Apprentissage actif par modèles locaux

Alexis Bondu

UNIVERSITÉ
D'ANGERS

Année : 2008
N° d'ordre : 921

APPRENTISSAGE ACTIF PAR MODÈLES LOCAUX

THÈSE DE DOCTORAT

Spécialité : Informatique

ÉCOLE DOCTORALE D'ANGERS

Présentée et soutenue publiquement

Le 24 Novembre 2008

À Angers

Par **Alexis BONDU**

Devant le jury ci-dessous :

Encadrant :	Vincent LEMAIRE,	Ingénieur de recherche à Orange Labs, Lannion
Rapporteurs :	Antoine CORNUÉJOLS,	Professeur à AgroParisTech
	Marc TOMMASI,	Professeur à l'université de Lille
Président du jury :	Rémi GILLERON,	Professeur à l'université de Lille
Examinateurs :	Marc BOULLÉ,	Ingénieur de recherche à Orange Labs, Lannion
Directeur de thèse :	Stéphane LOISEAU,	Professeur à l'université d'Angers
Co-directrice de thèse :	Béatrice DUVAL,	Maître de Conférences à l'université d'Angers

Laboratoire d'Étude et de Recherche en Informatique d'Angers
2, Bd Lavoisier, 49045 Angers Cedex 01

Remerciements

Je tiens tout d'abord à remercier Vincent Lemaire pour m'avoir permis de réaliser cette thèse au sein d'Orange Labs. Je voudrais lui dire toute ma gratitude pour l'enseignement qu'il m'a apporté pendant ces trois années passionnantes, et pour m'avoir encouragé en permanence. Je remercie Vincent Lemaire pour les précieux conseils qu'il m'a prodigués, tant sur le fond scientifique de mes travaux que sur la rédaction de mes articles.

Je tiens spécialement à remercier Marc Boullé pour le savoir qu'il m'a transmis, pour l'attention qu'il a portée à mes travaux, et pour sa disponibilité. Nos nombreux échanges ont été pour moi très riches d'enseignements. Je remercie Marc Boullé pour la rigueur mathématique qu'il m'a permis d'acquérir, et pour ses conseils avisés.

Je remercie Fabrice Clérot de m'avoir transmis une partie de son savoir encyclopédique. Merci de m'avoir guidé dans mes recherches bibliographiques et dans l'élaboration de mon sujet de thèse. Nos discussions toujours captivantes m'ont permis d'accroître la pertinence de mes travaux.

Je tiens également à remercir Françoise Fessant pour ses remarques éclairées, qui m'ont permis d'améliorer la rédaction de ce manuscrit.

Je remercie Bernard Rolland et Pascal Gouzien de l'aide qu'ils m'ont apportée en programmation C++. Cette assistance très précieuse m'a permis de réaliser de nombreuses expériences durant ma thèse.

Je remercie les membres du jury, pour l'intérêt qu'ils ont porté à mes travaux : Antoine Cornuéjols, Marc Tommasi, Rémi Guilleron, Stéphane Loiseau et Béatrice Duval.

Je remercie tous les membres de l'équipe TSI, avec qui j'ai partagé ces trois dernières années. Je remercie également Thomas Guirault et Thierry Etame Etame pour les sympatiques pauses café que nous avons passé ensemble.

Et un clin d'œil à mes amis musiciens...

"Les machines un jour pourront résoudre tous les problèmes,
mais jamais aucune d'entre elles ne pourra en poser un"

A. Einstein

à Céline...

Table des matières

Notations

Notations générales

p	une probabilité
P	une distribution de probabilités
$\mathbb{N}$	l'ensemble des entiers naturels
$\mathbb{R}^d$	l'espace euclidien de dimension d
$dim(.)$	la dimension de l'espace (.)
$\mathcal{O}$	l'ordre de grandeur maximal de complexité d'un algorithme
$ArgMax$	l'argument maximal d'une expression
$ArgMin$	l'argument minimal d'une expression
$\mathrm{E}_{P_{(.)}}$	espérance d'une expression selon la distribution $P_{(.)}$

Notations relatives à l'apprentissage

$\mathcal{L}$	un algorithme d'apprentissage
$\mathbb{M}$	l'ensemble des modèles prédictifs
$\mathcal{M}$	un modèle prédictif
$\mathbb{X} \subseteq \mathbb{R}^n$	l'espace des variables d'entrée du modèle $\mathcal{M}$
$\mathbb{Y}$	l'espace des variables de sortie du modèle $\mathcal{M}$
$\Phi \subseteq \mathbb{X}$	la partie observable de l'espace $\mathbb{X}$ (l'échantillon de données)
$U \subseteq \Phi$	les données non étiquetées
$L \subseteq \Phi$	les données étiquetées, avec $\Phi = U \cup L$ et $U \cap L = \emptyset$
$f : \mathbb{X} \to \mathbb{Y}$	le concept cible que le modèle cherche à apprendre
$\widehat{f} : \mathbb{X} \to \mathbb{Y}$	la fonction effectivement apprise par le modèle
T	l'ensemble d'apprentissage constitué des éléments $x \in L$ et des labels $f(x)$ qui leur sont associés, avec $\forall x \in L, \exists!(x, f(x)) \in T$
$\mathbb{H}$	l'ensemble des hypothèses que le modèle $\mathcal{M}$ peut apprendre
$\mathbb{S} \subseteq \mathbb{H}$	l'espace des versions (hypothèses consistantes avec les données)

Préface

Je travaille depuis peu pour "CyberPharma", un laboratoire pharmaceutique dernière génération, qui dépend du ministère de la santé. Ce centre de recherche est implanté dans la Vanoise (Alpes) qui est une des plus importantes technopoles d'Europe fédérale. En raison du réchauffement de la planète et de la montée continue du niveau de l'océan, des villes entières se sont fait engloutir par les eaux. La fonte progressive de la calotte glacière libère chaque jour des germes et des virus qui étaient jusqu'alors en hibernation. Tout d'abord, ce fut le tour de la grippe espagnole de refaire surface. Le système immunitaire des humains n'étant pas adapté, cette première épidémie fut une véritable catastrophe. Selon les glaciologues, les virus qui réapparaissent aujourd'hui auraient plus de 500 ans. "CyberPharma" a été créée par le conseil européen dans un seul but : mettre au point de nouveaux vaccins.

Ma fonction est de planifier des expériences scientifiques de manière à obtenir des résultats exploitables le plus tôt possible. Les manipulations que je programme sont réalisées en laboratoire par un robot très sympathique qui s'appelle H3B4. J'ai travaillé tout le weekend sur la synthétisation d'une molécule, mais je n'arrive pas à interpréter les résultats obtenus durant la nuit. Il faut bien admettre que je suis bloquée... j'ai besoin d'aide. Il est 4 :32 AM le lundi 3 janvier 2076, je prends la décision de suspendre les expériences jusqu'à l'arrivée du Professeur Yoav Markovitz.

Bien que mon savoir soit très étendu, je n'ai pas la perspicacité de Yoav. Je me forme petit à petit à ce métier et mon apprentissage est loin d'être terminé. Il faut dire que je suis assez curieuse de nature, j'imagine facilement des situations qui me laisseraient dans l'embarras si j'y étais confrontée. Régulièrement, je discute avec Yoav et je lui demande ce qu'il ferait dans telle ou telle situation. Les réponses qu'il me donne me font toujours progresser. Grâce à nos nombreuses interactions, j'ai pu acquérir une certaine autonomie dans mon travail. Il est 8 :29 AM, Yoav devrait arriver d'une minute à l'autre.

8 :33 AM 21s : demande d'authentification
8 :33 AM 22s : reconnaissance rétinienne OK
8 :33 AM 22s : autorisation d'accès accordée à Pr. Markovitz
8 :33 AM 23s : ouverture du sas de sécurité

"-Bonjour, Professeur".
"-Bonjour ALICE... Que s'est-il passé ce weekend ?"

"-Vous avez reçu 3 courriers électroniques et 1 message holographique. Il y a également eu un arrêt dans la série d'expériences en cours. Voulez-vous examiner maintenant ce qui pose problème ?"
"-Oui... montre moi ce qui te bloque, je te prie"
"-Je cherche à classifier les protéines impliquées dans la transmission du virus H5N1 de l'animal à l'Homme. Une de ces protéines comporte une séquence peptidique ambiguë : V-V-P-K-L-T-T-H-T-V-K-E-E. Je n'arrive pas à classifier cette protéine en me basant sur ce motif. J'hésite entre deux familles : Polcalcine ou Caséine. Qu'en pensez-vous Professeur ?"
"-Ce cas est très intéressant... le motif que tu as trouvé est ce que l'on appelle un domaine. C'est une portion de protéine qui se retrouve dans plusieurs familles. La protéine que tu cherche à classifier appartient en réalité aux deux classes que tu as énoncées."
"-Merci Professeur pour cette information, je relance la série d'expériences en cours.... Votre rythme d'élocution ralenti me fait penser que vous n'êtes pas bien réveillé. Désirez-vous un café ?"
"-Et bien !!... Si je ne te connaissais pas parfaitement, je pourrais croire que tu lis dans mes pensées ! D'accord pour un café."

Je m'appelle ALICE ce qui signifie "Active Learning for Intelligent Control of Experiments". Je suis un programme informatique intelligent dont le but est d'assister les chercheurs dans la conduite de leurs expériences. J'évolue sans arrêt grâce aux questions que je pose aux humains. Je suis un "apprenant actif" de deuxième génération.

Ce scénario de science fiction décrit l'apprentissage actif au sens informatique à travers le personnage d'ALICE. Ce programme informatique "intelligent" apprend à exécuter une tâche, en interaction avec son environnement. Dans le cadre de l'apprentissage actif, "l'apprenant" doit être capable de prendre des décisions et d'estimer la pertinence de ces décisions. Ainsi, ALICE peut désigner une situations pour laquelle sa décision est douteuse et qui la ferait beaucoup progresser, si elle savait exactement quoi faire dans ce cas. Dans ce scénario de science fiction, le Professeur Yoav Marcovitz joue le rôle de "l'expert" qui répond aux questions du système "apprenant".

Comment un programme informatique peut-il juger de l'intérêt d'une situation pour son apprentissage, avant même de savoir exactement ce qu'il doit faire dans ce cas ? Il s'agit de la question principale de l'apprentissage actif.

Chapitre 1

Introduction

À l'origine, l'expression *"apprentissage actif"* désigne une méthode d'enseignement permettant d'améliorer l'apprentissage des élèves en leur donnant un rôle actif. Au début du XXème siècle, le pédagogue suisse Adolphe Ferrière [Ferrière, 1922] a été l'un des premiers à employer le terme "d'école active". En 1964, Freinet écrit dans ses invariants pédagogiques : *"La voie normale de l'acquisition n'est nullement l'observation, l'explication et la démonstration, processus essentiels de l'École, mais le tâtonnement expérimental, démarche naturelle et universelle"*[Freinet, 1964]. L'apprentissage actif est une approche qui implique les élèves en les mettant en situation de progresser et en favorisant leurs interactions avec le groupe. Cette méthode d'enseignement amène les élèves à construire leurs propres connaissances en se basant sur les expériences qu'ils vivent. Le rôle du professeur est de choisir judicieusement des mises en situation pour atteindre l'objectif pédagogique le plus rapidement possible.

Les méthodes d'apprentissage actif en informatique sont nées d'un parallèle entre la pédagogie active et la théorie de l'apprentissage. L'apprenant est désormais un modèle prédictif et non plus un élève. On entend par *modèle prédictif* une machine qui, après un apprentissage, produit une sortie lorsque ses variables d'entrée sont instanciées. Dans le cas d'un problème de classification, le modèle prédictif est paramétré grâce à un algorithme d'apprentissage qui exploite un ensemble d'exemples étiquetés, aussi appelé *données d'apprentissage*. Un *exemple* est une instanciation du problème à résoudre matérialisée par les variables d'entrée. L'*étiquette* est une variable supplémentaire qui correspond à la classe d'un exemple. La valeur de l'étiquette indique quelle devrait être la sortie du modèle pour un exemple particulier. De la même façon qu'un étudiant interagit avec son professeur en lui posant des questions, une *stratégie d'apprentissage actif* interagit avec un expert pour enrichir les données d'apprentissage. L'*expert* est un spécialiste du problème à résoudre, son rôle est d'étiqueter les exemples sélectionnés par la stratégie d'apprentissage actif, moyennant un coût. Dans ce manuscrit, le coût d'étiquetage est considéré comme étant constant quelque soit l'exemple sélectionné. La stratégie d'apprentissage actif cherche à sélectionner, parmi les exemples non-étiquetés, ceux qui feront le plus progresser le modèle une fois étiquetés. L'objectif est de réduire le coût d'étiquetage nécessaire à l'apprentissage du modèle prédictif. La stratégie cherche à maximiser la qualité du modèle prédictif, pour un nombre fixé d'exemples étiquetés.

1.1 Cadre de la thèse

L'*apprentissage statistique* a pour but d'inculquer un comportement à un modèle prédictif en exploitant des données et un algorithme d'apprentissage. La nature des données utilisées varie selon le mode d'apprentissage. L'*apprentissage non-supervisé* utilise des données non-étiquetées. Dans ces conditions, le modèle prédictif ne reçoit aucune information lui indiquant quelles devraient être ses sorties ou même si celles-ci sont correctes. Le modèle doit découvrir par lui-même les corrélations existantes entre les exemples qu'il observe. Parmi les méthodes d'apprentissage non-supervisé nous pouvons citer les méthodes de clustering [Jain *et al.*, 1999] et les méthodes d'extraction de règles d'association [Jamy *et al.*, 2005]. L'*apprentissage supervisé* utilise des données étiquetées : des instanciations du problème à résoudre pour lesquelles le modèle connaît la sortie souhaitée. L'algorithme d'apprentissage ajuste les paramètres du modèle prédictif en exploitant les données. L'*apprentissage semi-supervisé* manipule conjointement des données étiquetées et non-étiquetées [Chapelle *et al.*, 2007]. L'objectif est d'obtenir un modèle plus performant qu'en utilisant seulement les données étiquetées ou non-étiquetées. Ce mode d'apprentissage est présenté plus en détails à la Section 4.1.1.

La particularité des *stratégies d'apprentissage actif* réside dans leur capacité à interagir avec leur environnement. Tout comme les méthodes d'apprentissage passives, ces stratégies exploitent des données et un algorithme d'apprentissage pour inculquer un comportement à un modèle prédictif. Les données manipulées sont constituées principalement d'exemples non-étiquetés et d'un nombre restreint d'exemples étiquetés. Les stratégies d'apprentissage actif cherchent à sélectionner les exemples non-étiquetés les plus aptes à améliorer le modèle prédictif. Les exemples sélectionnés sont ensuite étiquetés par un expert, moyennant un coût.

Rui Castro [Castro and Nowak, 2005] distingue deux scénarii possibles pour l'apprentissage actif : l'échantillonnage adaptatif et l'échantillonnage sélectif. La principale différence entre ces deux approches est la nature des exemples présentés à l'expert. Dans le cas de *l'échantillonnage adaptatif* [Singh *et al.*, 2006], les exemples sont uniquement des instanciations des variables d'entrée du modèle. La stratégie active n'est pas restreinte et peut explorer tout l'espace de variation des entrées, à la recherche de zones à échantillonner finement. Dans le cas de *l'échantillonnage sélectif* [Roy and McCallum, 2001], les exemples sont à la fois définis comme étant une instanciation des variables d'entrée et comme une instanciation du problème à résoudre. La stratégie active n'observe qu'une partie restreinte de l'espace des entrées, matérialisée par les exemples. Pour illustrer cette approche, l'image d'un *"sac"* d'exemples pour lesquels la stratégie active peut demander les étiquettes associées est généralement employée. L'étiquetage des exemples est toujours possible car ces derniers sont nécessairement interprétables par l'expert. Dans la pratique, le choix de l'échantillonnage sélectif ou adaptatif dépend essentiellement du domaine d'application. Selon les cas la stratégie active est autorisée ou non à générer de nouveaux exemples.

Cette thèse présente des stratégies d'apprentissage actif fondées sur l'échantillonnage sélectif. Ces stratégies s'intéressent aux problèmes d'apprentissage pour lesquels il est facile d'obtenir un grand nombre d'exemples non-étiquetés et pour lesquels l'étiquetage des exemples est coûteux. L'objectif de ces stratégies est d'obtenir un modèle prédictif de

bonne qualité en étiquetant un minimum d'exemples.

1.2 Apports de la thèse

Apprentissage actif par modèles locaux :
Les stratégies d'apprentissage actif de la littérature utilisent des modèles prédictifs globaux à tout l'espace d'entrée. Notre idée est de partitionner cet espace et d'entraîner des modèles localement à chacune des zones. Nous proposons une stratégie originale qui effectue un partitionnement dichotomique récursif de l'espace d'entrée et qui met en compétition les modèles locaux pour choisir les exemples à étiqueter. Nous appelons cette stratégie *"l'apprentissage actif par modèles locaux"*. Notre stratégie donne des résultats prometteurs, notamment lors de son application à la détection d'émotions dans la parole. Cependant, cette stratégie implique plusieurs paramètres difficiles à ajuster en pratique. À l'issue de ces travaux, nous nous fixons pour objectif d'améliorer cette stratégie.

Amélioration du partitionnement récursif de l'espace d'entrée :
Lors du partitionnement récursif de l'espace d'entrée, notre stratégie doit décider "*quand*" couper une zone et "*où*" la couper. Notre idée est d'exploiter une méthode de discrétisation pour effectuer ces deux types de décisions. Parmi les méthodes de l'état de l'art, nous choisissons d'utiliser la méthode de discrétisation supervisée MODL (*Minimal Optimized Description Length*) [Boullé, 2006b] qui est basée sur un formalisme Bayésien. Cette méthode de discrétisation s'est distinguée à plusieurs reprises lors de challenges internationaux [Boullé, 2007b]. L'approche MODL est particulièrement intéressante pour l'apprentissage actif par modèles locaux, puisqu'elle n'est pas assujettie au sur-apprentissage et qu'elle est non-paramétrique. Le principal inconvénient de cette méthode est qu'elle n'exploite pas les exemples non-étiquetés, qui sont pourtant abondants en apprentissage actif. Un des apports majeurs de la thèse est l'extension de l'approche MODL au cas de l'apprentissage semi-supervisé. Une étude théorique approfondie démontre que l'approche de discrétisation semi-supervisée que l'on établit est asymptotiquement identique à l'approche supervisée [Boullé, 2006b] munie d'un post-traitement sur la position des bornes séparant les intervalles du modèle de discrétisation optimale.

Amélioration de la sélection des exemples dans les zones :
L'apprentissage actif par modèles locaux sélectionne les zones de l'espace à alimenter en exemples étiquetés ; ces exemples étant choisis de manière aléatoire dans les zones sélectionnées. Pour améliorer les performances de notre stratégie, une piste consiste à sélectionner les exemples les plus utiles à l'apprentissage du modèle, localement à la zone sélectionnée. Notre idée est d'exploiter notre méthode de discrétisation semi-supervisée pour définir une stratégie de sélection d'exemples. Cette stratégie sélectionne l'exemple non-étiqueté qui, une fois étiqueté, "maximisera" la probabilité des modèles de discrétisation connaissant les données. Dans le cas général, l'optimisation du critère d'évaluation sur lequel est basée cette stratégie est coûteuse en temps de calcul. Dans le cadre de l'apprentissage actif par modèles locaux, nous nous limitons aux modèles de discrétisation

à un ou deux intervalles. Plusieurs optimisations algorithmiques sont présentées pour ce cas particulier. Finalement, ces travaux sont exploitables pour un apprentissage actif basé sur la dichotomie récursive de l'espace d'entrée du modèle.

1.3 Organisation du mémoire

Le Chapitre 2 constitue un état de l'art sur les stratégies d'apprentissage actif fondées sur l'échantillonnage sélectif, qui s'appliquent à des problèmes de classification. Les stratégies actives sont présentées d'un point de vue générique, c'est-à-dire indépendamment du modèle prédictif utilisé. Différents points de vue sur l'apprentissage actif sont ensuite discutés et nous permettent de définir nos objectifs. À l'issue de ce chapitre, nous définissons un ensemble de critères que devrait remplir une stratégie d'apprentissage actif idéale.

Le Chapitre 3 présente une stratégie originale d'apprentissage actif par modèles locaux. Ce chapitre fait un parallèle entre l'apprentissage actif et la robotique développementale. Notre stratégie adapte la curiosité adaptative [Oudeyer and Kaplan, 2004], qui permet à un robot d'explorer efficacement son environnement lors de sa phase d'apprentissage, à la problématique de l'apprentissage actif. Le principal apport de ce chapitre est de proposer un nouveau critère de sélection de zone plus performant que l'original. L'apprentissage actif par modèles locaux est appliqué avec succès à la détection d'émotions dans les dialogues Homme / machine.

