a une indétermination moyenne qui est beaucoup moins sensible à la variation de la quantité des données d'apprentissage : l'indétermination moyenne est presque constante quelle que soit la proportion des données utilisées pour l'apprentissage. En revanche, l'approche IP est en mesure de régler le niveau d'indétermination en fonction de la quantité d'informations disponibles : pour une valeur donnée de s, plus il y a de données, plus l'indétermination est faible.

Cela révèle la différence fondamentale entre les approches PP et IP. L'approche PP est seulement sensible à l'indétermination due à l'ambiguïté. Cela signifie que l'indétermination n'est basée que sur les valeurs des probabilités (ou le coût affaibli généralisé dans notre cas), *i.e.* une prédiction indéterminée n'est faite que lorsque les probabilités (ou le coût affaibli généralisé)

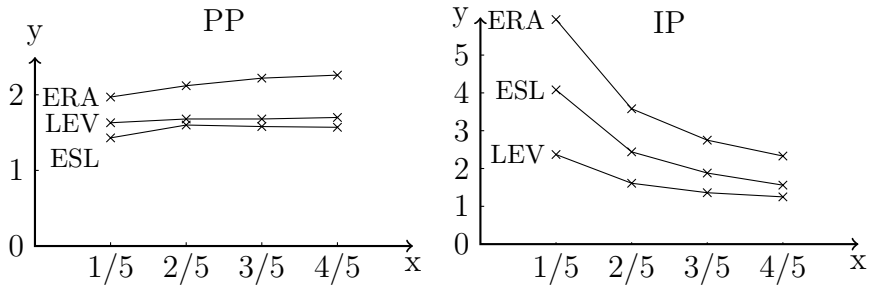


FIGURE 6.3: Taille des prédictions (axe y) selon la proportion des données utilisées pour l'apprentissage

sont trop similaires pour qu'une seule classe d'être prédite. L'approche IP considère un autre facteur : l'indétermination due au manque d'information qui est modélisé en estimant la confiance de chaque estimation de probabilités sous forme d'intervalles. Par exemple, si la probabilité d'une classe est estimée à $[0.1; 0.9]$ en raison de l'absence de données, alors il est fort possible qu'elle sera toujours prédite par l'approche IP car cette possibilité ne peut pas être exclue en toute confiance.

Cette considération est absente lorsque nous utilisons l'approche PP, de sorte que tant que les proportions des classes restent similaires, la quantité de données d'apprentissage a peu d'impact sur les prédictions. Cela peut aussi expliquer pourquoi l'approche PP peut maintenir un bon pouvoir prédictif quand il y a moins de données, même si elle est censée donner des résultats moins fiables. Ainsi, l'aversion à la prudence prend un sens précis ici, puisqu'elle est étroitement liée à la taille moyenne des prédictions. Cette approche est donc plus intéressante quand nous voulons que la taille des prédictions soit stables, *i.e.*, indépendante de la quantité des données.

Inversement, l'aversion à la prudence prend un sens plus large avec l'approche par probabilités imprécises, puisqu'elle est liée à la fois à l'ambiguïté des estimations et au manque d'information. Cependant, à moins de fixer s à un niveau très élevé, plus il y a de données, plus les bornes inférieure et supérieure des intervalles de probabilités vont se rapprocher, et plus les résultats de l'approche IP seront similaires à ceux d'une approche déterminée. Donc, nous pensons que l'approche IP serait moins sensible à l'ambiguïté

que PP dans le cas où il y a beaucoup de données. En effet, avec une infinité de données, IP donnera toujours des prédictions déterminées, alors que PP pourra encore donner des prédictions indéterminées. Ce sera vérifiée dans le prochain paragraphe.

6.3.3 Utiliser l'indétermination pour déduire les spécificités des approches et des données

Nous avons vu dans le paragraphe précédent que les comportements des deux approches indéterminées peuvent être très différents. Ici, nous précisons leurs différences en examinant plus en détail comment et quand les prédictions indéterminées sont faites. Pour cela, nous calculons les pourcentages d'instances de données où, soit les deux approches donnent des prédictions indéterminées (ou déterminées), soit seulement l'une des deux donne des prédictions indéterminées.

Cette connaissance nous permet alors de déduire l'origine de l'indétermination (*i.e.*, soit l'ambiguïté soit le manque d'information) d'un jeu de données, puisque si les deux approches choisissent d'être indéterminées principalement sur les mêmes instances de données, cela signifie qu'une grande partie d'instances "difficile à classer" sont les mêmes pour les deux approches. La Figure 6.4 montre que c'est bien le cas pour le jeu de données ERA, où 80% des prédictions indéterminées se font sur les mêmes instances de données.