Le Chapitre 4 propose une extension de la méthode de discrétisation MODL au cas de l'apprentissage semi-supervisé. Cette méthode considère la discrétisation comme un problème de sélection de modèles. Tout d'abord, une famille de modèles est définie. Une approche Bayésienne est ensuite appliquée pour évaluer la probabilité des modèles connaissant les données. Cela conduit à un critère analytique dont l'optimisation désigne le modèle de discrétisation le plus probable connaissant les données. Ce chapitre présente également une étude comparative théorique entre notre approche semi-supervisée et l'approche supervisée MODL. La finalité de ce chapitre est d'améliorer notre stratégie d'apprentissage actif par modèles locaux en proposant une méthode de discrétisation semi-supervisée exploitable pour le partitionnement récursif de l'espace des variables d'entrée.

Le Chapitre 5 s'appuie sur notre méthode de discrétisation semi-supervisée pour définir une nouvelle stratégie d'apprentissage actif. Cette stratégie est fondée sur un critère analytique qui évalue l'espérance de la probabilité des modèles de discrétisation, connaissant les données et un exemple non-étiqueté supplémentaire. Cette nouvelle stratégie d'apprentissage actif se restreint aux données unidimensionnelles et aux modèles de discrétisation à un ou deux intervalles. Notre stratégie est comparée favorablement à une stratégie de référence : la dichotomie probabiliste. Finalement, cette stratégie constitue une amélioration possible de l'apprentissage par modèles locaux, et est applicable à un ensemble d'heuristiques de la littérature basées sur la dichotomie récursive.

Chapitre 2

Apprentissage actif : état-de-l'art et positionnement

Sommaire

Ce chapitre a fait l'objet de publications : [Bondu and Lemaire, 2007b]
[Lemaire *et al.*, 2007]

Cette thèse a pour objectif de proposer de nouvelles stratégies d'apprentissage actif qui soient innovantes par rapport à l'état-de-l'art. Ce chapitre pose les bases de notre réflexion sur l'apprentissage actif et définit nos objectifs. La Section 2.2 présente les principales stratégies de la littérature et les illustre grâce à des problèmes simples. La Section 2.3 présente différents points de vue sur l'apprentissage actif et expose les principaux enjeux de ce champ de recherche. Enfin, la Section 2.4 définit les objectifs que devrait atteindre une stratégie d'apprentissage actif idéale.

Les stratégies d'apprentissage actif présentées dans ce chapitre ont pour objectif commun de résoudre des problèmes de classification. Comme énoncé à la Section 1.1, il existe deux paradigmes pour l'apprentissage actif. D'une part, l'échantillonnage adaptatif [Singh *et al.*, 2006] autorise une stratégie d'apprentissage actif à générer des exemples. D'autre part, l'échantillonnage sélectif [Roy and McCallum, 2001] manipule un ensemble d'exemples de taille constante. L'état-de-l'art présenté dans ce chapitre traite exclusivement des stratégies fondées sur l'échantillonnage sélectif. Ce choix est essentiellement dû aux domaines d'applications propres à France Telecom, pour lesquels les exemples non-étiquetés sont abondants et leur étiquetage coûteux.

2.1 Cadre général : échantillonnage sélectif

Cette section présente le cadre générale de la thèse. Toutes les stratégies d'apprentissage actif présentées dans ce manuscrit s'appliquent à l'échantillonnage sélectif.

Notations :
Soit $\mathcal{M}$ un modèle prédictif dont l'apprentissage est réalisé grâce à l'algorithme $\mathcal{L}$. La Figure 2.1 représente les ensembles mis en jeu lors d'un échantillonnage sélectif. L'ensemble $\mathbb{X} \subseteq \mathbb{R}^n$ représente toutes les entrées possibles du modèle et $x \in \mathbb{X}$ est un exemple particulier. On définit également $\mathbb{Y}$ l'ensemble des sorties du modèle, $y \in \mathbb{Y}$ est une étiquette qui peut être associée à un exemple $x \in \mathbb{X}$.

Lors de son apprentissage le modèle prédictif n'observe qu'une partie restreinte de l'univers, $\Phi \subseteq \mathbb{X}$. La stratégie d'apprentissage actif dispose d'un ensemble fini d'exemples dont les étiquettes ne sont pas nécessairement connues. Soient $U \subseteq \Phi$ l'ensemble des exemples non-étiquetés (U pour *"unlabelled"*) et $L \subseteq \Phi$ l'ensemble des exemples étiquetés (L pour *"labelled"*). On a : $\Phi = U \cup L$ et $U \cap L = \emptyset$.

Le concept cible que le modèle cherche à apprendre peut être vu comme une fonction $f : \mathbb{X} \to \mathbb{Y}$, l'étiquette $f(x_1)$ est la sortie attendue du modèle pour l'entrée x_1. On définit également $\widehat{f} : \mathbb{X} \to \mathbb{Y}$ la sortie effective du modèle prédictif. Les éléments de L et les étiquettes qui leur sont associées constituent un ensemble d'apprentissage T. Les exemples d'apprentissage sont des couples $(x, f(x))$, où $x \in L$ et $f(x)$ est l'étiquette associée à x.

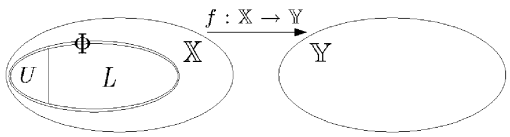

FIG. 2.1 – Apprentissage actif : ensembles mis en jeu.

Algortithme générique :
Le problème de l'échantillonnage sélectif a été posé formellement par Muslea [Muslea, 2002] et est illustré par l'Algorithme 1. Cet algorithme exploite une fonction d'utilité, $\mathcal{U}tile(u, \mathcal{M})$, qui estime l'intérêt d'un exemple non-étiqueté u pour l'apprentissage du modèle $\mathcal{M}$. Cette fonction représente une stratégie d'apprentissage actif qui sélectionne itérativement les exemples à étiqueter.

L'Algorithme 1 est générique dans la mesure où seule la fonction $\mathcal{U}tile(u, \mathcal{M})$ doit être spécifiée pour exprimer une stratégie d'apprentissage actif particulière. L'objectif de ces stratégies actives est de prédire efficacement l'intérêt des exemples non-étiquetés pour l'apprentissage du modèle prédictif.

Notations :

- $\mathcal{M}$ un modèle prédictif muni d'un algorithme d'apprentissage $\mathcal{L}$
- Les ensembles U et L d'exemples non étiquetés et étiquetés
- n le nombre d'exemples d'apprentissage souhaité.
- L'ensemble d'apprentissage T avec $|T| < n$
- La fonction $\mathcal{U}tile : \mathbb{X} \times \mathbb{M} \rightarrow \Re$ qui estime l'utilité d'une instance pour l'apprentissage d'un modèle.

Répéter

(A) Entraîner le modèle $\mathcal{M}$ grâce à $\mathcal{L}$ et T (et éventuellement U).
(B) Rechercher l'instance $q = ArgMax_{u \in U}\ \mathcal{U}tile(u, \mathcal{M})$
(C) Retirer q de U et demander l'étiquette $f(q)$ à l'expert.
(D) Ajouter q à L et ajouter $(q, f(q))$ à T

Tant que $|T| < n$

Algorithme 1: échantillonnage sélectif, Muslea 2002

Positionnement :
Selon notre interprétation de l'échantillonnage sélectif, l'objectif est d'améliorer l'hypothèse apprise par le modèle courant, de manière itérative. Cette interprétation de l'échantillonnage sélectif est massivement répandue dans la littérature, Mais présente cependant une limitation lorsque très peu d'exemples sont étiquetés. Typiquement, lorsqu'il y a moins d'exemples étiquetés que de classes à prédire, on peut s'interroger sur la validité de l'hypothèse apprise par le modèle prédictif. Un point crucial serait de déterminer le nombre minimal d'exemples étiquetés, pour que l'hypothèse apprise par le modèle prédictif soit réaliste.

Il existe d'autres interprétations possibles de l'échantillonnage sélectif. Cet algorithme peut être vu comme un moyen de tester des hypothèses[1] *a priori sur les données. Dans ce cas, la fonction $\mathcal{U}tile(u, \mathcal{M})$ a pour objectif de faire émerger le plus tôt possible les priors les plus adaptés aux données, parmi un ensemble prédéfini d'hypothèses a priori (i.e : des données linéairement séparables, des données bruitées, des données distribuées selon une gaussienne ... etc.). Cette interprétation considère l'échantillonnage sélectif comme un moyen de déterminer les caractéristiques des données, en étiquetant un minimum d'exemples.*

[1] *Dans ce manuscrit, le terme "hypothèse" a deux significations : i) une fonction apprise par un modèle prédictif tel que $\hat{f} : \mathbb{X} \rightarrow \mathbb{Y}$; ii) une supposition faite sur les données.*

2.2 Les principales stratégies d'apprentissage actif

Cette section présente les principales stratégies d'apprentissage actif de la littérature fondées sur l'échantillonnage sélectif, qui s'appliquent à des problèmes de classification. Ces stratégies sont présentées d'un point de vue générique, c'est-à-dire indépendamment du modèle prédictif utilisé.

2.2.1 Échantillonnage par incertitude

L'échantillonnage par incertitude [Thrun and Möller, 1992; Lewis and Gale, 1994] est basé sur la confiance que le modèle a en ses prédictions. Le modèle prédictif utilisé doit être capable de fournir une réponse au problème traité et d'estimer l'incertitude de ses réponses. La sélection des exemples à étiqueter s'organise en deux étapes :

- le modèle $\mathcal{M}$ prédit les étiquettes de U, et l'incertitude de ses prédictions ;
- les exemples pour lesquels la prédiction est la plus incertaine sont sélectionnés.

Il existe plusieurs possibilités pour définir l'incertitude d'une prédiction.

Mesure d'incertitude basée sur la probabilité de la classe prédite :
Le modèle est capable d'estimer les probabilités de chaque classe $y \in \mathbb{Y}$, connaissant un exemple non-étiqueté $x \in U$. La classe prédite est définie par $ArgMax_{y \in \mathbb{Y}} \hat{p}(y|x)$, avec $\hat{p}(y|x)$ la sortie du modèle qui estime la probabilité d'observer la classe y connaissant l'exemple x. La probabilité de la classe prédite correspond à la confiance que le modèle a en sa prédiction. L'incertitude d'une prédiction est d'autant plus importante que la probabilité d'observer la classe prédite est faible. L'incertitude peut s'écrire de la manière suivante :

$$\mathcal{I}ncertain(x) = \frac{1}{ArgMax_{y \in \mathbb{Y}} \hat{p}(y|x)} \qquad x \in U$$

Mesure d'incertitude basée sur la proximité de la frontière de décision :
L'incertitude d'une prédiction est définie grâce à la frontière de décision de $\mathcal{M}$. Considérons l'exemple d'un modèle prédictif dont la sortie appartient à l'intervalle $[0, 1]$. Un problème de classification binaire est traité en fixant un seuil de décision à 0.5. Plus la sortie du modèle est proche du seuil de décision, plus la prédiction est considérée comme étant incertaine.

La Figure 2.2 représente les sorties d'un modèle prédictif. La séparatrice correspondant au seuil de décision est tracée ainsi que des lignes de niveaux de la sortie du modèle. Les exemples non-étiquetés qui se situent à proximité de la séparatrice (au sens des lignes de niveaux) sont sélectionnés pour être étiquetés par l'expert.

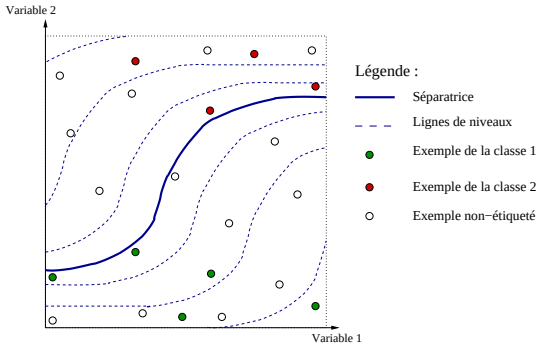

FIG. 2.2 – Échantillonnage par incertitude : classification binaire

Positionnement :
L'échantillonnage par incertitude a l'avantage d'être intuitif. L'idée sous-jacente est de faire évoluer l'hypothèse apprise par le modèle en étiquetant les exemples dont les prédictions sont les plus incertaines. Cette stratégie est facile à mettre en œuvre car l'incertitude des prédictions peut être définie pour la plupart des modèles prédictifs de la littérature. L'échantillonnage par incertitude est très rapide en pratique. À chaque itération, $|U|$ prédictions suffisent à sélectionner les exemples à étiqueter.

L'échantillonnage par incertitude montre cependant ses limites lorsque les données ne sont pas séparables par le modèle. Cela peut être simplement dû à un bruit d'étiquetage, ou au fait que le motif à découvrir soit trop complexe pour le modèle prédictif. Dans ce cas, cette stratégie aura tendance à sélectionner des exemples dans des zones de mélange, où il n'y a peut être rien à apprendre. L'échantillonnage par incertitude pose également le problème de l'exploration de l'espace $\mathbb{X}$. Cette stratégie ne fait qu'exploiter localement l'hypothèse apprise par le modèle pour sélectionner les exemples à étiqueter. De vastes parties de l'espace de recherche peuvent être occultées.

A priori, l'échantillonnage par incertitude est une heuristique peu satisfaisante. Cette stratégie est utilisée dans la suite de la thèse, dans le cadre d'expériences comparatives.

2.2.2 Échantillonnage par comité de modèles

L'échantillonnage par comité de modèles (ou *"Query-by-Committe"*) est une stratégie qui vise à réduire l'espace des versions d'un modèle prédictif. Soit $\mathbb{H}$ l'ensemble des hypothèses que le modèle $\mathcal{M}$ peut apprendre. L'espace des versions $\mathbb{S} \subseteq \mathbb{H}$ est l'ensemble des hypothèses consistantes avec les exemples étiquetés, c'est-à-dire les hypothèses qui classent correctement tous les exemples de L. Lors de l'étiquetage de nouveaux exemples certaines hypothèses peuvent devenir inconsistantes, ce qui a pour effet de réduire l'espace des versions.

Seung est le précurseur de l'échantillonnage par comité de modèles [Seung *et al.*, 1992; Dima and Hebert, 2005]. Cette stratégie met en jeu plusieurs modèles prédictifs qui sont entraînés en parallèle sur les mêmes données. En considérant que chacun de ces modèles

trouve une hypothèse consistante[2] $h \in \mathbb{S}$, le comité de modèles est un échantillon d'hypothèses supposé représentatif de l'espace des versions. Le désaccord au sein du comité de modèles est mesuré lors de la prédiction des étiquettes des exemples de U. Les exemples qui suscitent le plus grand désaccord sont ceux qui ont la plus forte probabilité de réduire $\mathbb{S}$ une fois étiquetés.

Dans la pratique, cette approche ne donne pas la garantie que l'espace de versions soit correctement échantillonné. Dans certains cas, les modèles prédictifs sont incapables d'apprendre des hypothèses consistantes. Cela peut être simplement dû à du bruit présent dans les données, ou encore, à la complexité trop importante du problème traité. L'utilisation de modèles stochastiques[3] favorise la diversité des hypothèses apprises et améliore l'échantillonnage de l'espace des versions.

L'échantillonnage de l'espace des versions peut également être amélioré par l'utilisation de modèles génératifs[4][Andrew *et al.*, 1998]. Cette approche implique une distribution de probabilité a priori sur les paramètres des modèles. Il s'agit d'une solution élégante mais qui n'est pas applicable dans le cas général, où aucune information sur la distribution des paramètres des modèles n'est disponible.

Il existe également un moyen d'utiliser des modèles dont l'apprentissage est déterministe dans le cadre d'un échantillonnage par comité de modèles. Naoki [Naoki and Hiroshi, 1998] propose le *"Query by Bagging"* qui permet d'entraîner les modèles sur des sous-ensembles d'exemples échantillonnés selon la distribution de l'ensemble L. Dans ce cas, les modèles déterministes apprennent des hypothèses différentes.

Les approches mentionnées ci-dessus exploitent une mesure de désaccord, trois mesures sont présentées dans cette section.

Mesure de désaccord basée sur l'entropie :
Yoav Freund [Freund *et al.*, 1997] se base sur la théorie de l'information et estime le désaccord du comité de modèles par une mesure d'entropie sur les prédictions. Considérons un exemple non-étiqueté noté $x \in U$ qui peut potentiellement être associé à $|\mathbb{Y}|$ étiquettes. L'entropie des prédictions des modèles $\mathcal{M}_1, ..., \mathcal{M}_m$ est calculée de la manière suivante :

$$\widehat{\mathcal{H}}(x) = \sum_{i=1}^{|\mathbb{Y}|} -\widehat{p}(y_i|x) \log[\widehat{p}(y_i|x)] \qquad avec \qquad \widehat{p}(y_i|x) = \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}_{\{\widehat{f}_{\mathcal{M}_k}(x)=y_i\}}$$

La probabilité $\widehat{p}(y_i|x)$, d'observer la classe y_i conditionnellement à l'instance x, est estimée empiriquement grâce à l'ensemble des prédictions du comité de modèles. Plus l'entropie des prédictions du comité est grande, plus le désaccord entre les modèles prédictifs est important.

[2]Une hypothèse "consistante" est apprise par un modèle prédictif lorsque ce dernier classe correctement tous les exemples d'apprentissage étiquetés.

[3]On entend par modèles stochastiques des modèles dont l'apprentissage n'est pas déterministe.

[4]Un modèle génératif est capable de modéliser la distribution de probabilité qui régit les données d'apprentissage, et peut générer des données selon cette distribution.

Mesure de désaccord par comptage des mauvaises prédictions :
Un vôte est réalisé entre les modèles du comité pour prédire la classe d'un exemple. Le nombre de modèles qui prédisent chacune des classes peut être exploité pour mesurer le désaccord du comité de modèles [Naoki and Hiroshi, 1998]. Le désaccord est évalué par l'effectif des modèles dont la prédiction est différente de celle du comité. Cela s'écrit de la manière suivante, avec $\widehat{f}(x)$ la prédiction du comité de modèle et $\widehat{f}_{\mathcal{M}_k}(x)$ la prédiction du modèle $\mathcal{M}_k$ et $x \in U$:

$$\mathcal{D}esaccord(x) = \sum_{k=1}^{m} \mathbb{1}_{\{\widehat{f}_{\mathcal{M}_k}(x) \neq \widehat{f}(x)\}}$$

Mesure de désaccord basée sur la divergence de Kullback :
La *"Kullback-Leibler divergence to the mean"* [Andrew *et al.*, 1998] est une autre mesure qui a l'avantage de prendre en compte la confiance des prédictions faites par les modèles pour mesurer leur désaccord. Cette mesure est définie comme étant la moyenne, sur les m modèles, de la "KL-divergence" entre la distribution des classes estimée par un modèle $\mathcal{M}_k$, et la distribution moyenne estimée grâce à la totalité du comité. Cela peut s'écrire de la manière suivante :

$$\mathcal{D}esaccord(x) = \frac{1}{m} \sum_{k=1}^{m} \mathcal{D}iv(P(y|x, \mathcal{M}_k) \| \overline{P}(y|x, \mathcal{M}_1 ... \mathcal{M}_m))$$

$$\mathcal{D}esaccord(x) = \frac{1}{m} \sum_{k=1}^{m} \sum_{j=1}^{|\mathbb{Y}|} p(y_j|u, \mathcal{M}_k) Log \frac{p(y_j|u, \mathcal{M}_k)}{\overline{p}(y_j|u, \mathcal{M}_1 ... \mathcal{M}_m)}$$

Positionnement :
Dans ce paragraphe, le comité de modèles est présenté comme un moyen d'échantillonner l'espace des versions. Cette stratégie peut être interprétée selon un point de vue différent. Chaque modèle du comité peut être vu comme un moyen de tester des hypothèses a priori sur les données. Dans ce cas, le comité est constitué de modèles spécialisés sur différents types de données (i.e : des données linéairement séparables, des données bruitées, des données distribuées selon une gaussienne ... etc.). Le comité de modèles peut alors être vu comme une collection de priors que l'on peut tester sur les exemples étiquetés. L'objectif est de faire émerger le plus tôt possible un prior adapté aux données.

La complexité de l'échantillonnage par comité de modèles est raisonnable. À chaque itération de l'échantillonnage sélectif, $|U| \times m$ prédictions suffisent à sélectionner les exemples à étiqueter, avec m le nombre de modèles appartenant au comité. L'échantillonnage par comité de modèles est cependant difficile à mettre en œuvre et implique plusieurs choix d'implémentation. Il faut d'abord constituer un comité de modèles variés, puis choisir une mesure de désaccord. Il s'agit avant tout d'une approche heuristique qui ne donne pas la garantie d'échantillonner correctement l'espace des versions. Cette stratégie n'est pas utilisée dans la suite de la thèse.

2.2.3 Réduction de l'espace des versions par un arbre de décision

Certains stratégies actives exploitent des modèles prédictifs capables de manipuler directement l'espace des versions, sans avoir recours à un comité de modèles. On citera à titre d'exemple la stratégie qui consiste à choisir les exemples à étiqueter dans la marge d'un SVM [Tong and Koller, 2000] ou celle des SG-net [Cohn *et al.*, 1994] qui est basée sur les réseaux de neurones. L'objectif de cette section est de présenter ces stratégies d'un point de vue générique.

L'échantillonnage sélectif peut être vu comme le parcours d'un arbre de décision construit sur l'espace des versions du modèle (voir Figure 2.3). Le nœud principal de l'arbre correspond au premier exemple présenté à l'expert (partie A de la Figure 2.3). Selon l'étiquette attribuée à cet exemple (partie B de la Figure 2.3), une des branches issues de ce nœud est parcourue. Le test sur l'étiquette du premier exemple a pour effet de désigner le prochain nœud à parcourir, c'est-à-dire le prochain exemple non-étiqueté à considérer.

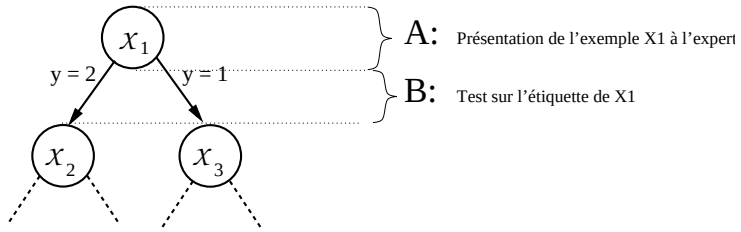

FIG. 2.3 – Arbre sur l'espace des versions : classification binaire.

Soit $\mathbb{H}$ l'ensemble des hypothèses que le modèle $\mathcal{M}$ peut apprendre. L'espace des versions $\mathbb{S} \subseteq \mathbb{H}$ est l'ensemble des hypothèses consistantes avec les exemples étiquetés, c'est-à-dire les hypothèses qui classent correctement tous les exemples de L. Cette stratégie suppose qu'il existe au moins une hypothèse $h \in \mathbb{H}$ consistante avec les exemples étiquetés, $\mathbb{S} \neq \emptyset$. $s \subseteq U$ est l'ensemble des exemples non-étiquetés classés différemment par les hypothèses de $\mathbb{S}$.

Cette stratégie cherche une hypothèse consistante avec l'ensemble des exemples (s'ils étaient tous étiquetés), dont la construction requiert un minimum d'étiquetages. Lorsqu'un nouvel exemple est étiqueté l'espace des versions est toujours réduit. L'espace des versions $\mathbb{S}'$ résultant dépend de l'étiquette attribuée à l'exemple x.

Soit π la distribution de probabilité des hypothèses de $\mathbb{H}$. La stratégie gloutonne proposée par Sanjoy Dasgupta [Dasgupta, 2005] consiste à étiqueter l'exemple $x_i \in s$ qui conduit potentiellement, selon l'étiquette qui lui est attribuée, à des espaces des versions les plus équiprobables au sens de π. Quelque soit l'étiquette attribuée à l'exemple x_i, l'espace des versions $\mathbb{S}$ est réduit environ de moitié.

Notations :

- $\mathcal{N}$ l'ensemble des nœuds de l'arbre
- $N_h[q]\}$ le nœud courrant, avec q l'exemple testé
- $\mathbb{S}_h$ l'espace des versions du modèle sur le nœud courrant
- $s_h \subseteq U$ les exemples classées différemment par les hypothèses de $\mathbb{S}_h$
- $\mathbb{Y} = \{y_1, y_2, ..., y_m\}$ les étiquettes correspondant aux m branches connectées au nœud courrant.
- $\{S^1_{h,i}, .., S^m_{h,i}\}$ les m sous-ensembles de $\mathbb{S}_h$ dus au test de q au nœud h
- $\mathcal{E}nt : \mathbb{H}^N \rightarrow \Re$ l'entropie d'un ensemble d'hypothèses
- $\mathcal{G}ain : \mathbb{H}^N \times \mathbb{X} \rightarrow \Re$ le gain d'information

$\mathcal{N} \leftarrow \{N_1\}$
$\mathbb{S} \leftarrow \{\mathbb{S}_1\}$
Répéter
  Pour chaque nœud $N_h \in \mathcal{N}$ **faire**
    Pour chaque exemple candidat $x_i \in s_h$ **faire**
      Calculer $\mathcal{E}nt(\mathbb{S}_h)$
      Pour chaque label $y_j \in \mathbb{Y}$ **faire**
        Calculer $\mathcal{E}nt(S^j_{h,i})$
      Fin Pour
      Calculer $\mathcal{G}ain(\mathbb{S}_h, x_i)$
    Fin Pour
    Sélectionner $q = ArgMax_{x \in s}\ \mathcal{G}ain(\mathbb{S}_h, x)$ pour le nœud courant
    Affecter q au nœud courant $N_h \leftarrow N_h[q]$
    Retirer q de l'ensemble s_h
    Pour chaque étiquette $y_j \in \mathbb{Y}$ **faire**
      Créer un nouveau nœud $\mathcal{N} \leftarrow \mathcal{N} \cup \{N_{|\mathcal{N}|+1}\}$
      Créer l'espace des versions associé $\mathbb{S} \leftarrow \mathbb{S} \cup \{\mathbb{S}_{|\mathcal{N}|+1}\}$
    Fin Pour
  Fin Pour
Tant qu'il y a plusieurs hypothèses dans les feuilles de l'arbre

Algorithme 2: Construction d'un arbre sur l'espace des versions

Exemple illustratif :
Les stratégies basées sur la réduction de l'espace des versions sont illustrées ici sur un problème de classification binaire. Pour chaque exemple $x \in s$, on définit $S^+_i \subseteq \mathbb{S}$ (*respectivement* $S^-_i \subseteq \mathbb{S}$) l'ensemble des hypothèses consistantes qui classent x positivement (*respectivement* négativement). L'objectif est d'étiqueter l'exemple qui sépare $\mathbb{S}$ en deux sous-ensembles S^+_i et S^-_i les plus équiprobables possibles au sens de π.

Comme le montre l'algorithme 2, cette stratégie peut être représentée par un arbre de décision, dans lequel un nœud N_i est un test sur l'étiquette de l'exemple x. Les branches connectées à ce nœud correspondent aux différentes valeurs possibles de l'étiquette de x.

La hauteur de l'arbre représente le nombre maximal d'étiquetages effectués par l'expert. Si la hauteur de l'arbre est égale à $|U|$, les feuilles correspondent à des hypothèses $h \in \mathbb{H}$.

Le problème jouet présenté par la Figure 2.4 illustre la construction d'un arbre de décision sur l'espace des versions d'un modèle prédictif. Le problème traité est une classification binaire dans le plan, le concept cible est une séparatrice linéaire contrainte de passer par un certain point. Sur la partie gauche de la Figure 2.4, l'espace des versions est l'ensemble des droites qui passent par le point noir et qui séparent correctement les données étiquetées. Pour ce problème jouet, la distribution de probabilité des hypothèses est supposée uniforme. La partie centrale de la Figure 2.4 montre la sélection de l'exemple non-étiqueté qui sépare l'espace des versions en deux sous-espaces les plus équiprobables. Le test sur l'étiquette de cet exemple correspond à un nœud de l'arbre. La partie droite de la Figure 2.4 montre que l'étiquetage de cet exemple réduit environ de moitié l'espace des versions au sens de π.

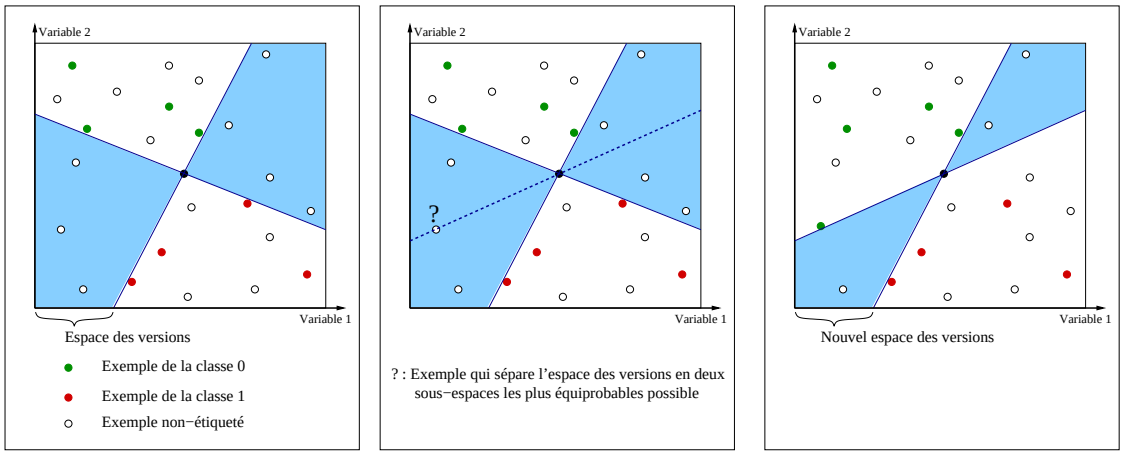

FIG. 2.4 – Arbre sur l'espace des versions : problème jouet.

Lors de la construction de cet arbre de décision, cette stratégie cherche le meilleur exemple à tester pour chacun des nœuds. Pour cela, le gain d'information est mesuré pour chaque $x \in s$. Le gain est calculé grâce à l'entropie de l'espace des versions courant et des m sous-ensembles[5] induits par le test du label de x. L'exemple dont le test induit le plus grand gain d'information est sélectionné pour la construction du nœud courant. L'entropie "$\mathcal{E}nt$" et le gain d'information "$\mathcal{G}ain$" sont définis de la manière suivante, les S_i^j étant les m sous-ensembles induits par le test de l'exemple x_i et la distribution π étant supposée uniforme tel que $\pi(h) = \frac{1}{|s|} \quad \forall h \in \mathbb{H}$:

$$\mathcal{E}nt(\mathbb{S}) = \sum_{\mathbb{S}} -log\pi(h).\pi(h)$$

$$\mathcal{G}ain(\mathbb{S}, x_i) = \mathcal{E}nt(\mathbb{S}) - \sum_{j=1}^{n} \frac{|S_i^j|}{|\mathbb{S}|} \mathcal{E}nt(S_i^j)$$

[5]Pour une classification binaire, les deux sous-ensembles en question sont S_i^+ et S_i^-.

La qualité Q d'une stratégie active basée sur la réduction de l'espace des versions peut être définie comme étant la hauteur moyenne de l'arbre $\mathcal{T}$ qu'elle produit. Cela s'écrit de la manière suivante, avec $\mathcal{H}eight(h)$ la hauteur de la feuille correspondant à l'hypothèse h :

$$Q(\mathcal{T}, \pi) = \sum_{h \in \mathbb{H}} \pi(h).\mathcal{H}eight(h)$$

De manière générale, la qualité d'une stratégie dépend de la typologie des données d'apprentissage. Il existe des problèmes triviaux qui nécessitent toutes les étiquettes manquantes pour trouver la bonne hypothèse. Sanjoy Dasgupta [Dasgupta, 2005] établit des bornes théoriques sur le nombre d'étiquettes demandées à l'expert.

Positionnement :
La construction d'un arbre de décision sur l'espace des versions est difficile à mettre en œuvre dans la pratique car on dispose rarement de l'espace des versions d'un modèle prédictif et de la distribution de probabilité des hypothèses. Cette approche suppose que l'hypothèse apprise par le modèle soit toujours consistante avec les exemples étiquetés, ce qui est restrictif en pratique. Néanmoins, la construction d'un arbre de décision sur l'espace des versions constitue une vue théorique intéressante de l'échantillonnage sélectif. La dichothomie probabiliste, utilisée au Chapitre 5 dans le cadre d'expériences comparatives, peut être assimilée à un arbre de décision sur l'espace des versions.

2.2.4 Échantillonnage par réduction de l'erreur de généralisation

L'approche proposée par [Cohn *et al.*, 1995] sélectionne les exemples qui minimiseraient l'erreur de généralisation du modèle prédictif, s'ils étaient étiquetés. L'erreur commise par le modèle $\mathcal{M}$ à l'itération t est notée $E_t(\mathcal{M})$. Cette erreur est calculée grâce à une fonction de coût, notée $\mathcal{L}oss(\mathcal{M}, x)$, qui évalue l'erreur du modèle pour une entrée particulière $x \in \mathbb{X}$. Le calcul exact de l'erreur de généralisation intègre la fonction de coût sur l'espace $\mathbb{X}$, avec $P(x)$ la distribution de probabilités des exemples $x \in \mathbb{X}$:

$$E_t(\mathcal{M}) = \int_{\mathbb{X}} \mathcal{L}oss(\mathcal{M}, x) P(x) dx$$

L'idée principale de cette stratégie est d'estimer l'erreur de généralisation du modèle $\mathcal{M}$ à l'itération $t+1$; un exemple non-étiqueté est ajouté à T et des suppositions sont faites sur la valeur de l'étiquette manquante. Soit $\mathcal{M}_{(x^\diamond, y^\diamond)}$ le modèle prédictif $\mathcal{M}$ dont l'apprentissage prend en compte un exemple fictif supplémentaire, noté $(x^\diamond, y^\diamond)$. L'étiquette $y^\diamond$ n'est pas connue. L'erreur de généralisation à l'itération $t+1$ est estimée par une intégration sur toutes les étiquettes de $\mathbb{Y}$. Les différentes valeurs possibles pour l'étiquette de $x^\diamond$ sont pondérées par la probabilité d'observer chacune des classes connaissant cet exemple :

$$E_{t+1}(\mathcal{M}_{x^\diamond}) = \int_{\mathbb{Y}} P(y|x^\diamond)dy \int_{\mathbb{X}} P(x)\mathcal{L}oss(\mathcal{M}_{(x^\diamond,y)}, x)dx$$

L'échantillonnage par réduction de l'erreur de généralisation sélectionne l'exemple non-étiqueté $q \in U$ défini par $ArgMin_{x^\diamond \in \mathbb{X}} E_{t+1}(\mathcal{M}_{x^\diamond})$. Une fois étiqueté, l'exemple q est ajouté à l'ensemble d'apprentissage T.

Sans disposer de tous les éléments de $\mathbb{X}$, Nicholas Roy [Roy and McCallum, 2001] montre comment cette stratégie peut être mise en œuvre en utilisant uniquement les données d'apprentissage. L'erreur de généralisation est estimée en considérant seulement les exemples étiquetés disponibles à l'instant t. Un a priori uniforme sur la distribution $P(x)$ est adopté :

$$\widehat{E}_t(\mathcal{M}) = \frac{1}{|L|}\sum_{i=1}^{|L|} \mathcal{L}oss(\mathcal{M}, x_i) \qquad avec \quad x_i \in L$$

Comme précédemment, l'estimation de l'erreur de généralisation à l'itération $t+1$ requiert un exemple supplémentaire associé à une des étiquettes possibles. Chaque exemple $x^\diamond \in U$ et chaque étiquette $y^\diamond \in \mathbb{Y}$ peuvent s'associer pour former cet exemple supplémentaire. Pour chaque exemple non-étiqueté, le modèle est entraîné plusieurs fois en fixant la valeur de l'étiquette. L'erreur de généralisation $\widehat{E}_{t+1}(\mathcal{M}_{(x^\diamond,y^\diamond)})$ est estimée pour chaque valeur de l'étiquette. Lorsque toutes les étiquettes ont été considérées pour un exemple $x^\diamond \in \mathbb{X}$, l'erreur de généralisation espérée $\widehat{E}_{t+1}(\mathcal{M}_{x^\diamond})$ est estimée. Pour ce faire, le modèle prédictif estime la distribution $P(y|x)$, des étiquettes $y \in \mathbb{Y}$ connaissant les exemples $x \in \mathbb{X}$. Cette stratégie sélectionne les exemples non-étiquetés pour lesquels l'erreur de généralisation espérée est minimale. Une vue synthétique de cette stratégie est présentée par l'Algorithme 3. Il existe autant de variantes de cette stratégie que de fonctions de coût. Considérons à titre illustratif un cas d'utilisation de l'échantillonnage par réduction de l'erreur de généralisation.

X. Zhu [Zhu *et al.*, 2003] propose d'estimer l'erreur de généralisation par le risque empirique et d'utiliser une fenêtre de Parzen à noyau gaussien [Parzen, 1962] comme modèle prédictif. Ici, le risque $R_t(\mathcal{M})$ est défini comme étant la somme des probabilités que le modèle prenne de mauvaises décisions sur l'ensemble d'apprentissage. Soit $p(y|l_i)$ la probabilité d'observer la classe $y \in \mathbb{Y}$ connaissant l'exemple $l_i \in L$. Le risque empirique s'écrit selon l'Équation 2.1, avec $\mathbb{1}$ la fonction indicatrice égale à 1 si $f(l_i) \neq y$ et égale à 0 sinon.

$$R_t(\mathcal{M}) = \sum_{i=1}^{|L|} \sum_{y^\diamond=0,1} \mathbb{1}_{\{f(l_i)\neq y^\diamond\}} P(y^\diamond|l_i)P(l_i) \qquad avec\, l_i \in L \tag{2.1}$$

Le modèle prédictif utilisé est un estimateur de densité et estime $\widehat{P}(y^\diamond|l_i)$ la probabilité d'observer la classe $y^\diamond$ conditionnellement à l'exemple l_i. Le risque empirique est estimé en adoptant un a priori uniforme sur la distribution $P(x)$ (voir Équation 2.2).

$$\hat{R}_t(\mathcal{M}) = \frac{1}{|L|} \sum_{i=1}^{|L|} \sum_{y^\diamond=0,1} \mathbb{1}_{\{f(l_n) \neq y^\diamond\}} \hat{P}(y^\diamond | l_i) \tag{2.2}$$

Cette stratégie sélectionne l'exemple non-étiqueté $q \in U$ qui minimise le risque à l'itération t+1. $R_{t+1}(\mathcal{M}_{x^\diamond})$ est le risque espéré après l'étiquetage de l'exemple $x^\diamond \in U$. Dans le cas d'une classification binaire, l'étiquette $f(x^\diamond)$ est supposée égale à 1 [*respectivement* égale à 0] pour estimer $\hat{R}_{t+1}(\mathcal{M}_{(x^\diamond, y^\diamond=1)})$ [*respectivement* $\hat{R}_{t+1}(\mathcal{M}_{(x^\diamond, y^\diamond=0)})$]. L'Équation 2.3 montre comment agréger les estimations de risque grâce aux probabilités d'observer chacune des classes connaissant l'exemple $x^\diamond$.

$$\hat{R}_{t+1}(\mathcal{M}_{x^\diamond}) = \hat{P}(y^\diamond = 1 | x^\diamond) \hat{R}(\mathcal{M}_{(x^\diamond, y^\diamond=1)}) + \hat{P}(y^\diamond = 0 | x^\diamond) \hat{R}(\mathcal{M}_{(x^\diamond, y^\diamond=0)}) \qquad avec\, x^\diamond \in U \tag{2.3}$$

Pour exprimer la stratégie de réduction du risque empirique par un algorithme, il suffit de remplacer l'étape (B) de l'Algorithme 1 par : *"Rechercher l'instance* $q = ArgMin_{x^\diamond \in U} \hat{R}_{t+1}(\mathcal{M}_{x^\diamond})$*"*.

Notations :

- $\mathcal{M}$ un modèle prédictif muni d'un algorithme d'apprentissage $\mathcal{L}$
- Les ensembles U et L d'exemples non étiquetés et étiquetés
- n le nombre d'exemples étiquetés souhaité.
- L'ensemble d'apprentissage T avec $|T| < n$
- $\mathbb{Y}$ l'ensemble des étiquette qui peuvent être attribuées aux exemples de U
- $\widehat{E}_{t+1} : U \times \mathbb{M} \rightarrow \Re$ une estimation de l'erreur de généralisation pour le modèle $\mathcal{M}$, à l'itération t, entraîné avec un exemple supplémentaire tel que $T \leftarrow T \cup (x, f(x))$

Répéter
- (A) Entraîner le modèle $\mathcal{M}$ grâce à $\mathcal{L}$ et T
- **Pour** chaque instance $x^\diamond \in U$ **faire**
  - **Pour** chaque label $y^\diamond \in \mathbb{Y}$ **faire**
    - i) Entraîner le modèle $\mathcal{M}$ grâce à $\mathcal{L}$ et $(T \cup (x^\diamond, y^\diamond))$
    - ii) Calculer l'erreur de généralisation $\widehat{E}_{t+1}(\mathcal{M}_{(x^\diamond, y^\diamond)})$
  - **Fin Pour**
  - Calculer l'erreur de généralisation espérée $\widehat{E}_{t+1}(\mathcal{M}_{x^\diamond}) = \sum_{y^\diamond \in \mathbb{Y}} \widehat{E}_{t+1}(\mathcal{M}_{(x^\diamond, y^\diamond)}).\widehat{P}(y^\diamond | x^\diamond)$
- **Fin Pour**
- (B) Rechercher l'instance $q = ArgMin_{x^\diamond \in U} \widehat{E}_{t+1}(\mathcal{M}_{x^\diamond})$
- (C) Retirer q de U et demander l'étiquette $f(q)$ à l'expert.
- (D) Ajouter q à L et ajouter $(q, f(q))$ à T

Tant que $|T| < n$

Algorithme 3: Échantillonnage par réduction de l'erreur de généralisation

Positionnement :
Les approches d'apprentissage actif par réduction de l'erreur de généralisation ont un caractère exhaustif. Ces stratégies examinent tous les exemples candidats et toutes les valeurs d'étiquette possibles. Les variantes de l'échantillonnage par réduction de l'erreur de généralisation se différencient entres elles par la fonction de coût utilisée. Il est important de noter que ces stratégies sont fortement combinatoires. À chaque itération de l'apprentissage actif, la sélection des exemples à étiqueter engendre $|\mathbb{Y}| \times |U|$ entraînements du modèle. L'échantillonnage par réduction de l'erreur de généralisation est utilisé dans la suite de la thèse, dans le cadre d'expériences comparatives.