Cependant, sur la même figure, le même constat ne peut être fait pour LEV et ESL : le niveau d'indétermination de l'approche PP est plus élevé que celui de IP, puisque le cas où seulement PP est indéterminée a un pourcentage nettement plus élevé que celui où seulement IP est indéterminée. Comme nous avons vu que l'approche PP ne peut tenir compte que de l'indétermination due à l'ambiguïté, cela signifie que les instances de données qui sont considérées comme difficiles à classer ne sont pas toujours les mêmes pour les deux approches. Il y a beaucoup plus d'instances de données (en particulier pour LEV et ESL) qui sont considérées comme difficiles à classer par l'approche PP mais pas par IP. Cependant, la comparaison

du pouvoir prédictif dans la Section 6.3.1 a montré que les deux approches ont des performances comparables. Cela nous amène à penser que, même si l'approche IP est moins sensible à l'ambiguïté, néanmoins elle peut (mieux) réussir à identifier les instances de données où l'ambiguïté est critique.

En outre, l'utilisation conjointe de ces deux approches permet de révéler des informations sur les jeux de données eux-mêmes. Sachant que l'approche PP est seulement sensible à l'indétermination due à l'ambiguïté, cette spécificité peut être exploitée par l'expert et le décideur pour comprendre et améliorer les jeux de données. Par exemple, nous avons une grande partie des instances de données qui ne sont indéterminées que pour l'approche PP avec les jeux de données LEV et ESL. Par conséquent, si l'expert ou le décideur veulent améliorer la justesse des prédictions, le fait d'essayer d'autres classifieurs de base et/ou l'ajout de variables d'entrée plus discriminantes peut s'avérer utile (afin de lever l'ambiguïté des classes).

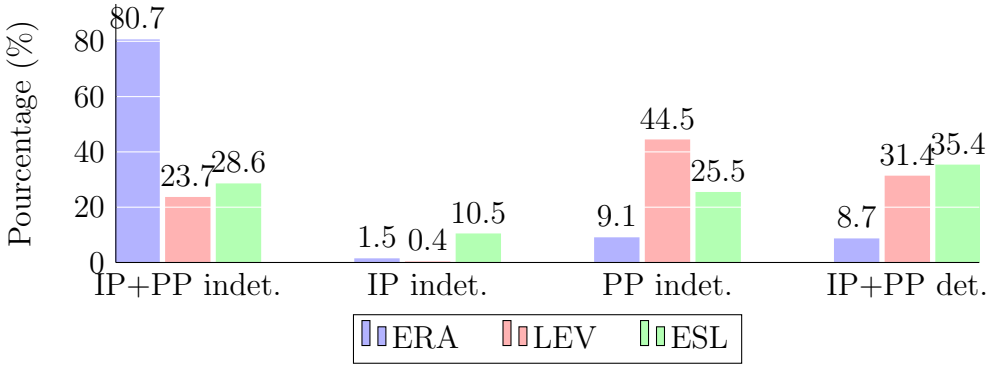


FIGURE 6.4: Pourcentage et répartition des prédictions indéterminées (et déterminées) selon les jeux de données, quand 4/5 des données sont utilisées pour l'apprentissage.

De même, nous savons que l'approche IP prend le manque d'information en considération, de sorte que si nous avons des instances de données où seulement IP est indéterminée (par exemple avec ESL), il peut être utile de recueillir plus de données (en particulier celles similaires aux instances de données où seulement IP est indéterminée) afin d'améliorer le pouvoir prédictif. Cette situation est visible sur la Figure 6.5, où nous montrons ce qui se passe si nous faisons la même expérience, mais avec seulement 1/5 des

données utilisées pour l'apprentissage. En fait, ici, le niveau d'indétermination de l'approche IP est significativement plus élevé que celui de PP, et une partie importante des instances de données ne sont indéterminées que pour IP. Par conséquent, il est possible de déduire que l'indétermination est due au manque d'information, et il serait utile de recueillir plus de données.

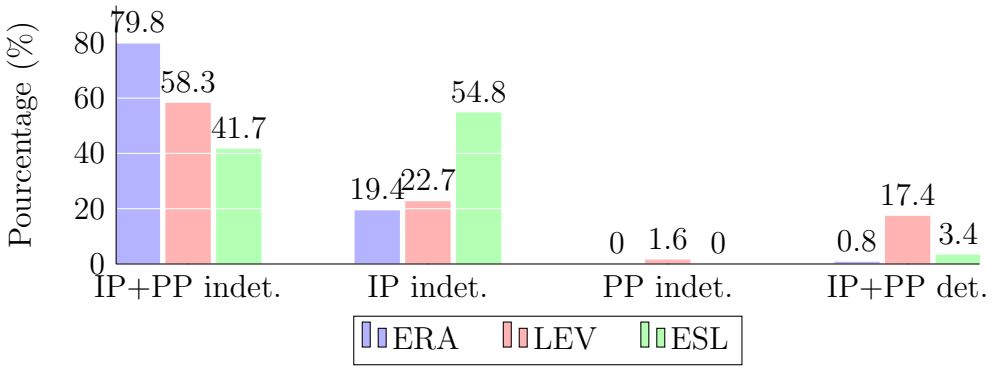


Figure 6.5: Pourcentage et répartition des prédictions indéterminées (et déterminées) selon les jeux de données, quand 1/5 des données sont utilisées pour l'apprentissage.