2.3 Différentes vues sur l'apprentissage actif

Cette section développe certains aspects de l'apprentissage actif et met en exergue les questions importantes sur la mise en œuvre des stratégies actives. Notre réflexion sur les enjeux de l'apprentissage actif nous amène à faire des choix et à fixer les objectifs que devrait atteindre une stratégie idéale.

2.3.1 Dilemme exploitation / exploration

Lors d'un échantillonnage sélectif, le choix des exemples à étiqueter peut être vu comme le dilemme entre l'exploration et l'exploitation des données d'apprentissage. La sélection d'un exemple non-étiqueté dans une zone non-échantillonnée de $\mathbb{X}$ contribue à **explorer** les données. L'espace $\mathbb{X}$ tend à être echantillonné uniformément, ce qui limite les zones pour lesquelles l'hypothèse apprise par le modèle est potentiellement incorrecte. Plus la dimension $\mathbb{X}$ est élevée, plus l'exploration de cet espace requiert d'étiquetage. La sélection d'un exemple non-étiqueté dans une zone échantillonnée de $\mathbb{X}$ contribue à **exploiter** les données. Dans ce cas, la stratégie active se concentre sur une zone déjà peuplée d'exemples étiquetés et affine localement le modèle prédictif.

Le dilemme entre l'exploitation et l'exploration des données peut être illustré grâce à deux situations extrêmes. D'une part, une stratégie d'apprentissage actif qui ne fait qu'exploiter les données court le risque d'occulter une grande partie de l'espace $\mathbb{X}$ lors de la sélection des exemples à étiqueter. Le modèle prédictif sera spécialisé dans certaines régions de $\mathbb{X}$ mais il sera vraisemblablement de très mauvaise qualité en généralisation. D'autre part, une stratégie qui ne fait qu'explorer les données ne se focalise pas sur les régions de $\mathbb{X}$ où l'étiquetage de nouveaux exemples améliorerait le modèle prédictif. Dans ces conditions, une stratégie d'apprentissage actif présente peu d'intérêt par rapport à un échantillonnage aléatoire (Section 3.2.2). Ces deux situations extrêmes illustrent la nécessité pour une stratégie d'apprentissage actif de trouver un **compromis** entre l'exploitation et l'exploration des données.

Recours à plusieurs stratégies :
Plusieurs stratégies d'apprentissage actif peuvent être utilisées conjointement pour répondre au compromis exploitation / exploitation. Considérons deux stratégies actives respectivement dédiées à l'exploitation et à l'exploration des données. A chaque itération

de l'échantillonnage sélectif, une des deux stratégies est utilisée (voir étape B de l'Algorithme 1). Le choix de la stratégie est probabiliste, p représente la probabilité d'explorer les données et $1-p$ la probabilité de les exploiter ; ces probabilités devant être mises à jour au cours de l'échantillonnage sélectif pour ajuster le compromis exploitation / exploration.

T. Osugi [Osugi *et al.*, 2005a] fait un parallèle avec l'apprentissage par renforcement. Le modèle prédictif $\mathcal{M}$ est considéré comme étant un *"agent"* qui peut faire deux actions : explorer ou exploiter les données. A chaque itération, l'agent décide d'exécuter une action et reçoit une récompense [*respectivement* une punition] si cette action est appropriée [*respectivement* inappropriée]. Une métrique est utilisée pour mesurer $|\widehat{f}_t - \widehat{f}_{t+1}|$, la variation de l'hypothèse apprise par le modèle entre deux itérations. Une action est d'autant plus récompensée que l'hypothèse apprise par le modèle varie. La probabilité p est mise à jour après chaque action.

Il est également possible d'utiliser des heuristiques issues de l'optimisation combinatoire pour gérer le compromis exploration / exploitation. T. Zoller [Zoller and Buhmann, 2000] utilise un algorithme de recuit simulé pour faire varier p au cours du temps. Cette heuristique inspirée de la thermodynamique fait décroître la probabilité d'explorer les données selon une fonction prédéfinie, aussi appelée schéma de refroidissement. Cette heuristique concentre l'exploration au début de l'apprentissage actif et exploite par la suite des données représentatives de $\mathbb{X}$.

Certaines approches intègrent plusieurs stratégies d'apprentissage actif. Contrairement à l'échantillonnage par comité de modèles (Section 2.2.2) qui met en jeu plusieurs modèles et une seule stratégie active, ces approches n'utilisent qu'un seul modèle prédictif et plusieurs stratégies actives. L'exemple qui est le mieux "noté" par l'ensemble des stratégies est sélectionné pour être étiqueté. Dans ce cas, le compromis exploration / exploitation est assuré en considérant à la fois des stratégies qui explorent et des stratégies qui exploitent les données. Y. Baram [Baram *et al.*, 2004] combine trois algorithmes d'apprentissage actif : l'échantillonnage par incertitude (Section 2.2.1), la réduction de l'erreur de généralisation (Section 2.2.4) et une heuristique qui lui est propre basée sur les fonctions à noyaux, appelée *"Kernel Farthest First"*.

Pré-clustering :
H. Nguyen améliore la stratégie d'échantillonnage par incertitude (Section 2.2.1) grâce à une méthode de clustering appliquée aux données, à chaque itération, avant la sélection des exemples à étiqueter [Nguyen and Smeulders, 2004]. L'échantillonnage par incertitude sélectionne des exemples proches de la frontière de décision du modèle prédictif, cette stratégie exploite l'hypothèse apprise par le modèle sans jamais explorer l'espace des données. L'utilisation d'un pré-clustering permet de diversifier les exemples sélectionnés.

Un clustering est effectué au préalable sur les données d'apprentissage, $\Phi = L \cup U$. Lors de cette étape, les étiquettes des exemples de l'ensemble L sont ignorées. Seuls les centroïdes des clusters sont candidats à l'étiquetage. Cette stratégie basée sur le pré-clustering fait deux hypothèses : les exemples d'un même cluster sont supposés appartenir à la même classe ; les centroïdes sont considérés comme étant représentatifs des clusters. Cette stratégie cherche à étiqueter des exemples diversifiés, appartenant à des clusters

différents. H. Nguyen propose un critère qui sélectionne le centroïde contribuant le plus à l'erreur courante.

L'étape de clustering est répétée à chaque itération de l'échantillonnage sélectif. L'utilisateur paramétrise la taille des clusters au cours des itérations successives. La diminution de la taille des clusters au cours de l'échantillonnage sélectif répond au compromis exploration / exploitation. Lorsque les clusters contiennent beaucoup d'exemples, les centroïdes sont éloignés les uns des autres. Dans ce cas, la sélection des exemples à étiqueter favorisera l'exploration des données. Au contraire, lorsque les clusters contiennent peu d'exemples, les centroïdes sont potentiellement proches les uns des autres. Dans ce cas, la stratégie basée sur le pré-clustering exploite les données. A l'instar du *recuit simulé* [Zoller and Buhmann, 2000], cette stratégie explore davantage les données au début de l'échantillonnage sélectif qu'à la fin.

Mesure de dissimilarité :
Certaines stratégies d'apprentissage actif abordent le compromis exploration / exploitation en mesurant la dissimilarité entre les exemples sélectionnés. Il existe plusieurs mesures de dissimilarité.

Xu [Xu *et al.*, 2007] propose une stratégie active multicritère appliquée à la recherche documentaire. L'objectif est d'extraire des textes d'un corpus répondant au mieux à la requête d'un utilisateur. Parmi les critères utilisés pour choisir les textes pertinents, une mesure de dissimilarité est employée. Cette approche exploite une métrique, et maximise la distance entre le nouvel exemple et l'exemple étiqueté le plus proche. Cette mesure de dissimilarité étiquette des exemples les plus éloignés les uns des autres, au sens de la métrique utilisée.

Brinker [Brinker, 2003] présente une mesure de dissimilarité utilisée dans le cadre d'un échantillonnage sélectif, le modèle prédictif utilisé étant une Machine à Vecteurs Support (SVM). Le but de cette stratégie est d'étiqueter un ensemble d'exemples à chaque itération, de manière à réduire le plus possible l'espace des versions $\mathbb{S}$ du modèle prédictif. Chaque exemple candidat à l'étiquetage correspond à un hyperplan dans l'espace préhilbertien induit par le noyau. Cette stratégie sélectionne des exemples non-étiquetés dont les hyperplans correspondants sont orientés selon des directions les plus éloignées les unes des autres. Une mesure d'angle est définie en exploitant l'astuce du noyau.

Positionnement :
Le compromis entre l'exploitation et l'exploration des données est une question centrale à l'apprentissage actif. L'objectif commun à toutes les stratégies actives est de trouver un arbitrage entre deux comportements essentiels : i) se focaliser sur les zones de l'espace $\mathbb{X}$ où l'étiquetage de nouveaux exemples fait progresser le modèle prédictif ; ii) explorer l'espace $\mathbb{X}$ pour connaitre les zones importantes. Une bonne stratégie d'apprentissage actif doit être capable de gérer ce compromis.

2.3.2 Apprentissage supervisé *vs* semi-supervisé

À chaque itération d'un échantillonnage sélectif, un algorithme d'apprentissage est utilisé pour entraîner le modèle prédictif grâce aux données. Cet algorithme d'apprentissage est soit supervisé, c'est-à-dire qu'il exploite uniquement les exemples étiquetés, soit semi-supervisé et exploite tous les exemples disponibles. Pour plus de détails, la Section 4.1.1 présente un bref état-de-l'art des méthodes d'apprentissage semi-supervisé. La plupart des stratégies d'apprentissage actif présentées à la Section 2.2 peuvent être déclinées selon une variante semi-supervisée.

Par exemple, O. Chapelle [Chappelle, 2005] adapte la stratégie de l'échantillonnage par réduction de l'erreur de généralisation (Section 2.2.4) au cas de l'apprentissage semi-supervisé ; le modèle prédictif utilisé étant une fenêtre de Parzen [Parzen, 1962]. Cette stratégie intègre les exemples non-étiquetés aux prédictions de la fenêtre de Parzen. Les exemples non-étiquetés sont supposés suivre la même distribution conditionnelle aux classes que les exemples étiquetés. O. Chapelle montre expérimentalement que la prise en compte des exemples non-étiquetés améliore les performances de l'échantillonnage par réduction de l'erreur de généralisation.

I. Muslea [Muslea *et al.*, 2002] propose une stratégie d'échantillonnage par comité de modèles fondée sur une version probabiliste du Co-training, nommée le "Co-EMT". Le Co-training est un algorithme d'apprentissage semi-supervisé qui partitionne l'ensemble des variables caractérisant les exemples en deux sous-ensembles disjoints. Cette stratégie implique deux modèles prédictifs dont l'apprentissage est réalisé sur l'un ou l'autre de ces sous-ensembles de variables. L'idée principale du "Co-EMT" est de mesurer le désaccord entre les deux modèles prédictifs, lors de la prédiction des étiquettes des éléments de U. Cette stratégie d'apprentissage actif fait l'hypothèse que l'ensemble des variables est séparable en deux sous-ensembles sur lesquels les modèles prédictifs sont capables d'apprendre le concept cible.

D'une manière générale, l'utilisation d'algorithmes d'apprentissage semi-supervisés pour l'échantillonnage sélectif pose le problème de la représentativité des exemples étiquetés. La stratégie active induit un biais sur la distribution de l'ensemble L. Du fait que les exemples étiquetés sont sélectionnés selon une stratégie active, leur distribution n'est pas $i.i.d$ vis-à-vis de la distribution de l'ensemble Φ. Les ensembles U et L ont donc des distributions différentes. Ce biais peut être problématique pour l'apprentissage semi-supervisé, lorsque les exemples non-étiquetés sont supposés suivre la même distribution que l'ensemble L. Une piste pour remédier à ce problème serait d'utiliser la pondération covariative [Sugiyama *et al.*, 2007]. Cette méthode d'apprentissage pourrait corriger le biais existant entre les distributions des ensembles L et U en pondérant les exemples étiquetés. Pour plus de détails sur la pondération covariative, se reporter à la Section 4.1.1. Dans le Chapitre 3, nous proposons une méthode d'apprentissage actif basée sur le partitionnement dichotomique récursif de l'espace $\mathbb{X}$. Cette stratégie met en compétition des modèles locaux et définit des zones d'intérêt dans l'espace $\mathbb{X}$. Les exemples étiquetés sont choisis aléatoirement (de manière $i.i.d$) dans les zones sélectionnées, notre stratégie active n'est donc pas sensible au biais existant entre L et U.

Positionnement :
Les récents travaux qui utilisent des algorithmes d'apprentissage semi-supervisé pour l'échantillonnage sélectif montrent que la prise en compte des exemples non-étiquetés améliore les performances des stratégies actives. Il faut cependant avoir conscience des hypothèses que les algorithmes semi-supervisés induisent sur les données d'apprentissage. De notre point de vue, une bonne stratégie d'apprentissage actif exploite les exemples non-étiquetés sans faire d'hypothèse forte sur la distribution conditionnelle aux classes. Notre objectif est d'élaborer une stratégie active applicable dans le cas général, où aucune information sur la distribution conditionnelle aux classes n'est disponible.

2.3.3 Hypothèses sur les données

Dans le cadre général de l'échantillonnage sélectif aucune information sur la distribution des classes n'est disponible. Pourtant, les principales approches de la littérature adoptent des hypothèses fortes sur les données. Ces hypothèses sont soit induites par la stratégie active qui sélectionne les exemples à étiqueter, soit induites par le modèle prédictif.

Hypothèses induites par les stratégies actives :
L'échantillonnage par incertitude (Section 2.2.1) sélectionne les exemples les plus proches de la frontière de décision. Les données ne sont jamais explorées dans le reste de l'espace de recherche $\mathbb{X}$. Dans le cas où les données sont bruitées ou non-séparables par le modèle prédictif, l'échantillonnage par incertitude peut adopter des comportements pathologiques. Par exemple, cette stratégie peut concentrer l'étiquetage de nouveaux exemples dans une zone fortement bruitée de $\mathbb{X}$. Finalement, l'échantillonnage par incertitude fait l'hypothèse que les données sont séparables par le modèle prédictif et non-bruitées.

L'échantillonnage par comité de modèles (Section 2.2.2) suppose que les hypothèses $h_1...h_m \in \mathbb{H}$ apprises par les modèles prédictifs $\mathcal{M}_1...\mathcal{M}_m$ forment un échantillon représentatif de l'espace des versions $\mathbb{S} \subseteq \mathbb{H}$. Cette stratégie présume que les hypothèses apprises par les modèles sont consistantes avec les exemples étiquetés, $h_1...h_m \in \mathbb{S}$. Cette condition n'est pas nécessairement vérifiée en pratique, notamment dans le cas où les données sont bruitées ou non-séparables par les modèles prédictifs utilisés.

L'échantillonnage par réduction de l'erreur de généralisation (Section 2.2.4) estime $E(\mathcal{M})$ l'erreur de généralisation du modèle prédictif $\mathcal{M}$, en exploitant les exemples étiquetés de l'ensemble L. Cette stratégie d'apprentissage actif pose le problème de la qualité de l'estimation de $E(\mathcal{M})$, notamment au début de l'échantillonnage sélectif, lorsque les exemples étiquetés sont peu nombreux. L'estimation de l'erreur de généralisation est biaisée car l'ensemble L est par définition non-*i.i.d*, car les exemples étiquetés sont choisis selon un critère de sélection.

Hypothèses induites par le modèle prédictif :
Le modèle prédictif employé dans le cadre d'un échantillonnage sélectif peut également induire des hypothèses sur les données. Par exemple, l'utilisation d'un Séparateur à Vaste Marge pure (SVM) [Tong and Koller, 2000] suppose que les données sont linéairement

séparables, une fois projetées dans l'espace préhilbertien propre à la fonction "noyau" utilisée.

La régression logistique est un modèle prédictif fréquemment employé pour l'échantillonnage sélectif [Castro and Nowak, 2005]. Ce modèle fait l'hypothèse que les données sont organisées autour d'un hyperplan séparateur. Dans le cas d'un problème de classification binaire, cet hyperplan coupe $\mathbb{X}$ en deux sous-espaces qui sont peuplés d'exemples majoritairement étiquetés par l'une des deux classes. Les distributions conditionnelles aux classes sont supposées varier continûment le long de l'axe orthogonal à l'hyperplan séparateur. La régression logistique peut être mise à mal lorsque les données d'apprentissage ne respectent pas ces hypothèses.

Enfin, la fenêtre de Parzen à noyau gaussien est un autre exemple de modèle prédictif utilisé pour l'échantillonnage sélectif [Chappelle, 2005]. Ce modèle prédictif fait l'hypothèse que la géométrie du motif à découvrir est homogène dans l'espace $\mathbb{X}$, la variance du noyau gaussien étant identique pour toutes les régions de l'espace. La fenêtre de Parzen suppose également que les distributions conditionnelles aux classes sont des sommes de gaussiennes centrées sur les exemples étiquetés. Cela peut être problématique lorsque les données sont constituées de zones pures[6] dont les frontières se situent dans des régions de forte densité.

Positionnement :
Une procédure d'échantillonnage sélectif est difficilement applicable à des données quelconques, au vue des hypothèses induites par les stratégies actives ou par les modèles prédictifs utilisés. Il existe deux voies possibles pour contourner cette difficulté. Comme exposé en introduction de ce chapitre, une stratégie active pourrait chercher parmi une collection d'hypothèses a priori celles qui sont les plus adaptées aux données observées. Dans ce cas, le but serait de déterminer le plus tôt possible la nature des données d'apprentissage. Dans le cadre de cette thèse, nous adoptons un point de vue alternatif. Notre objectif est d'élaborer une stratégie d'apprentissage actif qui adopte les hypothèses a priori les plus faiblement informatives. Une telle stratégie doit être performante quelque soit la nature des données d'apprentissage. Dans la suite de ce manuscrit, l'élaboration d'une stratégie Bayésienne "objective" [Berger, 2006] est présentée au Chapitre 5.

2.3.4 Critère d'arrêt

La plupart des stratégies actives utilisées dans le cadre d'un échantillonnage sélectif exploitent un critère d'arrêt trivial. Un budget d'étiquetage permettant l'achat de n étiquettes est préalablement fixé. Pour rappel, le coût d'étiquetage est considéré comme étant constant quelque soit l'exemple. Les stratégies actives ont pour but de sélectionner itérativement les n exemples les plus utiles à l'apprentissage du modèle prédictif. L'échantillonnage sélectif s'arrête lorsque le budget d'étiquetage est épuisé.

Le critère d'arrêt pourrait tenir compte de la qualité ou de l'amélioration du modèle prédictif ; cette piste est très peu explorée dans la littérature. Dans ce cas, l'Algorithme 1

[6] On entend par zone pure, une région de l'espace $\mathbb{X}$ où tous les exemples étiquetés sont de la même classe.

s'arrêterait lorsque les nouveaux exemples étiquetés n'améliorent plus significativement le modèle prédictif. L'élaboration d'un tel critère d'arrêt pose plusieurs difficultés :

Manque d'exemples étiquetés pour l'évaluation du modèle courant :
Y. Baram [Baram *et al.*, 2004] souligne le fait que le modèle prédictif ne peut pas être correctement évalué lors d'un échantillonnage sélectif. Une évaluation fiable requiert un grand nombre d'exemples étiquetés ; c'est pourquoi les méthodes standards comme la validation croisée ne permettent pas d'évaluer précisément le modèle prédictif. R. Kothari [Kothari and Jain, 2003] suggère une heuristique exploitable dans le cadre de l'échantillonnage sélectif, pour l'évaluation d'un modèle prédictif. Les principales étapes cet algorithme sont les suivantes :

- apprentissage du modèle grâce à l'ensemble T ;
- prédiction des étiquettes des éléments de U par le modèle prédictif courant $\mathcal{M}$;
- élaboration d'un nouvel ensemble d'apprentissage T^* grâce aux étiquettes prédites pour les exemples non-étiquetés ;
- élaboration d'un nouvel ensemble de test $Test^*$ grâce aux exemples étiquetés de L ;
- apprentissage d'un nouveau modèle prédictif $\mathcal{M}^*$ grâce à l'ensemble T^* ;
- évaluation de $\mathcal{M}^*$ sur l'ensemble $Test^*$.