6.3.4 Conclusion sur l'expérience

Avec cette expérience, nous avons pu comparer et révéler des différences de comportement sur les classifieurs indéterminés construits soit dans le cadre des probabilités précises soit imprécises. Les résultats ont montré l'intérêt du modèle probabiliste imprécis pour produire des prédictions indéterminées sensibles aux coûts pour plusieurs raisons :

- La performance d'un modèle imprécis (même s'il agit d'un modèle assez approximatif) peut atteindre un niveau comparable à celle d'un modèle glouton (et exhaustif) utilisant les probabilités standards.
- La sensibilité à la quantité de données disponibles. Il peut refléter le manque d'information (à travers le niveau d'indétermination) quand il n'y a pas suffisamment de données pour prédire de manière fiable.

Le pré-requis à ces avantages est que, si un modèle paramétrique est utilisé, le paramètre du classifieur a besoin d'être optimisé en fonction de

l'aversion au risque. En effet, un point essentiel que nous avons remarqué est que, lorsque le coût affaibli généralisé est utilisé pour la comparaison de classifieurs variés, il est crucial d'assurer que les modèles paramétriques soient appliqués avec des paramètres adéquats en accord avec l'aversion à la prudence fixée par le coût affaibli généralisé, sinon la comparaison est biaisée dès le départ.

Un autre résultat intéressant est que, nous pouvons combiner l'utilisation du cadre de probabilités imprécises et précises pour expliciter des informations sur les données elles-mêmes et déduire des informations utiles (ambiguïté ou manque d'information sur les données existantes) aux experts et aux décideurs.

6.4 Expérience 3 : pouvoir prédictif des dichotomies emboîtées imprécises

Dans le chapitre précédent, nous avons pu tester l'intérêt d'utiliser les dichotomies emboîtées avec les probabilités imprécises, cependant, dû au manque de référentiel adapté et de critère d'évaluation, nous n'avons pas pu évaluer son pouvoir prédictif en présence de coûts génériques. Comme nous avons validé notre formule d'évaluation avec l'Expérience 1 et défini une version sensible aux coûts du NCC avec l'Expérience 2, nous pouvons maintenant évaluer le pouvoir prédictif des dichotomies imprécises avec une structure de coûts.

6.4.1 Cadre expérimental

Pour cette expérience, nous allons comparer nos modèles se basant sur les dichotomies emboîtées avec la version sensible aux coûts du NCC, qui est utilisée comme méthode de référence. Tous les modèles seront appliqués avec le paramètre s fixé à 2. C'est d'ailleurs la raison pour laquelle nous n'allons pas comparer nos modèles avec l'approche gloutonne utilisant les probabilités précises, car ceci nous obligerait à optimiser s pour avoir des comparaisons non-biaisées. Comme nous voulons avant tout juger

Nom	Nb. instances	Nb. attributs	Nb. classes
autoPrice	159	16	7
boston housing	506	14	7
california housing	20640	9	7
delta ailerons	7129	6	7
delta elevators	9517	7	7
kinematics	8192	9	7
ERA	1000	5	9
ESL	488	5	9
LEV	1000	5	5

Table 6.5: Jeux de données ordinales utilisés pour l'Expérience 3

le pouvoir prédictif des dichotomies emboîtées, l'utilisation d'une procédure d'optimisation de s rajouterait une source d'influence parasitaire pour notre objectif.

Cependant, comme nous comparons exclusivement des modèles imprécis ici (ils possèdent tous les avantages du modèle imprécis décrits précédemment), il faut que nos modèles aient de meilleurs résultats et ces derniers doivent être statistiquement significatifs pour confirmer l'intérêt des dichotomies emboîtées. Ainsi, il nous faut plus de jeux de données. Comme les données naturellement ordinales sont difficiles à trouver, nous allons utiliser ici des données initialement utilisées pour des problèmes de régression. La variable de sortie de ces données est définie sur un ensemble continu (*e.g.*, $\mathbb{R}$), infini (*e.g.*, $\mathbb{Z}$) ou trop grand (*e.g.*, $[0; 100]$) pour qu'on puisse considérer chaque élément comme une classe. Malgré ceci, il existe un ordre naturel sur cet ensemble, ce qui nous permet de les transformer en données ordinales avec une discrétisation. Les jeux de données utilisées ici sont détaillés dans le Tableau 6.5. La variable de sortie est discrétisées en 7 classes de fréquence égale. S'il y a des attributs continus, ils sont discrétisés en 5 valeurs de fréquence égale.