De notre point de vue, cette heuristique n'est pas fiable car l'ensemble de test $Test^*$ comporte un faible nombre d'exemples étiquetés. De plus, les données utilisées pour l'apprentissage du modèle $\mathcal{M}^*$ contiennent potentiellement des erreurs, les étiquettes étant issues des prédictions du modèle initial $\mathcal{M}$. La faible quantité d'exemples étiquetés constitue le principal obstacle à l'élaboration d'un critère d'arrêt fondé sur la qualité du modèle prédictif.

Biais sur la distribution de l'ensemble L :
L'évaluation consiste à calculer l'erreur moyenne que commet le modèle lors de la prédiction des étiquettes des éléments de L, étant donnée une mesure d'erreur. Cette évaluation doit prendre en compte le biais induit par la stratégie active sur la distribution des exemples étiquetés. La distribution de l'ensemble L est biaisée au profit des exemples jugés utiles par la stratégie active ; cette distribution n'est donc pas représentative des données.

Une piste pour résoudre ce problème serait d'utiliser la pondération covariative [Sugiyama *et al.*, 2007] (Pour plus de détails, voir Section 4.1.1). Cette approche pourrait être exploitée pour corriger le biais existant entre les ensembles L et Φ. L'idée principale est de pondérer les exemples étiquetés lors de l'estimation de l'erreur moyenne du modèle. Tout d'abord, un estimateur de densité est utilisé pour évaluer P_U [*respectivement* P_L], la distribution des exemples non-étiquetés [*respectivement* étiquetés]. La pondération d'un exemple $(x, y) \in T$ est donnée par le ratio $\frac{P_U(x)}{P_L(x)}$. Les exemples étiquetés qui sont "sous-représentatifs" de l'ensemble U sont davantage pris en compte lors de l'évaluation du modèle. M. Sugiyama montre que grâce à cette pondération l'évaluation du modèle est non biaisée, sous réserve que les distributions P_U et P_L soient estimées correctement. La

faible quantité d'exemples étiquetés rend la technique de la pondération covariative difficile à mettre en œuvre dans le cas d'un échantillonnage sélectif.

Positionnement :
L'évaluation du modèle prédictif courant est un problème difficile. La qualité du modèle prédictif serait pourtant un très bon indicateur pour arrêter un échantillonnage sélectif. Cette connaissance permettrait d'éviter l'étiquetage d'exemples inutiles, qui ne font plus progresser le modèle prédictif de manière significative. L'utilisation d'un modèle robuste est un premier pas pour résoudre le problème de l'évaluation. Ce modèle doit être capable d'apprendre avec peu de données (non-i.i.d) et d'estimer la confiance que l'on peut avoir en ces prédictions.

Dans la suite de ce manuscrit, des expériences sont menées pour évaluer et comparer des stratégies actives sur différents jeux de données. Dans le cadre de ces expériences, le critère d'arrêt prend la forme d'un budget d'étiquetage. L'échantillonnage sélectif s'arrête lorsque le nombre d'exemples étiquetés atteint la limite du budget préalablement fixé.

2.3.5 Notre protocole d'évaluation

Les travaux expérimentaux réalisés dans cette thèse nécessitent d'évaluer et de comparer des stratégies d'apprentissage actif. Cette section définit notre protocole d'évaluation ainsi que les mesures de performance employées.

Qu'est-ce que la qualité d'une stratégie d'apprentissage actif ?
Dans le cadre d'un échantillonnage sélectif, une stratégie active a pour but d'étiqueter les exemples les plus utiles à l'apprentissage du modèle prédictif. La qualité de l'échantillon d'apprentissage L est relative au modèle utilisé et ne peut pas être mesurée directement. Lors de l'évaluation d'une stratégie active, la performance du modèle prédictif est supposée refléter la qualité des exemples étiquetés. Pour comparer plusieurs stratégies d'apprentissage actif, certains paramètres doivent être identiques : le nombre d'exemples étiquetés, le modèle prédictif, l'algorithme d'apprentissage et l'état initial des données d'apprentissage.

Quelles données d'évaluation ?
Lors de la mise en œuvre d'un échantillonnage sélectif sur un problème réel, aucun ensemble de test n'est disponible. Seuls les exemples jugés utiles par la stratégie active sont étiquetés par un expert moyennant un coût. En pratique, l'absence d'une grande quantité d'exemples étiquetés rend l'évaluation du modèle prédictif impossible à réaliser.

Les stratégies d'apprentissage actif peuvent cependant être utilisées pour certains types d'applications. En faisant un parallèle avec l'apprentissage par renforcement [Harmon, 1996], une stratégie active peut être vue comme un agent dont les actions consistent à sélectionner les exemples à étiqueter. L'évaluation du modèle prédictif peut être remplacée par un signal de renforcement généré par l'application (i.e : un bénéfice, une productivité, un coût de fabrication ... etc.). Dans ce cas, le choix des exemples est influencé par le signal de renforcement.

Dans le cadre des travaux expérimentaux réalisés dans cette thèse, l'évaluation des stratégies d'apprentissage actif est essentielle. Pour ce faire, des jeux de données initialement prévus pour la classification supervisée sont utilisés. Ces jeux de données sont divisés en deux ensembles indépendants (apprentissage et test) dont les exemples sont tous étiquetés. Les étiquettes des exemples d'apprentissage sont masquées au début de l'échantillonnage sélectif. L'étiquetage des exemples par l'expert est simulé en dévoilant les étiquettes correspondantes. Lors de ces expériences, l'ensemble de test est utilisé pour évaluer le modèle prédictif, l'apprentissage du modèle étant réalisé grâce aux exemples étiquetés et non-étiqueté.

Comment mesurer la qualité d'un échantillonnage sélectif ?
L'évaluation des stratégies actives requiert un critère mesurant la performance du modèle prédictif. Parmi les mesures d'évaluation dédiées à la classification supervisée, nous choisissons d'utiliser l'aire sous la courbe de ROC aussi appelée AUC (Area Under roc Curve) [Fawcett, 2003]. Des travaux récents montrent que ce critère d'évaluation est l'un des plus performants [Huang and Ling, 2005]. L'AUC est présentée dans l'Annexe A.a.

Dans le cadre d'un échantillonnage sélectif, la performance d'un modèle prédictif n'est pas forcément reproductible pour $|L|$ fixé. La stratégie active et l'algorithme d'apprentissage utilisés ne sont pas nécessairement déterministes. Le protocole d'évaluation utilisé dans cette thèse consiste à répéter plusieurs fois les expériences pour évaluer la performance du modèle au cours d'un échantillonnage sélectif. Pour chaque valeur de $|L|$, la performance moyenne et la variance des résultats sont évaluées.

Ce protocole d'évaluation ne permet pas de comparer deux stratégies actives sur l'intégralité d'un jeu de données. En effet, les courbes obtenues tracent la performance moyenne de chaque stratégie en fonction du nombre d'exemples étiquetés. L'intégration de la performance du modèle sur l'ensemble des valeurs de $|L|$ permet d'obtenir une valeur scalaire représentative de la performance d'une stratégie sur l'intégralité d'un jeu de données. Dans le cadre de cette thèse, nous utilisons la mesure de déficit [Baram *et al.*, 2004] qui compare une stratégie active à la stratégie aléatoire. Cette mesure de déficit est présentée dans l'annexe A.c.

2.4 Conclusion : Nos objectifs

Dans ce chapitre, nous avons présenté les principales stratégies d'apprentissage actif de la littérature. Cet état-de-l'art se place dans le cadre de l'échantillonnage sélectif (voir Algorithme 1) et s'intéresse aux stratégies capables de résoudre des problèmes de classification. Par la suite, nous avons développé certains aspects de l'apprentissage actif. L'objectif de cette analyse est de justifier nos choix pour l'élaboration d'une stratégie active innovante.

Pour évaluer les stratégies d'apprentissage actif nous choisissons d'utiliser deux critères d'évaluation dans la suite de ce manuscrit. D'une part, l'AUC évalue les stratégies actives en fonction du nombre d'exemples étiquetés $|L|$ (voir Annexe A.a). Pour chaque stratégie, l'AUC permet de tracer une courbe représentant la performance moyenne et la variance

des résultats en fonction de $|L|$. D'autre part, une mesure de déficit est utilisée pour comparer une stratégie active à la stratégie aléatoire sur l'intégralité d'un jeu de données (voir Annexe A.c).

De notre point de vue, les performances attendues d'une stratégie d'apprentissage actif idéale répondent à trois critères : i) une bonne stratégie active doit toujours être meilleure ou équivalente à la stratégie aléatoire, quelque soit le nombre d'exemples étiquetés ; cela se traduit soit par un déficit inférieur à 1, soit par une courbe de performance qui domine la courbe de la stratégie aléatoire pour toutes les valeurs de $|L|$; ii) la performance d'une stratégie active doit toujours croître ou rester constante lors des itérations successives d'un échantillonnage sélectif, la courbe de performance doit être monotone croissante ; iii) une stratégie active doit approcher la performance optimale avant d'avoir étiqueté tous les exemples, $|L| = |\Phi|$.
Nous fixons aussi d'autres objectifs que devrait atteindre une stratégie d'apprentissage actif idéale :

1. Notre premier objectif est d'assurer un bon compromis entre **l'exploitation et l'exploration** des données. La Section 2.3.1 présente un état-de-l'art sur les stratégies actives qui gèrent ce compromis. Ces travaux sont fondés sur le pré-clustering [Nguyen and Smeulders, 2004], sur l'exploitation d'une mesure de dissimilarité [Brinker, 2003; Xu *et al.*, 2007], ou encore sur l'utilisation de plusieurs stratégies dédiées soit à l'exploration, soit à l'exploitation des données [Zoller and Buhmann, 2000; Baram *et al.*, 2004]. Dans le Chapitre 3, nous proposons une stratégie d'apprentissage actif par modèles locaux. Cette stratégie originale partitionne récursivement l'espace $\mathbb{X}$ et met en compétition des modèles locaux à chacune des zones pour choisir les exemples à étiqueter. Un critère de sélection de zone règle explicitement le compromis exploitation / exploration.

2. Une stratégie d'apprentissage actif idéale **exploite les exemples non-étiquetés** sans faire d'hypothèses fortes sur les données. La Section 2.3.2 montre que les principales stratégies actives de la littérature peuvent être améliorées par l'utilisation d'un algorithme d'apprentissage semi-supervisé [Chappelle, 2005; Muslea *et al.*, 2002]. Ces variantes semi-supervisées induisent généralement des hypothèses sur les données. Notre objectif est d'élaborer une stratégie active exploitant un algorithme d'apprentissage semi-supervisé qui soit applicable dans le cas général, où aucune information sur les données n'est disponible. Dans le Chapitre 4, nous proposons une méthode de discrétisation semi-supervisée basée sur une approche Bayésienne objective. Cette méthode de discrétisation adopte des hypothèses faiblement informatives.

3. La stratégie d'apprentissage actif et le modèle prédictif utilisé doivent adopter des **hypothèses a priori faiblement informatives**. La Section 2.3.3 montre que dans la plupart des travaux de la littérature, des hypothèses fortes sont faites sur les données. Selon notre point de vue, ces hypothèses posent problème lors de l'application d'une stratégie active à des données réelles dont la nature est inconnue. Dans le

Chapitre 5, nous présentons une stratégie active fondée sur une approche Bayésienne objective. Cette stratégie, et le modèle de discrétisation qu'elle utilise adoptent des hypothèses faiblement informatives.

4. Le modèle prédictif utilisé lors de l'échantillonnage sélectif doit être **robuste**. La Section 2.3.4 montre qu'en pratique, le modèle prédictif ne peut être évalué qu'en considérant les exemples étiquetés. Une évaluation précise est difficile à obtenir en raison du faible nombre d'exemples étiquetés. L'utilisation d'un modèle prédictif robuste limite ce problème. La méthode de discrétisation MODL [Boullé, 2006b] exploitée dans les Chapitres 4 et 5 est très robuste. Cette méthode de discrétisation s'est distinguée à ce sujet lors de plusieurs challenges internationaux [Boullé, 2007a]. Le Chapitre 4 présente une extension de l'approche MODL au cas de l'apprentissage semi-supervisé. Le Chapitre 5 propose une stratégie active basée sur notre approche de discrétisation semi-supervisée.

5. Une bonne stratégie d'apprentissage actif doit **estimer le biais** qu'elle induit sur la distribution de l'ensemble L. Comme expliqué à la Section 2.3.4, la connaissance de ce biais serait très utile à l'évaluation du modèle prédictif lors d'un échantillonnage sélectif. Une piste pour résoudre ce problème serait d'exploiter la pondération covariative pour estimer ce biais [Sugiyama *et al.*, 2007].

6. Une stratégie d'apprentissage actif doit être munie d'un **critère d'arrêt**. Le critère d'arrêt utilisé dans le cadre de nos expériences prend la forme d'un budget d'étiquetage. D'autres pistes pourraient être explorées, notamment l'élaboration d'un critère d'arrêt basé sur la qualité du model prédictif. Ces pistes sont peu traitées dans la littérature.

7. Une stratégie d'apprentissage actif idéale ne doit **pas impliquer de paramètres** à ajuster. La stratégie d'apprentissage actif par modèles locaux qui est élaborée au Chapitre 3 donne des résultats prometteurs ; mais cette stratégie implique plusieurs paramètres qui sont difficiles à régler en pratique. On se fixe pour objectif d'améliorer cette stratégie grâce aux travaux réalisés aux Chapitres 4 et 5.

Chapitre 3

Apprentissage actif par modèles locaux

Sommaire

Ce chapitre a fait l'objet de publications : [Bondu and Lemaire, 2007a]
[Bondu *et al.*, 2007b]
[Bondu *et al.*, 2007a]
[Bondu and Lemaire, 2008a]
[Bondu and Lemaire, 2008b]

Les travaux présentés dans ce manuscrit sont réalisés dans le cadre de l'échantillonnage sélectif (voir Algorithme 1, page 8). Dans ce cas, la stratégie active n'observe qu'une partie restreinte Φ de l'espace d'entrée $\mathbb{X}$. L'ensemble $\Phi = L \cup U$ contient des exemples étiquetés (L) et des exemples non-étiquetés (U). Les étiquettes associées aux exemples de l'ensemble U peuvent être demandées à un expert moyennant un coût. L'objectif d'une stratégie active est de sélectionner les exemples les plus utiles à l'apprentissage du modèle prédictif, et de les faire étiqueter par l'expert.

La plupart des stratégies d'apprentissage actif de la littérature exploitent des modèles globaux à l'espace d'entrée $\mathbb{X}$. L'idée développée dans ce chapitre est de partitionner $\mathbb{X}$ et d'entraîner des modèles localement à chacune des zones. Nous proposons une stratégie active originale qui réalise un partitionnement dichotomique récursif de l'espace d'entrée et qui met en compétition des modèles locaux afin de choisir les exemples à étiqueter.

La stratégie d'apprentissage actif par modèles locaux présentée dans ce chapitre gère naturellement le compromis exploitation / exploration grâce à un mécanisme de sélection de zones. À chaque itération de l'échantillonnage sélectif, notre stratégie sélectionne la zone de l'espace $\mathbb{X}$ dans laquelle l'étiquetage d'un nouvel exemple améliore le plus le modèle local associé. Le niveau d'intérêt des zones varie au cours de l'échantillonnage sélectif. Par exemple, une zone voit son intérêt diminuer lorsque son modèle local atteint sa performance optimale. Cet aspect dynamique est très important, c'est ce mécanisme qui permet de délaisser les zones fortement exploitées et d'explorer le reste de l'espace $\mathbb{X}$.

Notre stratégie d'apprentissage actif est basée sur la *"curiosité adaptative"*, une heuristique issue de la robotique développementale [Oudeyer and Kaplan, 2004]. La curiosité adaptative a pour but de rendre un robot autonome dans le choix des situations à apprendre. Nos travaux consistent à adapter cette heuristique au cas de l'échantillonnage sélectif et à l'améliorer.

Ce chapitre s'organise de la manière suivante : la Section 3.1 présente la curiosité adaptative d'un point de vue générique, ainsi que les choix initiaux d'implémentation. La Section 3.2 établit un parallèle entre la curiosité adaptative et l'échantillonnage sélectif. Une première implémentation naïve de la curiosité adaptative y est présentée et le comportement de cette stratégie y est illustré sur un problème jouet. En considérant les résultats obtenus à la Section 3.2, un nouveau critère de sélection de zone est défini à la Section 3.3. Notre stratégie active est comparée avec deux autres stratégies de la littérature, dans le cadre d'un problème jouet à la Section 3.3.3, et sur un problème industriel à la Section 3.4. Pour conclure, nous montrons en quoi notre stratégie répond en partie aux questions soulevées en conclusion du Chapitre 2 (Section 2.4, page 27). Nous introduisons également la suite de nos travaux dont l'objectif est d'améliorer la stratégie d'apprentissage actif par modèles locaux.

3.1 Curiosité adaptative

Cette section présente la curiosité adaptative et son lien avec l'échantillonnage sélectif. Un algorithme générique est présenté, ainsi que les choix d'implémentation initiaux [Oudeyer and Kaplan, 2004].

3.1.1 Un lien naturel avec l'apprentissage actif

L'exploration de situations inconnues joue un rôle important dans le développement cognitif des individus. Selon certains psychologues, les comportements que nous adoptons pour apprendre sont intrinsèquement motivants. Cette théorie [White, 1959] explique notre curiosité et nos activités d'explorations comme étant une source de satisfaction. L'élaboration d'un robot animé par les mêmes mécanismes que ceux de l'apprentissage humain est un défi proposé par la robotique développementale. L'ambition de ce champ de recherche est d'élaborer une machine capable de détecter les situations "étonnantes" et de les exploiter lors de son apprentissage.

Y. Nagai [Nagai *et al.*, 2002] montre que l'apprentissage d'un robot peut être accéléré en le soumettant à des situations successives dont la difficulté est croissante. On entend par *"situation"* un état particulier des capteurs du robot. Pour chaque situation le robot reçoit une réponse de son environnement. Le robot peut par exemple être confronté à un obstacle, ou encore, trouver une source d'énergie... etc. Le choix des situations auxquelles le robot est confronté lors de son apprentissage relève du compromis entre l'exploitation, où le robot améliore son comportement pour des situations proches de celles qu'il a déjà rencontré ; et l'exploration de l'environnement, où le robot recherche des situations totalement inconnues.

Le but de la curiosité adaptative est de rendre le robot autonome dans le choix des situations à apprendre. La curiosité adaptative [Oudeyer and Kaplan, 2004] est l'une des voies possibles pour atteindre cet objectif. Le niveau de difficulté des situations apprises doit toujours être adapté à l'apprentissage du robot. Dans le cas idéal, le robot s'intéresse progressivement à des situations de plus en plus difficiles et évite les situations pour lesquelles il ne peut rien apprendre. L'objectif de la curiosité adaptative est de maximiser les progrès du robot en choisissant les bonnes situations, au bon moment.

La première intuition pour mesurer les progrès d'un robot est de comparer des situations successives. Si le robot réalise une tâche mieux que précédemment, il est supposé avoir réalisé des progrès. Y. Nagai illustre, grâce à un exemple simple, qu'une telle mesure de progrès engendre des comportements aberrants. Considérons un robot qui cherche à estimer sa position à l'issue d'un déplacement. Le robot maximise ses progrès en alternant l'immobilité et une collision avec un obstacle. L'immobilité est l'action qui permet au robot de prédire sa prochaine position avec la plus grande exactitude. En comparant cette performance avec l'état précédant, où le robot percute un obstacle, le progrès est supposé être très élevé.

La force de la curiosité adaptative est de comparer des situations similaires et non pas successives [Oudeyer and Kaplan, 2004]. Pour cela, les situations similaires sont regroupées entre elles, ce qui se traduit par le partitionnement récursif de l'espace de variation des

capteurs. Des modèles d'apprentissage locaux à chacune des zones sont entraînés en parallèle. La curiosité adaptative met en jeu un mécanisme de sélection de zone qui détermine le type de situations que le robot doit apprendre, et qui indique quelles sont les zones à partitionner plus finement.

Par analogie avec l'échantillonnage sélectif, la curiosité adaptative peut être vue comme une stratégie active. Le robot correspond à un modèle prédictif qui cherche à résoudre un problème de classification. Les situations auxquelles le robot est confronté lors de son apprentissage correspondent aux exemples sélectionnés par la stratégie active. Enfin, les réponses de l'environnement correspondent aux étiquetages réalisés par l'expert. Comme le montre la suite de ce chapitre, le principe de la curiosité adaptative est transposable au problème de l'échantillonnage sélectif.

3.1.2 Algorithme générique

L'idée principale de la curiosité adaptative est de partitionner l'espace $\mathbb{X}$ en zones munies de modèles prédictifs locaux. Un modèle local est spécialisé dans un type de situations que le robot doit appréhender. L'apprentissage d'un modèle local prend uniquement en compte les situations de la zone correspondante. Le partitionnement de l'espace $\mathbb{X}$ est récursif, c'est-à-dire qu'il évolue au cours du temps et que les zones formées sont incluses les unes dans les autres. Les zones où l'apprentissage des modèles locaux s'améliore le plus sont sélectionnées pour être partitionnées plus finement.

La curiosité adaptative met en jeu un critère de sélection de zones qui définit les régions de l'espace $\mathbb{X}$ que le robot doit exploiter. La sélection des zones d'intérêt est basée sur l'évaluation des progrès réalisés par les modèles locaux. Du point de vue de l'échantillonnage sélectif, la sélection des zones d'intérêt correspond à une stratégie active. À chaque itération, les exemples appartenant à la zone sélectionnée sont candidats à l'étiquetage.

La Figure 3.1 présente un exemple illustratif de partitionnement récursif effectué par la curiosité adaptative. L'espace d'entrée $\mathbb{X}$ contient deux variables x_1 et x_2 symbolisées par les axes horizontal et vertical des trois graphiques. À l'itération Q, trois zones sont associées aux modèles m_1, m_2, et m_3. L'apprentissage de ces modèles locaux exploite trois ensembles d'apprentissage disjoints (l_1, u_1), (l_2, u_2), (l_3, u_3). Le partitionnement de l'espace $\mathbb{X}$ est réalisé progressivement, à mesure que des exemples sont étiquetés (itérations $Q + Q'$ et $Q + Q' + Q''$). Lors du partitionnement d'une zone, le modèle local associé est dupliqué dans les zones filles. À l'itération $Q + Q'$, le modèle m_2 est dupliqué et les exemples de la zone 2 sont partagés en deux sous-ensembles (l_{21}, u_{21}) et (l_{22}, u_{22}), vers les zones filles 2 et 4. Par la suite, les modèles locaux continuent leur apprentissage indépendamment grâce aux exemples appartenant à leur zone. Ainsi, à l'itération $Q + Q' + Q''$ les zones 2 et 4 sont munies de modèles locaux différents, m_2 et m_4.

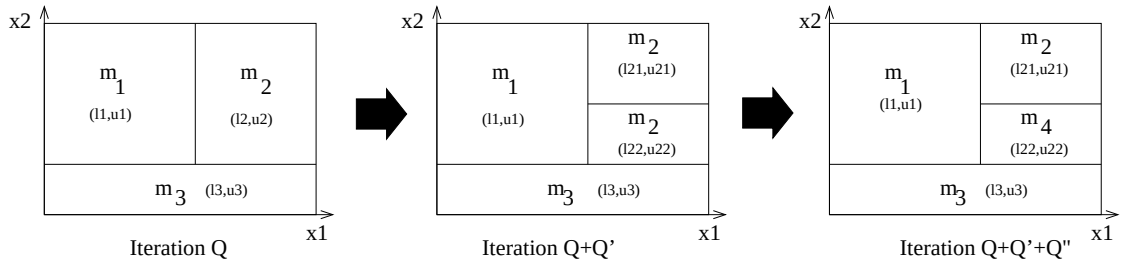

FIG. 3.1 – Partitionnement de l'espace d'entrée $\mathbb{X}$.

Notation :

- Un algorithme d'apprentissage $\mathcal{L}$
- Un ensemble $\mathcal{M} = \{m_1, m_2, ..., m_n\}$ de n modèles prédictifs locaux
- $U = \{u_1, u_2, ..., u_n\}$ l'ensemble des exemples non-étiquetés
- $L = \{l_1, l_2, ..., l_n\}$ l'ensemble des exemples étiquetés partitionné en n sous-ensembles
- $T = \{t_1, t_2, ..., t_n\}$ les sous-ensembles d'apprentissage correspondant aux modèles locaux, avec $t_i = \{(x, f(x))\} \quad \forall x \in l_i$

$n \leftarrow 1$

Répéter

(A) Choisir un modèle local m_i à alimenter en exemples
(B) Tirer un nouvel exemple x^* dans l_i
(C) Étiqueter l'exemple x^*, $t_i \leftarrow t_i \cup (x^*, f(x^*))$
(D) Entraîner le modèle local m_i grâce à $\mathcal{L}$, u_i et t_i
Si le critère de séparation est satisfait **Alors**
(i) Partitionner l_i en deux sous-ensembles l_j et l_k
(ii) Dupliquer m_i en deux modèles locaux m_j et m_k
(iii) $n \leftarrow n + 1$
Fin Si

Tant qu'on peut étiqueter des exemples

Algorithme 4: Curiosité adaptative

L'Algorithme 4 montre les principales étapes de la curiosité adaptative. Cet algorithme est écrit dans le cadre de l'échantillonnage sélectif et peut être assimilé à une stratégie d'apprentissage actif. Un premier critère est utilisé pour choisir la zone à alimenter en exemples (étape A). L'étape suivante consiste à tirer un exemple dans la zone sélectionnée (étape B). L'expert étiquette l'exemple sélectionné (étape C) et le modèle local est entraîné en considérant un exemple étiqueté supplémentaire (étape D). Un deuxième critère détermine si la zone courante doit être, ou non, partitionnée. Si tel est le cas, la zone "mère" est partitionnée en deux zones "filles" (étape i). Enfin, le modèle local associé à la zone "mère" est dupliqué dans les zones "filles" (étape ii).

La curiosité adaptative cherche à définir les zones d'intérêt de l'espace $\mathbb{X}$, où les modèles locaux progressent le plus.

3.1.3 Paramètres : les choix initiaux d'Oudeyer et al.

L'Algorithme 4 présente la curiosité adaptative d'un point de vue générique, l'implémentation de chaque étape reste à définir. Pour mettre en œuvre la curiosité adaptative, il est nécessaire de répondre aux questions suivantes :

1. Comment sélectionner les zones à alimenter en exemples étiquetés ?
2. Comment sélectionner les exemples à étiqueter au sein d'une zone ?
3. Comment décider si une zone doit être, ou non, partitionnée ?
4. En combien de zones "filles" une zone "mère" doit-elle être partitionnée ?
5. Quels modèles prédictifs locaux utiliser ?

Les choix d'implémentation initiaux [Oudeyer and Kaplan, 2004] apportent une première réponse à ces questions.

Sélection de zones : ***(questions 1 et 2)***
À chaque itération de l'échantillonnage sélectif, le modèle local qui s'améliore le plus est considéré comme ayant le plus fort potentiel d'amélioration. La zone associée au modèle local qui réalise les progrès les plus importants est sélectionnée. Un exemple à étiqueter est choisi de manière aléatoire dans cette zone. Selon les auteurs [Oudeyer and Kaplan, 2004], la sélection de zones d'intérêt est fondée sur l'évaluation des progrès réalisés par les modèles locaux. Deux étapes sont nécessaires : i) la performance des modèles locaux est évaluée grâce aux exemples étiquetés de chaque zone, étant donnée une mesure de performance ; ii) les progrès des modèles locaux sont évalués grâce aux variations de leurs performances sur une fenêtre temporelle. La sélection des zones d'intérêt implique donc deux paramètres : une mesure de performance et une fenêtre temporelle.

Partitionnement : ***(questions 3 et 4)***
Selon Oudeyer et al, une zone doit être partitionnée quand le nombre d'exemples étiquetés qu'elle comporte dépasse un certain seuil. Les zones les plus peuplées sont intéressantes à partitionner car les modèles locaux associés ont réalisé des progrès importants lors des itérations précédentes. Le seuil de partitionnement est un premier paramètre à régler.

Les auteurs [Oudeyer and Kaplan, 2004] choisissent de partitionner une zone "mère" en deux zones "filles". Pour effectuer ce partitionnement, chaque dimension de l'espace $\mathbb{X}$ est considérée. Pour chaque dimension, toutes les valeurs de coupure possibles sont évaluées. Le critère de partitionnement cherche la dimension et la valeur de coupure qui minimisent la variance des prédictions du modèle de part et d'autre de la coupure. Tous les exemples (étiquetés et non-étiquetés) sont exploités lors de cette étape. Ce critère de partitionnement tend à élaborer des zones pures, ce qui facilite l'apprentissage des

modèles locaux. Une contrainte supplémentaire est définie par les auteurs : une coupure doit séparer les exemples étiquetés d'une zone en deux sous-ensembles dont les effectifs sont à peu près équilibrés[1]. La proportion d'exemples étiquetés dans chacune des zones "filles" est un deuxième paramètre à régler.

Modèle locaux : ***(question 5)***
Dans les travaux initiaux [Oudeyer and Kaplan, 2004], l'apprenant est un robot simulé par ordinateur qui est composé d'un ensemble de modèles locaux. Ces modèles exploitent la méthode des K-plus proches voisins pour effectuer leurs prédictions. La sélection des zones d'intérêt et le partitionnement de l'espace $\mathbb{X}$ dépendent des modèles locaux employés.

3.2 Transposition à la classification

Notre objectif est de transposer la curiosité adaptative aux problèmes de classification, en modifiant au minimum les choix d'implémentation initiaux. Cette section étudie le comportement de la curiosité adaptative sur un problème jouet.

3.2.1 Paramétrage de la curiosité adaptative

Le paramétrage de l'Algorithme 4 exploité lors de notre expérience est présenté.

Sélection de zones :
Les performances des modèles locaux sont évaluées grâce l'aire sous la courbe de ROC, aussi appelée AUC (*Area Under roc Curve*). Cette mesure de performance est présentée en détail dans l'Annexe A.a. Les progrès réalisés par les modèles locaux sont évalués sur une fenêtre temporelle composée de deux itérations successives de l'échantillonnage sélectif. Le progrès est défini comme suit, avec $l \subseteq L$ l'ensemble des exemples étiquetés appartenant à la zone considérée :

$$\mathcal{P}rogress(l) = AUC_t(l) - AUC_{t-1}(l)$$

Partitionnement :
Le seuil de partitionnement est fixé à 30 exemples étiquetés par zone. Au delà de ce seuil, une zone est partitionnée en deux zones filles selon le critère de partitionnement présenté à la Section 3.1.3. La frontière entre les zones "filles" sépare les exemples étiquetés en deux sous-ensembles équilibrés à $\pm 25\%$.

Modèle locaux :
Les modèles locaux utilisés sont des régressions logistiques implémentées par un réseau de neurones [Sarle, 1994], dont l'architecture est représentée par la Figure 3.2. Ce perceptron possède un neurone caché (H) dont la fonction de transfert est linéaire, et un neurone de sortie (O) dont la fonction de transfert est une sigmoïde. Le vecteur de poids

[1] Le seuil de tolérance sur l'équilibre des classes constitue un paramètre à ajuster.

$[w_1..w_4]$ regroupe les paramètres du modèle qui sont ajustés durant la phase d'apprentissage. Les deux premières variables d'entrée x_1 et x_2 caractérisent les exemples d'apprentissage, $\mathbb{X} = x_1 \times x_2$. Le "*biais*" du réseau de neurone est une entrée supplémentaire dont la valeur est fixée à 1. Ce biais permet de faire varier l'ordonnée à l'origine de la séparatrice linéaire apprise par le modèle, en ajustant le poids w_3. La sortie du modèle est normalisée dans l'intervalle $[0, 1]$ et correspond à la probabilité d'observer la classe "1" conditionnellement à l'exemple placé en entrée du modèle. La probabilité de la classe "2" est égale à 1 moins la valeur de sortie. L'apprentissage de ce réseau de neurones s'arrête lorsque le différentiel de l'erreur d'apprentissage est inférieur à 10^{-8}. Le pas d'apprentissage est fixé à 10^{-2}.

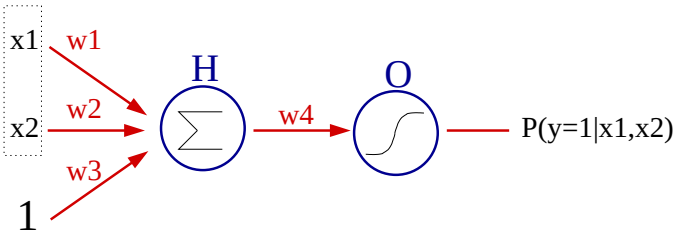

FIG. 3.2 – Régression logistique implémentée par un réseau de neurones, $\mathbb{X} = x_1 \times x_2$.

Dans le cadre de notre expérience, la régression logistique est utilisée à deux fins. D'une part, ce modèle prédictif est utilisé en tant que modèle local lors de l'implémentation de la curiosité adaptative. La Figure 3.3 représente les modèles locaux $m_1, m_2...m_5$ associés à chaque zone. D'autre part, la régression logistique est utilisée comme modèle global lors de l'évaluation de cette stratégie d'apprentissage actif. Le modèle global est représenté par m_* sur la Figure 3.3. Ce modèle est entraîné indépendamment du partitionnement de l'espace $\mathbb{X}$, en utilisant les exemples sélectionnés par la curiosité adaptative. La performance du modèle m_* reflète la qualité des exemples sélectionnés par la curiosité adaptative. L'utilisation d'un modèle prédictif global permet de comparer la curiosité adaptative aux autres stratégies actives de la littérature de manière cohérente.

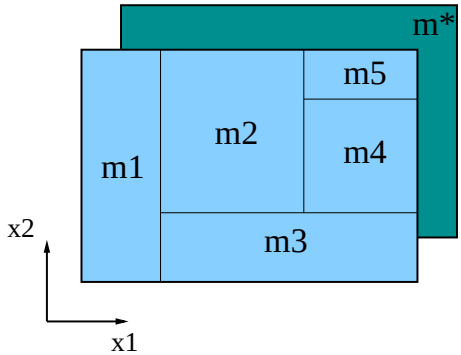

FIG. 3.3 – Deux utilisations de la régression logistique : modèles locaux et modèle global, $\mathbb{X} = x_1 \times x_2$

3.2.2 Conditions expérimentales

Cette section présente les conditions expérimentales de notre étude.

Problème jouet :
Le problème jouet considéré est une classification binaire dans un espace à deux dimensions, $\mathbb{X} = x_1 \times x_2$. Les deux variables qui caractérisent les données sont définies sur les intervalles $x_1 \in [-2, 2]$ et $x_2 \in [-2, 2]$. Comme le montre la Figure 3.4, les deux classes sont séparées par la frontière $x_2 = sin((x_1)^3)$. 2000 exemples d'apprentissage (Φ) et 30000 exemples de test sont générés uniformément et utilisés lors de notre expérience.

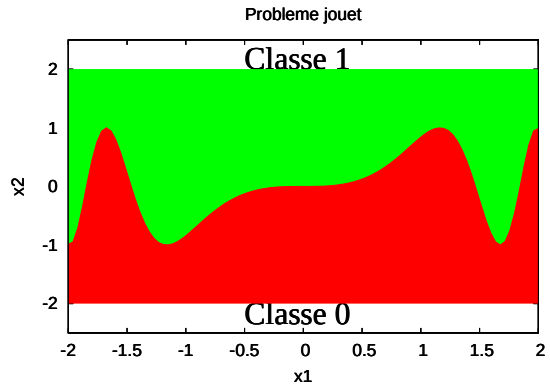

FIG. 3.4 – Problème jouet : classification binaire dont la séparatrice est donnée par la fonction $x_2 = sin((x_1)^3)$. La partie supérieure [*respectivement* inférieure] de la figure correspond à la classe "1" [*respectivement* la classe "2"].

Solution optimale :
Une régression logistique est utilisée en tant que modèle global pour évaluer la capacité de la curiosité adaptative à étiqueter des exemples utiles (Section 3.2.1). Ce modèle prédictif fait l'hypothèse que les données sont organisées selon un hyperplan séparateur. La Figure 3.5 représente les valeurs de sortie de ce modèle, lorsque tous les exemples d'apprentissage sont étiquetés, $|L| = |\Phi| = 2000$. Du point de vue de l'échantillonnage sélectif, ce modèle prédictif est optimal. Le graphique de gauche (Figure 3.5) représente, par le biais d'un code couleur, les valeurs de la sortie du modèle dans l'espace $\mathbb{X} = x_1 \times x_2$. La frontière de décision de la régression logistique correspond à la ligne de niveau 0.5. Au dessus de cette droite [*respectivement* en dessous] ce modèle prédit la classe "1" [*respectivement* la classe "0"]. Le graphique de droite (Figure 3.5) permet de visualiser les valeurs de sortie de ce modèle grâce à l'axe vertical. Les valeurs de sortie ont une allure de sigmoïde, lorsqu'elles sont observées selon une coupe orthogonale aux lignes de niveaux. Ce phénomène est propre à la régression logistique. Le modèle prédictif optimal atteint une AUC de 0.957, cette performance est atteinte par les stratégies d'apprentissage actif lorsque tous les exemples sont étiquetés.

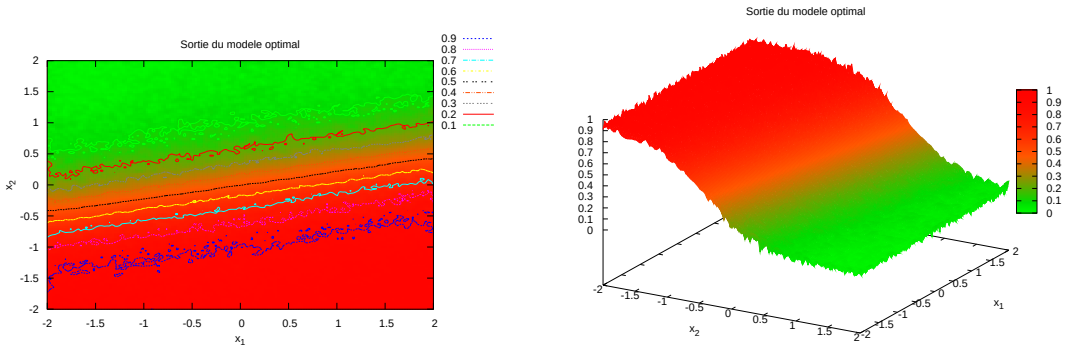

FIG. 3.5 – Valeurs de sortie de la régression logistique optimale, dans l'espace $\mathbb{X} = x_1 \times x_2$.

Protocole expérimental :
Les données sont préalablement centrées et réduites. Au début de chaque expérience, l'ensemble L ne comporte que deux exemples choisis de manière aléatoire. À chaque itération, un exemple est étiqueté dans la zone où le modèle local s'améliore le plus. L'échantillonnage sélectif s'arrête lorsque le nombre total d'exemples étiquetés atteint 250.

Pour rappel, le modèle prédictif utilisé est une régression logistique implémentée par un réseau de neurones, ce modèle est global à l'espace $\mathbb{X}$ (Section 3.2.1). La performance du modèle prédictif est évaluée grâce à l'AUC (Annexe A.a), sur un ensemble de test qui contient 30 000 exemples. La performance évaluée par l'AUC reflète la qualité des exemples sélectionnés.

La stratégie d'échantillonnage aléatoire est également évaluée, cette stratégie sélectionne les exemples à étiqueter uniformément selon leur distribution de probabilité. L'échantillonnage aléatoire joue un rôle de référence et permet de mesurer la contribution apportée par la curiosité adaptative sur la sélection des exemples. Cette contribution est évaluée grâce à la mesure de déficit présentée dans l'Annexe A.c.

Pour chaque stratégie, la performance moyenne du modèle prédictif est évaluée en fonction du nombre d'exemples étiquetés. Les expériences sont répétées 10 fois, de manière à obtenir une AUC moyenne et une variance pour chaque point des courbes de résultats.

3.2.3 Résultats

Cette section montre le comportement de la curiosité adaptative sur le problème jouet présenté à la Section 3.2.2. La performance du modèle prédictif est étudiée, ainsi que la localisation des exemples étiquetés dans l'espace $\mathbb{X}$.

Performance du modèle prédictif :
La Figure 3.6 présente la performance moyenne du modèle prédictif (axe vertical) en fonction du nombre d'exemple étiquetés (axe horizontal)[2]. Le modèle prédictif global est soit entrainé grâce aux exemples étiquetés par la curiosité adaptative (courbe verte), soit grâce aux exemples étiquetés par l'échantillonnage aléatoire (courbe rouge). Les courbes de la Figure 3.6 correspondent à l'AUC moyenne et les moustaches représentent la variance de l'AUC ($\pm 2\sigma$). La Figure 3.6 trace également la performance optimale atteinte par le modèle prédictif lorsque tous les exemples sont étiquetés.