	coût affaibli généralisé (rang)			
Données	NCC	F_m	F_v	ND+NCC
autoPrice	2.81 (4)	1.26 (2)	1.24 (1)	1.76 (3)
1 boston	3.3 (4)	1.27 (2)	1.26 (1)	2.4 (3)
california	3.62 (4)	1.47 (1.5)	1.47 (1.5)	1.56 (3)
delta ailerons	3.19 (4)	1.12 (2)	1.1 (1)	1.26 (3)
delta elevators	2.78 (4)	0.86 (1.5)	0.86 (1.5)	0.93 (3)
kinematics	3.45 (4)	1.52 (1.5)	1.52 (1.5)	1.59 (3)
ERA	3.79 (4)	1.93 (1.5)	1.93 (1.5)	2.0 (3)
LEV	2.2 (4)	0.81 (2)	0.8 (1)	0.87 (3)
ESL	3.43 (4)	0.88 (1.5)	0.88 (1.5)	1.18 (3)
rang moyen	4	1.72	1.28	3

TABLE 6.6: Coûts affaiblis généralisés des différents jeux de données ordinales

6.4.2 Comparaison de performance

Le Tableau 6.6 montre les résultats obtenus en termes du coût affaibli généralisé et du rang de chaque classifieur. Comme les méthodes comparées sont similaires à celles utilisées dans le cas multiclasse, nous n'allons donc pas les présenter à nouveau (*cf.* Section 5.3.2). Les résultats des Section 5.3.3 et 5.3.4 sur les spécificités de l'approche imprécise et des dichotomies emboîtées restent valides. Par conséquent, nous allons nous concentrer sur ce qui diffère du cas avec coûts unitaires.

Nous pouvons constater que le NCC, le ND+NCC et les approches par forêts ont des performances très distinctes, contrairement au cas multiclasse avec des coûts unitaires. Les approches par dichotomies emboîtées réalisent des performances systématiquement supérieures au NCC ("approché") sur tous les jeux de données. De plus, le coût affaibli généralisé obtenu pour le NCC est très souvent deux fois plus grand (voire plus) que celui obtenu avec les approches par forêts de dichotomies. Pour assurer que les différences soient statistiquement significatives, nous suivons l'approche recommandée par Demšar (2006) et nous appliquons le test de de Friedman Friedman (1937) sur les rangs obtenus par les classifieurs sur chaque jeu de données. Nous trouvons une p-valeur inférieure à 10^{-4}, l'hypothèse nulle peut être

être rejetée avec confiance. Ceci signifie que les performances sont significativement différentes pour ces classifieurs.

Comme l'hypothèse nulle est rejetée, nous pouvons utiliser un test de Nemenyi (Nemenyi, 1962) comme test *post-hoc*. Nous obtenons que, pour que deux classifieurs soient significativement différents (avec un niveau de confiance de 95%), il faut avoir une différence de rang moyen de plus de 1.56. Ainsi, sur la Figure 6.6, nous constatons que les approches ensemblistes avec dichotomies (F_m, F_v) surpassent significativement le NCC. Les deux approches ensemblistes ont une performance très comparables entre elles. En revanche, l'approche avec une structure unique est nettement moins performant par rapport aux approches ensemblistes en terme de rang, même si la différence en terme de coût affaibli généralisé est plutôt faible. Cette différence est intéressante notamment parce que nous avons vu qu'elles ont des performances similaires dans le cas des coûts unitaires. Ceci peut notamment être expliqué par le choix probablement non optimal de la structure de dichotomies.

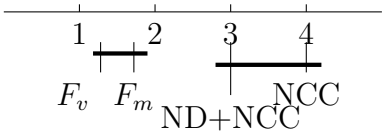


Figure 6.6: Test post hoc de Nemenyi sur les rangs des performances des classifieurs. Les groupes de classifieurs non significativement différents (avec une chance de plus de 5%) sont reliés avec une ligne en gras

Par ailleurs, en mettant ces résultats en perspective avec les pourcentages de prédictions indéterminées produites par ces approches, illustrés dans le Tableau 6.7, nous pouvons voir que l'approche avec une structure unique de dichotomies a un niveau d'indétermination presque toujours inférieur (sauf pour *delta elevators*) aux approches ensemblistes. Ceci peut également expliquer la différence de performances. ND+NCC a choisi d'être moins prudent sur certaines instances de données "difficiles à classer". Cependant, un niveau d'indétermination élevé n'est pas non plus synonyme de bonne performance (par exemple ND+NCC est plus performant que NCC

	Pourcentages de prédictions indéterminées (rang)			
Données	NCC	F_m	F_v	ND+NCC
autoPrice	74 (2)	90 (4)	92 (3)	58 (1)
boston	47 (2)	66 (4)	63 (3)	39 (1)
california	4 (1)	8 (3.5)	8 (3.5)	7 (2)
delta ailerons	4 (1)	7 (2.5)	12 (4)	7 (2.5)
delta elevators	1 (1)	4 (2)	9 (4)	5 (3)
kinematics	16 (1)	24 (3)	27 (4)	21 (2)
ERA	60 (1)	80 (3)	82 (4)	69 (2)
LEV	39 (4)	30 (3)	29 (2)	23 (1)
ESL	41 (3)	38 (2)	43 (4)	29 (1)
rang moyen	1.8	3	3.5	1.8

TABLE 6.7: Pourcentages d'instances de données où des prédictions indéterminées sont faites par chaque classifieur

tout en gardant le même niveau d'indétermination), ceci montre que le coût affaibli généralisé reste équitable et ne favorise pas l'indétermination. Nous notons également que l'approximation faite sur le NCC ne fait pas de lui le plus indéterminé, ainsi les "mauvaises" performances ne sont pas dues à cette approximation.

Chapitre 7

Conclusion et perspectives

Dans ce travail de thèse, nous avons examiné deux approches existantes de la littérature de l'apprentissage statistique pour rendre les modèles et les prédictions plus prudents et plus fiables : le cadre des probabilités imprécises et l'apprentissage sensible aux coûts. Nous avons étudié leurs spécificités à la fois sur le plan théorique et sur le plan pratique. Cette étude détaillée nous a notamment mené à penser qu'il serait intéressant de combiner les avantages de deux approches afin de construire un modèle prédictif plus compétent pour représenter, traiter et inférer les connaissances d'experts et les informations *a priori*. Au cours de notre travail, nous avons pu remarquer que la combinaison de ces deux cadres représente un défi à la fois théorique et pratique.

Sur le plan théorique, peu de travaux existants ont abordé la manière de quantifier les différentes erreurs de classification quand des prédictions sous forme d'ensembles sont produites et quand ces erreurs ne se valent pas (en terme de conséquences). Notre première contribution, dans le Chapitre 4, a donc été d'établir des propriétés générales et des lignes directrices permettant la quantification des coûts d'erreurs de classification pour les prédictions sous forme d'ensembles. Ces propriétés nous ont permis de dériver une formule générale, le coût affaibli généralisé, qui rend possible la comparaison des classifieurs quelle que soit la forme de leur prédictions (singleton ou ensemble) en tenant compte d'un paramètre d'aversion à la

prudence.

Sur le plan pratique, la plupart des classifieurs utilisant les probabilités imprécises ne permettent pas d'intégrer des coûts d'erreurs de classification génériques de manière simple. Nous avons notamment examiné le cas du Classifieur Crédal Naïf, où la simplicité originelle du classifieur est perdue lorsque nous faisons intervenir des coûts non unitaires. Ce problème a mené à notre deuxième contribution, la mise en place d'un classifieur qui permet de gérer les intervalles de probabilités produits par les probabilités imprécises et les coûts d'erreurs génériques avec le même ordre de complexité que dans le cas où les probabilités standards et les coûts unitaires sont utilisés. Il s'agit d'utiliser une technique de décomposition binaire, les dichotomies emboîtées. Les propriétés et les pré-requis de ce classifieur ont été étudiés en détail dans le Chapitre 5. Nous avons notamment pu voir que les dichotomies emboîtées permettent de réduire le niveau d'indétermination du modèle imprécis de base (NCC) sans perte de pouvoir prédictif.

Au final, des expériences variées ont été menées sur des jeux de données ordinales pour appuyer nos contributions. Dans le Chapitre 6, nous avons d'abord caractérisé le comportement du coût affaibli généralisé et ensuite donné des indications sur son utilisation en pratique. Nous avons notamment vu que son application pour la comparaison des classifieurs requiert une attention particulière quand ces derniers ont également des paramètres calibrant l'aversion à l'indétermination. De plus, nous avons mis en évidence les différences entre un modèle basé sur les probabilités standards pour produire des prédictions indéterminées et un modèle utilisant les probabilités imprécises. Ce dernier est en général plus compétent car il permet de distinguer deux sources d'indétermination (l'ambiguïté et le manque d'informations), même si l'utilisation conjointe de ces deux types de modèles présente également un intérêt particulier dans l'optique d'assister le décideur à améliorer les données ou les classifieurs. Enfin, nous avons vu que l'utilisation des dichotomies emboîtées permet d'améliorer significativement le pouvoir prédictif d'un modèle imprécis avec des coûts génériques.