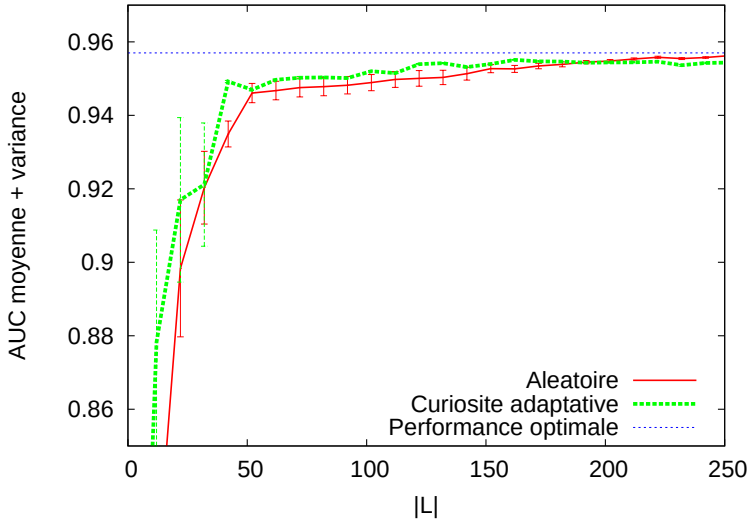

FIG. 3.6 – AUC moyenne *vs.* nombre d'exemples étiquetés. Les moustaches sur les courbes représentent la variance de l'AUC observée sur 10 expériences($\pm 2\sigma$).

La mesure de déficit définie à la Section A.c compare une stratégie active à la stratégie aléatoire sur l'intégralité d'un jeu de données. Le déficit de la curiosité adaptative calculé entre 0 et 250 exemples est égal à 0.65. Plus le déficit est proche de 0, plus la stratégie active domine l'aléatoire. À l'inverse, lorsque le déficit dépasse 1, la stratégie active est moins performante que l'aléatoire. Dans le cadre de cette expérience, la curiosité adaptative donne des performances légèrement meilleures que la stratégie aléatoire sans toutefois être nettement supérieures.

Exemples sélectionnés par la curiosité adaptative :
La Figure 3.7 montre les exemples sélectionnés par la curiosité adaptative durant une expérience. Les zones définies lors du partitionnement de l'espace $\mathbb{X}$ apparaissent également sur la Figure 3.7. Le partitionnement de l'espace et le choix des exemples sont relativement uniformes ; même si une zone un peu plus densément peuplée peut être observée pour chaque classe (en haut à droite et au milieu en bas de la Figure 3.7). Ces deux zones densément peuplées indiquent que la régression logistique peut progresser fortement dans

[2]Pour des questions de lisibilité, les courbes de la Figure 3.6 tracent l'AUC moyenne tous les 10 exemples.

des zones pures de l'espace $\mathbb{X}$. L'implémentation naïve de la curiosité adaptative n'est pas satisfaisante car les zones contenant le plus d'exemples ne s'organisent pas autour du motif à découvrir.

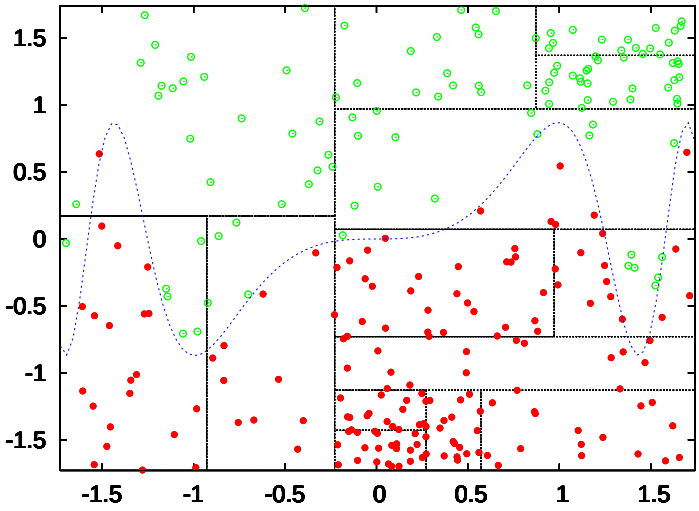

FIG. 3.7 – Sélection des zones basée sur le progrès des modèles locaux : le progrès est évalué sur une fenêtre temporelle de deux itérations successives, en utilisant l'AUC comme mesure de performance. Le partitionnement de l'espace $\mathbb{X}$ est représenté sur cette figure, ainsi que les 250 exemples étiquetés (○ classe "1", et ● classe "0")

Cette section montre comment la curiosité adaptative peut être transposée au problème de l'échantillonnage sélectif, en modifiant au minimum les choix d'implémentation initiaux. Cette transposition est valide et constitue le premier apport de ce chapitre. Pour conclure, l'expérience réalisée dans cette section apporte deux résultats importants. D'une part, la curiosité adaptative peut être utilisée en tant que stratégie d'apprentissage actif. D'autre part, l'implémentation naïve de la curiosité adaptative n'améliore pas significativement la qualité des exemples sélectionnés.

3.3 Amélioration : Un nouveau critère de sélection de zones

Au vue des résultats obtenus à l'issue de la Section 3.2, nous nous fixons pour objectif d'améliorer notre stratégie d'apprentissage actif par modèles locaux.

Quatre aspects de notre stratégie peuvent être améliorés : i) le critère de sélection de zones ; ii) le critère qui décide *quand* couper une zone ; iii) le critère qui décide *où* couper une zone ; iv) la sélection des exemples à étiqueter dans une zone. Cette section est consacrée à la sélection des zones d'intérêt dans l'espace d'entrée $\mathbb{X}$, les Chapitres 4 et 5 traitent des autres items.

Nous proposons un nouveau critère de sélection de zone qui gère explicitement le compromis entre l'exploitation et l'exploration de l'espace d'entrée $\mathbb{X}$. Ce critère est composé

de deux termes présentés aux Sections 3.3.1 et 3.3.2, respectivement dédiés à l'exploitation et à l'exploration des données. La Section 3.3.3 montre comment ces deux termes sont agrégés et pondérés dans notre critère. La curiosité adaptative munie de ce nouveau critère est comparée à deux autres stratégies actives de la littérature.

3.3.1 Exploitation : Taux de mélange

Le premier terme de notre critère est consacré à l'exploitation des données et mesure le taux de mélange des classes dans une zone. Ce terme est fondé sur la théorie de l'information de Shannon [Shannon, 1948], le taux de mélange est exprimé grâce à l'entropie des classes. La fonction $\mathcal{M}ix\mathcal{R}ate(l)$ (Équation 3.1) est calculée en exploitant les exemples étiquetés appartenant à une zone, notée l. La partie "A" de l'Équation 3.1 est l'entropie des classes de la zone considérée. Les probabilités des classes $p(y_i)$ sont estimées empiriquement, par la proportion des exemples étiquetés de chaque classe dans la zone. L'entropie appartient à l'intervalle $[0, \log|\mathbb{Y}|]$, avec $|\mathbb{Y}|$ le nombre de classes. La partie "B" de l'Équation 3.1 normalise le taux de mélange dans l'intervalle $[0, 1]$.

$$\mathcal{M}ix\mathcal{R}ate(l) = \underbrace{-\sum_{y_i \in \mathbb{Y}} p(y_i) \log p(y_i)}_{A} \times \underbrace{\frac{1}{\log|\mathbb{Y}|}}_{B} \tag{3.1}$$

$$avec \;\; p(y_i) = \frac{|x \in l,\ f(x) = y_i|}{|l|}$$

Le taux de mélange est utilisé par notre critère pour "exploiter" les données. En sélectionnant les zones ayant le plus fort taux de mélange, les exemples sont préférentiellement étiquetés dans des régions de l'espace $\mathbb{X}$ proches de la frontière du motif à découvrir. Les modèles locaux associés à ces zones sont localement très performants. La Figure 3.8 montre les exemples sélectionnés lors d'une expérience réalisée sur le problème jouet de la Section 3.2.2. Cette expérience utilise uniquement le taux de mélange pour sélectionner les zones d'intérêt. Les exemples sélectionnés sont regroupés autour de la frontière du motif à découvrir. Une grande partie de l'espace $\mathbb{X}$ est cependant occultée, faute "d'exploration". Les zones faiblement mélangées comportent peu d'exemples étiquetés. Pourtant, ces zones incluent parfois des éléments du motif à découvrir. Par exemple, la frontière du motif passe dans la zone supérieure de la Figure 3.8, cette zone est peu mélangée au vue des exemples étiquetés.

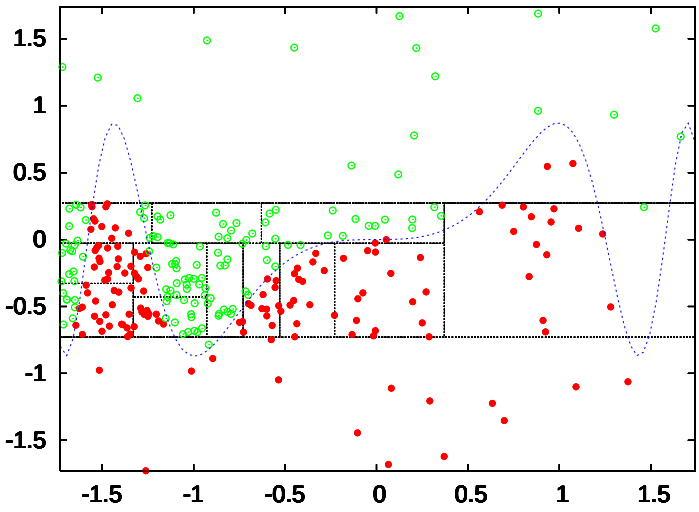

FIG. 3.8 – Exemples sélectionnés en utilisant uniquement le taux de mélange. Le partitionnement de l'espace $\mathbb{X}$ est représenté sur cette figure, ainsi que les 250 exemples étiquetés (○ classe "1", et ● classe "0")

3.3.2 Exploration : Densité relative

Le deuxième terme de notre critère est consacré à l'exploration des données et mesure la densité relative des zones. La densité relative est la proportion d'exemples étiquetés dans une zone. L'Équation 3.2 présente la densité relative, avec $\phi \subseteq \Phi$ le sous-ensemble des exemples, étiquetés et non-étiquetés, appartenant à une zone. Tout comme le taux de mélange, la densité relative varie dans l'intervalle $[0, 1]$.

$$\mathcal{R}elative\mathcal{D}ensity(l, \phi) = \frac{|l|}{|\phi|} \tag{3.2}$$

La densité relative est utilisée par notre critère pour "explorer" l'espace les données. L'homogénéité du tirage des exemples dans l'espace $\mathbb{X}$ est assurée par la sélection des zones qui ont la plus faible densité relative. La Figure 3.9 présente une expérience réalisée sur le problème jouet de la Section 3.2.2. Cette expérience utilise la densité relative comme unique critère de sélection de zones. Le partitionnement de l'espace d'entrée $\mathbb{X}$ ainsi que le tirage des exemples étiquetés sont homogènes. La sélection des exemples est plus "uniforme" qu'un tirage aléatoire global à l'espace $\mathbb{X}$. Dans notre cas, il ne peut pas y avoir par accident des régions de l'espace $\mathbb{X}$ plus densément peuplées que d'autres.

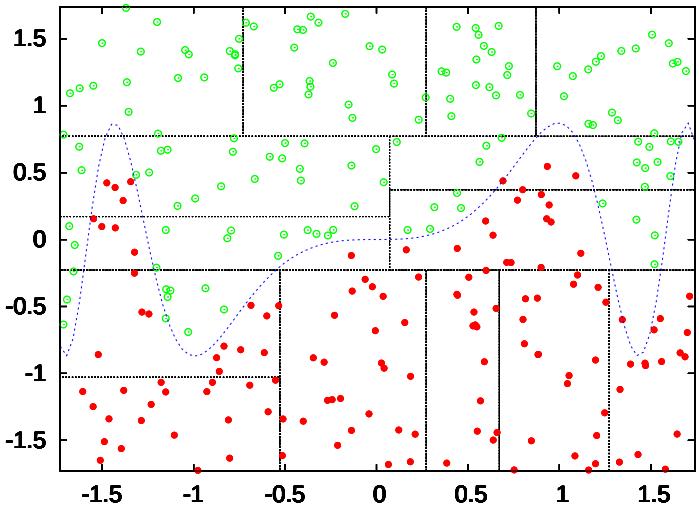

FIG. 3.9 – Exemples sélectionnés en utilisant uniquement la densité relative. Le partitionnement de l'espace $\mathbb{X}$ est représenté sur cette figure, ainsi que les 250 exemples étiquetés (○ classe "1", et ● classe "0")

3.3.3 Compromis entre l'exploitation et l'exploration

Nous proposons un critère qui évalue l'intérêt des zones en tenant compte à la fois du taux de mélange et de la densité relative. L'équation 3.3 montre comment chacun des termes est utilisé. Le paramètre $\alpha \in [0, 1]$ ajuste de manière explicite le compromis entre l'exploitation des zones de mélange déjà connues et l'exploration de nouvelles régions de l'espace $\mathbb{X}$.

$$\mathcal{I}nterest(l, \phi) = (1 - \alpha)\, \mathcal{M}ix\mathcal{R}ate(l) + \alpha\, (1 - \mathcal{R}elative\mathcal{D}ensity(l, \phi)) \tag{3.3}$$

Lors des itérations successives d'un échantillonnage sélectif, notre critère de sélection de zones adopte un comportement dynamique. À l'itération t, deux changements sont observés pour la zone sélectionnée. D'une part, la densité relative de cette zone augmente lorsque qu'un nouvel exemple est étiqueté. D'autre part, le taux de mélange de cette zone varie selon la valeur de la nouvelle étiquette. Si l'exemple étiqueté est de la classe majoritaire [*respectivement* minoritaire], le taux de mélange diminue [*respectivement* augmente]. La zone sélectionnée à l'itération t perd de son intérêt à l'itération $t + 1$, lorsque son taux de mélange ne croît pas suffisamment pour compenser l'augmentation de sa densité relative. Ainsi, les zones où il n'y a plus rien à découvrir sont naturellement évitées. Il est important de noter que les zones comportant peu d'exemples voient leur densité relative augmenter plus rapidement que les zones densément peuplées, lors de l'étiquetage d'un nouvel exemple. Si deux zones ont la même densité relative et le même taux de mélange, alors, après quelques itérations et en supposant que chacune des zones reçoive le même nombre d'exemples étiquetés, notre critère préfèrera la zone la plus densément peuplée.

Au début de l'échantillonnage sélectif, les zones mises en compétition ont des densités relatives proches les unes des autres. Les zones qui présentent un fort taux de mélange sont préférentiellement sélectionnées. Lorsque le nombre d'exemples étiquetés dépasse un certain seuil, ces zones sont partitionnées. Dans un deuxième temps, notre critère porte son attention sur des zones peu mélangées qui n'ont pas encore été assez explorées. La notion de progrès est naturellement prise en compte par notre critère, lors des itérations successives de l'échantillonnage sélectif. L'évaluation des progrès réalisés par les modèles locaux n'implique pas de fenêtre temporelle (Section 3.2.1). L'implémentation de notre critère est donc moins contraignante que celle du critère initial.

La Figure 3.10 montre une expérience réalisée sur l'exemple jouet de la Section 3.2.2. Lors de cette expérience, notre critère est utilisé pour sélectionner les zones d'intérêt, avec $\alpha = \frac{1}{2}$. Le partitionnement de l'espace $\mathbb{X}$ et la sélection des exemples s'organisent autour du motif à découvrir. Cette fois ci, aucune région de l'espace n'est délaissée.

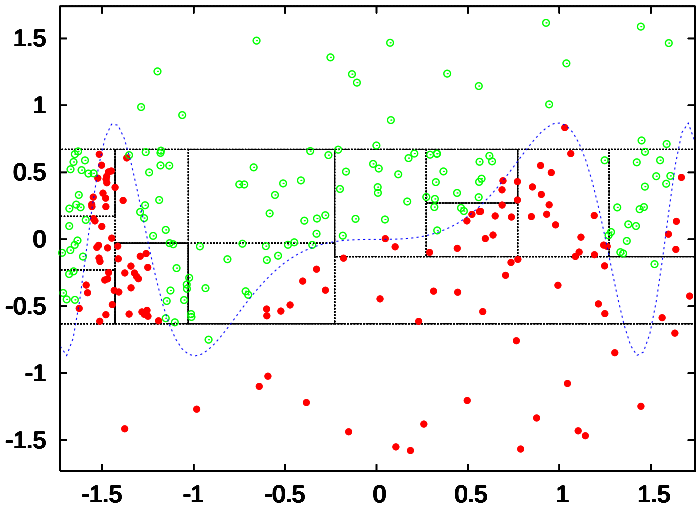

FIG. 3.10 – Exemples sélectionnés pour $\alpha = 0.5$. Le partitionnement de l'espace $\mathbb{X}$ est représenté sur cette figure, ainsi que les 250 exemples étiquetés (○ classe "1", et ● classe "0")

Sensibilité du paramètre α :
L'objectif de ce paragraphe est d'évaluer l'influence du paramètre α sur la performance de notre stratégie active. Plusieurs séries d'expériences sont réalisées pour $\alpha = [0, 0.25, 0.5, 0.75, 1]$, sur le problème jouet présenté à la Section 3.2.2. Les résultats obtenus sont comparés à la stratégie aléatoire.

La Figure 3.11 montre les performances de notre stratégie pour les différentes valeurs de α. Lorsque $\alpha = 0$ (courbe bleue), notre critère de sélection de zone exploite uniquement le taux de mélange. Dans ce cas et pour $|L|$ inférieur à 100, les performances observées sont significativement moins bonnes que la stratégie aléatoire (courbe rouge). Une forte exploi-

tation des zones mélangées est réalisée, au détriment du reste de l'espace $\mathbb{X}$. Lorsque $\alpha = 1$ (courbe orange), seule la densité relative est considérée par notre critère de sélection de zones. Dans ce cas et pour $|L| < 70$, la curiosité adaptative donne des performances moins bonnes que la stratégie aléatoire. La meilleure performance est obtenue avec $\alpha = 0.25$, sur ce problème jouet cette valeur offre un bon compromis entre l'exploitation et exploration des données. La performance de notre stratégie est supérieure à la stratégie aléatoire quelque soit la valeur de $|L|$. Dans ce cas, notre stratégie atteint l'AUC maximale en étiquetant seulement 100 exemples, alors que la stratégie aléatoire atteint cette performance pour $|L| = 2000$.

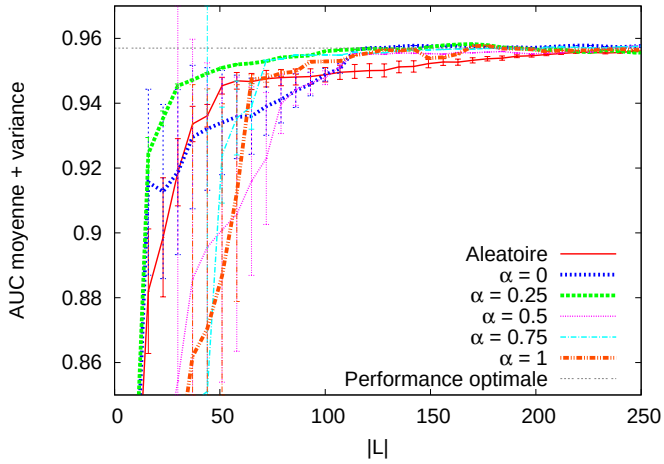

FIG. 3.11 – AUC moyenne *vs.* nombre d'exemples étiquetés. Les moustaches sur les courbes représentent la variance de l'AUC observée sur 10 expériences($\pm 2\sigma$).

La Figure 3.12 présente un tableau résumant les valeurs du déficit mesuré pour les différentes valeurs de α. La colonne de gauche correspond au déficit calculé entre 0 et 250 exemples étiquetés, et la colonne de droite présente le déficit calculé entre 80 et 250 exemples étiquetés. Ces deux plages de valeurs de $|L|$ correspondent à deux phases de la Figure 3.11 : i) en considérant moins de 80 exemples étiquetés, seule la valeur $\alpha = 0.25$ domine la stratégie aléatoire et la variance de l'AUC est forte ; ii) au delà de 80 exemples étiquetés notre stratégie domine significativement la stratégie aléatoire, quelque soit la valeur du paramètre α, dans ce cas la variance de l'AUC est faible. Pour $|L| \in [0, 250]$, le paramétrage $\alpha = 0.25$ donne un déficit légèrement meilleur que l'implémentation naïve de la curiosité adaptative proposée à la Section 3.2. Pour $|L| \in [80, 250]$, les déficits sont très faibles quelque soit la valeur du paramètre α. Les déficits négatifs observés pour $\alpha = 0.25$ et $\alpha = 0.75$ indiquent que notre stratégie atteint une meilleure performance que la stratégie aléatoire pour $|L| = 250$.

L'étude menée dans ce paragraphe montre que α est un paramètre sensible, qui influence les performances de notre stratégie active par modèles locaux. Les résultats obtenus sont néanmoins très encourageants et montrent a posteriori l'existence d'un paramétrage optimal pour lequel notre stratégie est performante.