Cependant, malgré les avantages des dichotomies emboîtées imprécises, plusieurs problèmes restent à résoudre. D'abord, le choix de la structure

de dichotomies est un problème important, nous avons vu qu'il peut grandement influencer la qualité des prédictions. Les approches ensemblistes et heuristiques que nous avons développées sont certes efficaces, mais le temps de calcul et le pouvoir prédictif peuvent pâtir d'un nombre de classes trop élevé. Trouver une approche efficace pour déterminer une structure unique d'arbre de dichotomies adéquate apportera une amélioration significative à notre approche.

Nous nous sommes limités dans ce travail à la classification multiclasse, cependant, le cadre des dichotomies imprécises que nous avons posé pourra également être efficace dans d'autres contextes. Par exemple, il pourrait s'avérer utile pour d'autres données structurées comme les données multilabel ou l'apprentissage des préférences. Ce point ouvre la voie à de nombreux travaux. De plus, les propriétés que nous avons proposées dans le Chapitre 4, quoique intuitives, n'étaient pas justifiées à l'aide d'un cadre théorique. Il serait donc intéressant de justifier ces propriétés par exemple dans le cadre de "paris" comme dans les travaux de Zaffalon et collab. (2012). De même, la formule du coût affaibli généralisé a été étendue à partir de la notion de justesse affaiblie selon l'utilité, mais la mesure F_β était également une autre proposition théoriquement solide sur laquelle nous aurions pu nous baser pour étendre au cadre des coûts génériques. Il serait également intéressant d'investiguer ces points dans un travail futur.

Bibliographie

Abellán, J. et A. R. Masegosa. 2012, «Imprecise classification with credal decision trees», *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 20, n° 05, p. 763–787.

Abellán, J. et S. Moral. 2005, «Upper entropy of credal sets. applications to credal classification», *International Journal of Approximate Reasoning*, vol. 39, n° 2, p. 235–255.

Allwein, E. L., R. E. Schapire et Y. Singer. 2001, «Reducing multiclass to binary : A unifying approach for margin classifiers», *Journal of Machine Learning Research*, vol. 1, p. 113–141.

Aly, M. 2005, «Survey on multiclass classification methods», cahier de recherche, California Institute of Technology.

Antonucci, A. et F. Cuzzolin. 2010, «Credal sets approximation by lower probabilities : application to credal networks», dans *Computational Intelligence for Knowledge-based Systems Design, 13th International Conference on Information Processing and Management of Uncertainty (IPMU)*, *Lecture Notes in Computer Science*, vol. 6178, édité par E. Hüllermeier, R. Kruse et F. Hoffmann, Springer, p. 716–725.

Bache, K. et M. Lichman. 2013, «UCI machine learning repository», URL `http://archive.ics.uci.edu/ml`.

Bailey, T. et C. Elkan. 1994, «Fitting a mixture model by expectation maximization to discover motifs in biopolymers.», dans *Proceedings of the International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 2, p. 28.

Bartlett, P. L. et M. H. Wegkamp. 2008, «Classification with a reject option using a hinge loss», *The Journal of Machine Learning Research*, vol. 9, p. 1823–1840.

Baum, L. E. et T. Petrie. 1966, «Statistical inference for probabilistic functions of finite state markov chains», *The annals of mathematical statistics*, p. 1554–1563.

Bernard, J.-M. 2005, «An introduction to the imprecise dirichlet model for multinomial data», *International Journal of Approximate Reasoning*, vol. 39, n° 2, p. 123–150.

Billingsley, P. 1979, *Probability and measure*, John Wiley & Sons.

Breiman, L., J. Friedman, C. J. Stone et R. A. Olshen. 1984, *Classification and regression trees*, CRC press.

Cesa-Bianchi, N., Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire et M. K. Warmuth. 1997, «How to use expert advice», *Journal of the ACM*, vol. 44, n° 3, p. 427–485.

Chai, X., L. Deng, Q. Yang et C. X. Ling. 2004, «Test-cost sensitive naive bayes classification», dans *Proceedings of the Fourth IEEE International Conference on Data Mining*, IEEE Computer Society, p. 51–58.

Chen, Y., M. M. Crawford et J. Ghosh. 2004, «Integrating support vector machines in a hierarchical output space decomposition framework», dans *Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. Proceedings. 2004 IEEE International*, vol. 2, IEEE, p. 949–952.

Chow, C. K. 1970, «On optimum recognition error and reject tradeoff», *Information Theory, IEEE Transactions on*, vol. 16, n° 1, p. 41–46.

Corani, G. et M. Zaffalon. 2008, «Credal model averaging : an extension of bayesian model averaging to imprecise probabilities», dans *Machine Learning and Knowledge Discovery in Databases*, Springer, p. 257–271.

Cortes, C. et V. Vapnik. 1995, «Support-vector networks», *Machine learning*, vol. 20, n° 3, p. 273–297.

Cox, D. R. 1958, «The regression analysis of binary sequences», *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 215–242.

del Coz, J. J., J. Díez et A. Bahamonde. 2009, «Learning nondeterministic classifiers», *Journal of Machine Learning Researchs*, vol. 10, p. 2273–2293.

Cozman, F. G. 2000, «Credal networks», *Artificial intelligence*, vol. 120, n° 2, p. 199–233.

De Cooman, G. et F. Hermans. 2008, «Imprecise probability trees : Bridging two theories of imprecise probability», *Artificial Intelligence*, vol. 172, n° 11, p. 1400–1427.

Deirdre, O. B., G. R. Maya et G. M. Robert. 2008, «Cost-sensitive multiclass classification from probability estimates», dans *Proceedings of the twenty-fifth International Conference on Machine learning*, p. 712–719.

Demšar, J. 2006, «Statistical comparisons of classifiers over multiple data sets», *The Journal of Machine Learning Research*, vol. 7, p. 1–30.

Destercke, S. et B. Quost. 2011, «Combining binary classifiers with imprecise probabilities», dans *Proceedings of the 2011 International Conference on Integrated Uncertainty in Knowledge Modelling and Decision Making*, Springer-Verlag, p. 219–230.

Dietterich, T. G. et G. Bakiri. 1995, «Solving multiclass learning problems via error-correcting output codes», *Journal of Artificial Intelligence Research*, vol. 2, p. 263–286.

Domingos, P. 1999, «Metacost : A general method for making classifiers cost-sensitive», dans *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, p. 155–164.

Domingos, P. et M. Pazzani. 1997, «On the optimality of the simple bayesian classifier under zero-one loss», *Machine learning*, vol. 29, n° 2-3, p. 103–130.

Drummond, C. et R. C. Holte. 2000, «Exploiting the cost (in)sensitivity of decision tree splitting criteria», dans *In Proceedings of the Seventeenth International Conference on Machine Learning*.

Dubuisson, B. et M.-H. Masson. 1993, «A statistical decision rule with incomplete knowledge about classes», *Pattern recognition*, vol. 26, n° 1, p. 155–165.

Eibe, F. et K. Stefan. 2004, «Ensembles of nested dichotomies for multi-class problems», dans *Proceedings of the 21st International Conference on Machine Learning*, ACM.

Elkan, C. 2001, «The foundations of cost-sensitive learning», *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*.

Fayyad, U. M. et K. B. Irani. 1993, «Multi-interval discretization of continuous-valued attributes for classification learning», dans *Proceedings of the International Joint Conference on Uncertainty in Artificial Intelligence*, p. 1022–1029.

Fox, J. 1997, *Applied regression analysis, linear models, and related methods.*, Sage Publications, Inc, 472-474 p..

Frélicot, C., M.-H. Masson et B. Dubuisson. 1995, «Reject options in fuzzy pattern classification rules», dans *Proceedings of 3rd European Congress on Intelligent Techniques and Soft Computing*, vol. 75.

Friedman, M. 1937, «The use of ranks to avoid the assumption of normality implicit in the analysis of variance», *Journal of the American Statistical Association*, vol. 32, n° 200, p. 675–701.

Friedman, M. 1940, «A comparison of alternative tests of significance for the problem of m rankings», *The Annals of Mathematical Statistics*, vol. 11, n° 1, p. 86–92.

Friedman, N., D. Geiger et M. Goldszmidt. 1997, «Bayesian network classifiers», *Machine learning*, vol. 29, n° 2-3, p. 131–163.

Fumera, G., F. Roli et G. Giacinto. 2000, «Reject option with multiple thresholds», *Pattern Recognition*, vol. 33, p. 2099–2101.

Gupta, S. S. 1965, «On some multiple decision (selection and ranking) rules», *Technometrics*, vol. 7, n° 2, p. 225–245.

Ha, T. M. 1997, «The optimum class-selective rejection rule», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, n° 6, doi :10.1109/34.601248, p. 608–615.

Hastie, T. et R. Tibshirani. 1998, «Classification by pairwise coupling», *The Annals of Statistics*, vol. 26, p. 451–471.

Hastie, T., R. Tibshirani et J. Friedman. 2001, *The elements of statistical learning : data mining, inference and prediction*, Springer-Verlag.

Herbei, R. et M. H. Wegkamp. 2006, «Classification with reject option», *Canadian Journal of Statistics*, vol. 34, n° 4, p. 709–721.

Hsu, C.-W. et C.-J. Lin. 2002, «A comparison of methods for multiclass support vector machines», *Neural Networks, IEEE Transactions on*, vol. 13, n° 2, p. 415–425.

Huhn, J. C. et E. Hullermeier. 2008, «Is an ordinal class structure useful in classifier learning ?», *International Journal of Data Mining, Modelling and Management*, vol. 1, n° 1, p. 45–67.

Japkowicz, N. et S. Stephen. 2002, «The class imbalance problem : a systematic study», *Intelligent data analysis*, vol. 6, n° 5, p. 429–449.

Kamp, H. 1981, «A theory of truth and semantic representation», *Formal Methods in the Study of Language*.

Kohavi, R. 1996, «Scaling up the accuracy of naive-bayes classifiers : a decision-tree hybrid», dans *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, p. 202–207.

Kumar, S., J. Ghosh et M. M. Crawford. 2002, «Hierarchical fusion of multiple classifiers for hyperspectral data analysis», *Pattern Analysis & Applications*, vol. 5, n° 2, p. 210–220.

Levi, I. 1983, *The enterprise of knowledge : An essay on knowledge, credal probability, and chance*, MIT press.

Ling, C. X., Q. Yang, J. Wang et S. Zhang. 2004, «Decision trees with minimal costs», dans *Proceedings of the twenty-first International Conference on Machine learning*, p. 69.

Lorena, A. C. et A. C. de Carvalho. 2010, «Building binary-tree-based multiclass classifiers using separability measures», *Neurocomputing*, vol. 73, n° 16, p. 2837–2845.

Maloof, M. A. 2003, «Learning when data sets are imbalanced and when costs are unequal and unknown», dans *Proceedings of the Workshop on Learning from Imbalanced Data sets at the International Conference on Machine*, vol. 2, p. 2–1.

Margineantu, D. D. 2002, «Class probability estimation and cost-sensitive classification decisions», dans *Proceedings of the 13th European Conference on Machine Learning*, Springer-Verlag, p. 270–281.

Mauá, D. D., C. P. de Campos, A. Benavoli et A. Antonucci. 2014, «Probabilistic inference in credal networks : new complexity results», *Journal of Artificial Intelligence Research*, vol. 50, n° 1, p. 603–637.

McCulloch, W. S. et W. Pitts. 1943, «A logical calculus of the ideas immanent in nervous activity», *The bulletin of mathematical biophysics*, vol. 5, n° 4, p. 115–133.

Mercer, J. 1909, «Functions of positive and negative type, and their connection with the theory of integral equations», *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, p. 415–446.

Murphy, K. P. 2002, *Dynamic Bayesian networks : representation, inference and learning*, thèse de doctorat, University of California, Berkeley.

Nemenyi, P. 1962, «Distribution-free multiple comparisons», dans *Biometrics*, vol. 18, INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210, p. 263.

Quinlan, J. R. 1993, «C4.5 : programs for machine learning», .

Rifkin, R. et A. Klautau. 2004, «Parallel networks that learn to pronounce english text», *Journal of Machine Learning Research*, p. 101–141.

Rokach, L. 2006, «Decomposition methodology for classification tasks : a meta decomposer framework», *Pattern Analysis and Applications*, vol. 9, n° 2-3, p. 257–271.

Rokach, L. 2010, «Ensemble-based classifiers», *Artificial Intelligence Review*, vol. 33, n° 1-2, p. 1–39.

Sheng, V. S. et C. X. Ling. 2006, «Thresholding for making classifiers cost-sensitive», dans *Proceedings of the Bational Conference on Artificial Intelligence*, vol. 21-1, AAAI Press, p. 476–481.

Ting, K. 1998, «Inducing cost-sensitive trees via instance weighting», *Principles of Data Mining and Knowledge Discovery*, p. 139–147.

Troffaes, M. C. 2007, «Decision making under uncertainty using imprecise probabilities», *International Journal of Approximate Reasoning*, vol. 45, n° 1, p. 17–29.

Tsoumakas, G. et I. Vlahavas. 2007, «Random k-labelsets : An ensemble method for multilabel classification», dans *Machine learning : ECML 2007*, Springer, p. 406–417.

Vovk, V., A. Gammerman et G. Shafer. 2005, *Algorithmic learning in a random world*, Springer Science & Business Media.

Walley, P. 1991, *Statistical reasoning with imprecise probabilities*, Chapman and Hall.

Walley, P. 1996, «Inferences from multinomial data : learning about a bag of marbles», *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 3–57.

Wu, T.-F., C.-J. Lin et R. C. Weng. 2004, «Probability estimates for multi-class classification by pairwise coupling», *Journal of Machine Learning Research*, vol. 5, p. 975–1005.

Zadrozny, B., J. Langford et N. Abe. 2003, «Cost-sensitive learning by cost-proportionate example weighting», dans *Proceedings of the Third IEEE International Conference on Data Mining*, IEEE Computer Society, p. 435–442.

Zaffalon, M. 2002, «The naive credal classifier», *Journal of Statistical Planning and Inference*, vol. 105, n° 1, p. 5–21.

Zaffalon, M., G. Corani et D. Mauá. 2012, «Evaluating credal classifiers by utility-discounted predictive accuracy», *International Journal of Approximate Reasoning*, vol. 53, n° 8, p. 1282 – 1301.