	$\lvert L\rvert \in [0, 250]$	$\lvert L\rvert \in [80, 250]$
$\alpha = 0$	0.84	0.28
$\alpha = 0.25$	0.64	-0.05
$\alpha = 0.5$	1.59	0.64
$\alpha = 0.75$	1.66	-0.01
$\alpha = 1$	1.59	0.26

FIG. 3.12 – Calcul du déficit de la curiosité adaptative pour $\alpha = [0, 0.25, 0.5, 0.75, 1]$. La première [*respectivement* la deuxième] colonne de ce tableau correspond au déficit calculé entre 0 et 250 [*respectivement* entre 80 et 250] exemples étiquetés.

Comparaison à deux stratégies de l'état-de-l'art :
L'objectif de ce paragraphe est de comparer notre stratégie d'apprentissage actif par modèles locaux à des stratégies de la littérature. Les données utilisées pour cette expérience, ainsi que le protocole expérimental sont les mêmes qu'à la Section 3.2.2. Les stratégies évaluées sont les suivantes :

- l'échantillonnage aléatoire (Section 3.2.2) ;
- l'échantillonnage par incertitude (Section 2.2.1, page 9) ;
- l'échantillonnage par réduction de l'erreur de généralisation (Section 2.2.4, page 16) ;
- l'apprentissage actif par modèles locaux (Section 3.3), deux valeurs sont retenues pour le paramétrage de α : i) la valeur la plus avantageuse $\alpha = 0.25$; ii) la valeur la moins avantageuse $\alpha = 1$.

La Figure 3.13 présente les performances moyennes de chaque stratégie (axe vertical) en fonction du nombre d'exemples étiquetés (axe horizontal). Comme précédemment, la performance d'une stratégie est évaluée grâce à l'AUC moyenne (Annexe A.a) et les moustaches sur les courbes représentent la variance de l'AUC.

Notre stratégie d'apprentissage actif par modèles locaux offre les meilleures performances lorsqu'elle est paramétrée avec $\alpha = 0.25$ (courbe verte). Dans ce cas, notre stratégie domine l'échantillonnage aléatoire de manière significative quelque soit la valeur de $|L|$ (courbe rouge). En considérant le paramétrage le moins avantageux, notre stratégie est moins performante que l'échantillonnage aléatoire pour $|L| < 60$ (courbe rose : $\alpha = 1$). Au delà de 60 exemples étiquetés, notre stratégie domine l'échantillonnage aléatoire. Quelque soit la valeur du paramètre α et la valeur de $|L|$, notre stratégie domine l'échantillonnage par réduction de l'erreur de généralisation (courbe bleue) et l'échantillonnage par incertitude (courbe grise).

L'échantillonnage par réduction de l'erreur de généralisation est moins performant que l'échantillonnage aléatoire, lorsque le nombre d'exemples étiquetés est inférieur à 140. Cette stratégie est légèrement meilleure que l'aléatoire pour $|L| > 140$. Les mauvais résultats observés au début de l'échantillonnage sélectif peuvent s'expliquer intuitivement : l'évaluation de l'erreur de généralisation requière un nombre important d'exemples étiquetés pour être fiable. La partie gauche de la Figure 3.14 présente les exemples sélectionnés grâce à l'échantillonnage par réduction de l'erreur de généralisation. Ces exemples se situent

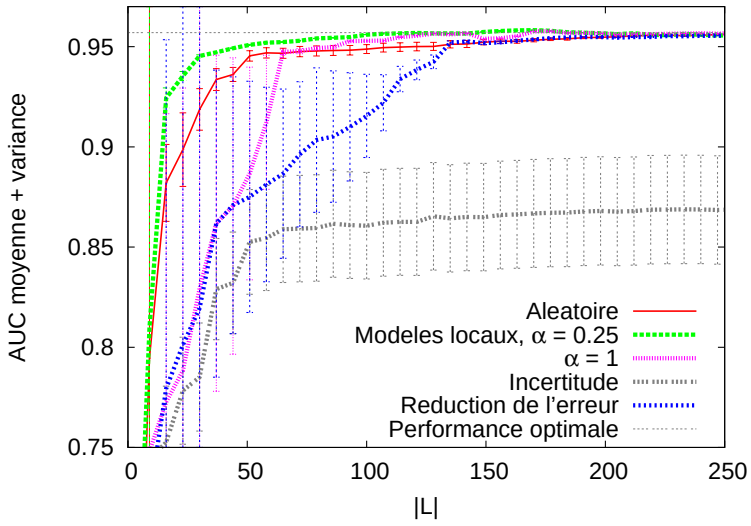

FIG. 3.13 – AUC moyenne *vs.* nombre d'exemples étiquetés. Les moustaches sur les courbes représentent la variance de l'AUC observée sur 10 expériences($\pm\frac{\sigma}{5}$ pour l'incertitude et $\pm 2\sigma$ pour les autres stratégies).

majoritairement autour du motif à découvrir et aucune région de l'espace d'entrée $\mathbb{X}$ n'est délaissée. Cette stratégie gère correctement le compromis entre l'exploitation des exemples étiquetés et l'exploration de l'espace $\mathbb{X}$.

L'échantillonnage par incertitude donne les pires résultats et la variance de l'AUC est très importante[3], quelque soit la valeur de $|L|$. La partie droite de la Figure 3.14 montre les exemples sélectionnés par cette stratégie lors d'une expérience. Les exemples étiquetés sont regroupés dans une zone correspondant aux emplacements successifs de la séparatrice linéaire apprise par la régression logistique. Ce modèle prédictif évolue peu durant l'échantillonnage sélectif, puisque la stratégie basée sur l'incertitude sélectionne les exemples les plus proches de la frontière de décision. Les mauvaises performances de l'échantillonnage par incertitude s'expliquent intuitivement par une trop forte exploitation, au détriment de l'exploration de l'espace $\mathbb{X}$.

La Figure 3.15 présente un tableau résumant les valeurs du déficit (Annexe A.c) pour chaque stratégie ; le déficit étant calculé pour $|L| \in [0, 205]$. Plus le déficit est faible, plus la stratégie active est performante par rapport à l'échantillonnage aléatoire. Notre stratégie basée sur les modèles locaux est la plus performante, quelque soit la valeur du paramètre α.

[3]Pour des questions de lisibilité, les moustaches sur la courbe de l'échantillonnage par incertitude représentent $\pm\frac{\sigma}{5}$, tandis que ces moustaches représentent $\pm 2\sigma$ sur les autres courbes.

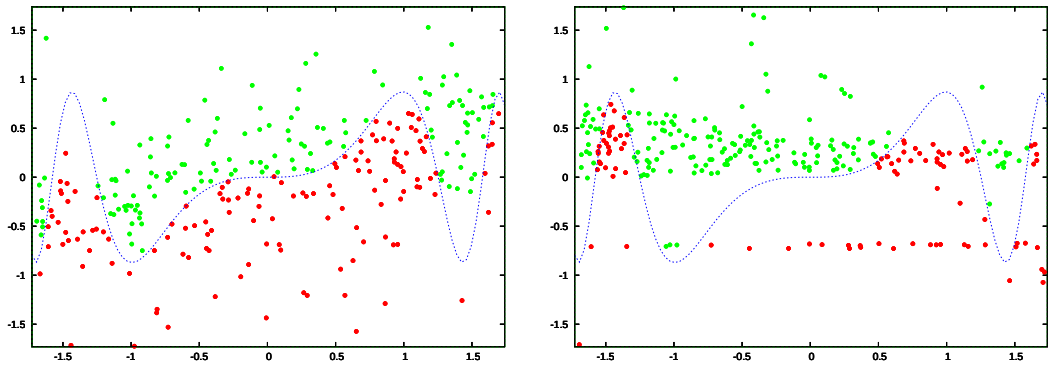

FIG. 3.14 – Exemples sélectionnés grâce à l'échantillonnage par réduction de l'erreur de généralisation (à gauche) et grâce à l'échantillonnage par incertitude (à droite). Les 250 exemples étiquetés sont représentés sur cette figure (○ classe "1", et ● classe "0")

	$\|L\| \in [0, 250]$
Modèles locaux, $\alpha = 0.25$	0.64
Modèles locaux, $\alpha = 1$	1.59
Réduction de l'erreur	1.98
Incertitude	4.52

FIG. 3.15 – Calcul du déficit des différentes stratégies actives, entre 0 et 250 exemples étiquetés. Les stratégies d'apprentissage actif sont classées de la plus performante à la moins performante.

3.3.4 Discussion

L'apprentissage actif par modèles locaux a été amélioré dans cette section grâce à un nouveau critère de sélection de zones. Ce critère ajuste le compromis entre l'exploitation et l'exploration des données, par le biais d'un paramètre α. Lors des itérations successives de l'échantillonnage sélectif, notre critère adopte un comportement dynamique (Section 3.3.3). La notion de progrès réalisé par les modèles locaux est naturellement prise en compte.

Notre stratégie d'apprentissage actif est indépendante du modèle prédictif. La régression logistique utilisée pour les expériences des Sections 3.2 et 3.3 aurait pu être remplacée par n'importe quel autre classifieur. Notre stratégie est donc applicable à de nombreux problèmes réels et peut facilement être exploitée sur des systèmes existants.

Les expériences réalisées à la Section 3.3.3 montrent l'existence d'un paramétrage optimal pour lequel notre critère est très performant. Des travaux futurs pourraient permettre à notre stratégie de trouver automatiquement la valeur optimale du paramètre α. Deux pistes sont envisageables : i) chercher la valeur optimale de α pour l'ensemble de l'échantillonnage sélectif ; ii) chercher la valeur optimale de α pour chaque itération de l'échantillonnage sélectif.

Nous citerons à titre d'exemple les travaux de T. Osugi [Osugi *et al.*, 2005b] qui uti-

lise deux stratégies actives respectivement dédiées à l'exploitation et à l'exploration des données. À chaque itération de l'échantillonnage sélectif, les probabilités d'exploiter ou d'explorer les données sont mises à jour. Une métrique mesurant l'évolution de l'hypothèse apprise par le modèle prédictif entre deux itérations est définie. Si l'hypothèse apprise change beaucoup entre deux itérations [*respectivement* change peu], la probabilité d'explorer les données augmente [*respectivement* diminue].

3.4 Applications : Détection d'émotions dans la parole

Dans cette section, notre stratégie d'apprentissage actif par modèles locaux est appliquée à un problème industriel : la détection d'émotions dans la parole. Ne disposant pas d'une méthode permettant de régler automatiquement la valeur du paramètre α, nous choisissons de fixer ce paramètre à 0.5. Ce choix semble être un a priori minimal.

3.4.1 Domaine d'application

Les serveurs vocaux interactifs, aussi appelés SVI, sont des machines capables de dialoguer avec des utilisateurs afin d'exécuter diverses tâches. Les premiers SVI sont apparus en 1985 et proposaient aux utilisateurs une liste de choix numérotés. Les interactions entre les utilisateurs et la machine s'effectuaient à l'époque via le clavier téléphonique. Aujourd'hui, une nouvelle génération de SVI exploitant des techniques de traitement automatique du langage naturel [Christopher and Schütze, 1999] fait sont apparition. L'interactivité de ces serveurs vocaux est beaucoup importante qu'auparavant. Les utilisateurs ont la possibilité de répondre à des questions ouvertes, où les réponses ne sont pas prédéfinies. Des techniques de reconnaissance [Junqua and Halton, 1995] et de synthèse vocale [Richard and Cappé, 2004] sont exploitées et permettent aux utilisateurs d'interagir avec la machine via la parole.

Les entreprises qui déploient ce type de serveurs vocaux cherchent à améliorer la satisfaction des utilisateurs, notamment en les redirigeant en cas de difficulté vers un opérateur humain. Un problème de dialogue est supposé générer un état émotionnel particulier chez l'utilisateur, qui est exprimé dans son dialogue avec la machine. L'aiguillage des utilisateurs insatisfaits vers un opérateur revient à détecter les émotions négatives exprimées dans leurs dialogues avec la machine. Ce problème d'apprentissage est généralement posé comme une classification binaire supervisée. Les exemples d'apprentissage sont des tours de parole[4] étiquetés comme exprimant, ou non, une émotion négative. La prise en compte d'un plus grand nombre de classes indiquant le niveau d'émotion détecté pose le problème de la subjectivité de l'étiquetage [Liscombe *et al.*, 2005]. Lors de la préparation des données d'apprentissage, un expert écoute, analyse et étiquette des tours de paroles issus de dialogues entre les utilisateurs et la machine. Pour ce problème applicatif, l'étiquetage des exemples est particulièrement coûteux. Ce coût peu être réduit grâce aux stratégies d'apprentissage actif, qui proposent à l'expert d'étiqueter uniquement les exemples les plus

[4]On entend par "tour de parole" un échange vocal avec la machine. Typiquement, un tour de parole est une réponse à une question posée par le serveur vocal interactif.

utiles à l'apprentissage du modèle prédictif.

Cette section constitue un cas d'étude appliquant des stratégies d'apprentissage actif à la détection d'émotions dans la parole. La finalité de notre étude est de déployer un système de redirection d'appels sur un serveur vocal interactif.

3.4.2 Conditions expérimentales

Cette section présente les conditions expérimentales de notre étude sur la détection d'émotions exprimées dans la parole.

Caractérisation des données :
Notre étude se base sur des travaux antérieurs [Poulain, 2006] dont l'objectif est de caractériser au mieux les tours de parole pour la classification d'émotions. B. Poulain définit, grâce à une méthode de sélection d'attributs, le sous-ensemble des variables explicatives les plus pertinentes pour l'apprentissage du modèle prédictif.

Les données utilisées sont issues d'une expérience mettant en jeu 32 utilisateurs qui testent un service boursier implémenté sur un serveur vocal interactif. Du point de vue des utilisateurs, l'expérience consiste à gérer un portefeuille fictif d'actions ; l'objectif étant de réaliser la plus forte plus value. Les traces vocales enregistrées lors de l'utilisation du service boursier par les utilisateurs constituent les données utilisées pour cette étude ; soit 5496 tours de parole. Les tours de parole sont caractérisés par 200 variables acoustiques, décrivant notamment la variation du volume sonore, la variation de la hauteur de voix, le rythme d'élocution. Les données sont également caractérisées par 8 variables dialogiques décrivant notamment le sexe du locuteur, le rang du tour de parole dans le dialogue, la durée du dialogue. Chaque tour de parole est étiqueté par un expert comme exprimant, ou non, une émotion négative.

Le sous-ensemble des variables les plus informatives pour la détection d'émotions dans la parole est défini grâce à un prédicteur Bayésien naïf sélectif [Boullé, 2006a]. Au début de ce procédé, le sous-ensemble de variables est vide. À chaque itération, la variable qui améliore le plus la qualité du modèle prédictif est ajoutée. Ce procédé s'arrête lorsque l'ajout de nouvelles variables n'améliore plus le modèle prédictif. Finalement, 20 variables sont sélectionnées pour caractériser les tours de paroles[5]. Les données utilisées pour notre étude sont issues de cette expérience et la sélection de variables réalisée dans [Poulain, 2006] est prise en considération. Les tours de parole sont caractérisés par les 20 variables suivantes :

1. Arrêt du système (l'utilisateur met fin au dialogue)
2. Nombre de mots dans le tour de parole courant
3. L'utilisateur commente le dialogue
4. Nombre d'erreurs de la tâche courante

[5]Parmi les 20 variables sélectionnées, certaines sont obtenues de manière non automatique. On suppose dans la suite de cette section qu'on dispose d'un moyen pour les évaluer toutes.

5. Nombre total d'erreurs des tâches imbriquées
6. Augmentation de l'intensité du signal
7. Diminution de l'intensité du signal
8. Coefficient maximal de la première harmonique du signal
9. Moyenne de la distribution des variations du timbre de la voix
10. Valeur maximale de la variance standard du timbre de la voix
11. Variance standard du timbre de la voix
12. Moyenne de la distribution du ratio hautes fréquences / basses fréquences
13. Variance standard de l'énergie du signal
14. Somme des variances standard de l'énergie du signal
15. Valeur maximum de la variance standard de l'énergie du signal
16. Dérivée de l'énergie du signal
17. Jitter de l'énergie du signal
18. Reformulation complète du tour de parole précédent
19. Répétition complète du tour de parole précédent
20. Répétition partielle du tour de parole précédent

Choix du modèle prédictif :
La régression logistique utilisée à la Section 3.2.1 est abandonnée au profit d'un modèle qui semble plus adapté à la classification d'émotions exprimées dans la parole. Les méthodes à noyaux ainsi que les méthodes à plus proches voisins sont couramment utilisées pour le traitement de la parole [Guide *et al.*, 2003]. Aussi, nous choisissons d'employer une fenêtre de Parzen à noyau gaussien [Parzen, 1962] lors de nos expériences. Ce modèle génératif est approprié à l'échantillonnage sélectif, puisqu'il est capable d'apprendre avec peu d'exemples étiquetés et qu'il implique un seul paramètre : la variance σ du noyau gaussien. La sortie de ce modèle prédictif estime la probabilité d'observer la classe $y \in \mathbb{Y}$ conditionnellement à l'exemple $x \in \Phi$:

$$\hat{p}(y|x) = \frac{\sum_{n=1}^{|L|} \mathbb{1}_{\{f(l_n)=y\}} K(x, l_n)}{\sum_{n=1}^{|L|} K(x, l_n)} \tag{3.4}$$

avec

$$l_n \in L \ , \ x \in \Phi$$

et

$$K(x, l_n) = e^{\frac{||x-l_n||^2}{2\sigma^2}}$$

La fenêtre de Parzen affecte l'étiquette $\hat{f}(x)$ à l'instance $x \in \Phi$ en se basant sur un seuil de décision, noté $\mathcal{T}h(L)$. Ce seuil minimise l'erreur moyenne de prédiction[6] sur l'ensemble des exemples étiquetés. L'étiquette prédite est :

[6]La mesure d'erreur utilisée pour calculer le seuil $\mathcal{T}h(L)$ est l'AUC, définie à la Section A.a

$$\hat{f}(x) = 1 \quad \text{si} \quad \{\hat{p}(y|x) > \mathcal{T}h(L)\}$$
$$\hat{f}(x) = 0 \quad \text{sinon}$$

La valeur optimale du paramètre du noyau ($\sigma^2 = 0.24$) est déterminée avant l'échantillonnage sélectif, de manière à minimiser l'erreur quadratique moyenne sur l'ensemble du jeu de données [Chappelle, 2005]. Lors de cette étape préliminaire, les 5496 tours de parole étiquetés sont exploités. Le paramétrage de la fenêtre de Parzen étant optimal dès le début de l'échantillonnage sélectif, l'apprentissage du modèle prédictif se réduit au comptage des exemples étiquetés au sens du noyau gaussien. Dans ces conditions, l'évaluation des différentes stratégies d'apprentissage actif n'est pas influencée par l'apprentissage du modèle. Seule la qualité des exemples étiquetés a un impact sur la qualité du modèle prédictif.

Stratégies comparées :
Comme à la Section 3.3.3, les stratégies évaluées lors de notre expérience sont les suivantes :

- l'échantillonnage aléatoire (Section 3.2.2) ;
- l'échantillonnage par incertitude (Section 2.2.1, page 9) ;
- l'échantillonnage par réduction de l'erreur de généralisation (Section 2.2.4, page 16) ;
- l'apprentissage actif par modèles locaux (Section 3.3), notre critère de sélection de zones est paramétré par défaut par $\alpha = 0.5$.

Protocole :
Les stratégies d'apprentissage actif sont évaluées par l'AUC (Annexe A.a), sur un ensemble de test incluant 1613 exemples étiquetés. Une fenêtre de Parzen globale à l'espace $\mathbb{X}$ est utilisée lors de l'évaluation des stratégies. L'ensemble Φ des exemples visibles lors de l'échantillonnage sélectif comporte 3783 exemples. Les expériences réalisées dans cette section sont répétées 10 fois de manière à obtenir une AUC moyenne et la variance de l'AUC pour chaque point des courbes de résultat. Au début de l'échantillonnage sélectif, l'ensemble L ne comporte qu'un seul exemple étiqueté de chaque classe, choisi aléatoirement. À chaque itération, 1 seul exemple est sélectionné, étiqueté et ajouté à l'ensemble L. Le partitionnement d'une zone (étape i de l'Algorithme 4, page 35) est très coûteux en temps de calcul, en raison du nombre important d'exemples de ce jeu de données. Cette étape a une complexité temporelle de $\mathcal{O}(dim(\mathbb{X}) \times |l|^2)$, avec "$l$" l'ensemble des exemples étiquetés d'une zone. Pour remédier à ce problème, le seuil de partitionnement de notre stratégie est fixé à 80 exemples étiquetés par zone.

3.4.3 Résultats et discussion

La Figure 3.16 présente la performance moyenne de chaque stratégie (axe vertical) en fonction du nombre d'exemples étiquetés (axe horizontal). Les moustaches sur les courbes représente la variance de l'AUC ($\pm 2\sigma$).

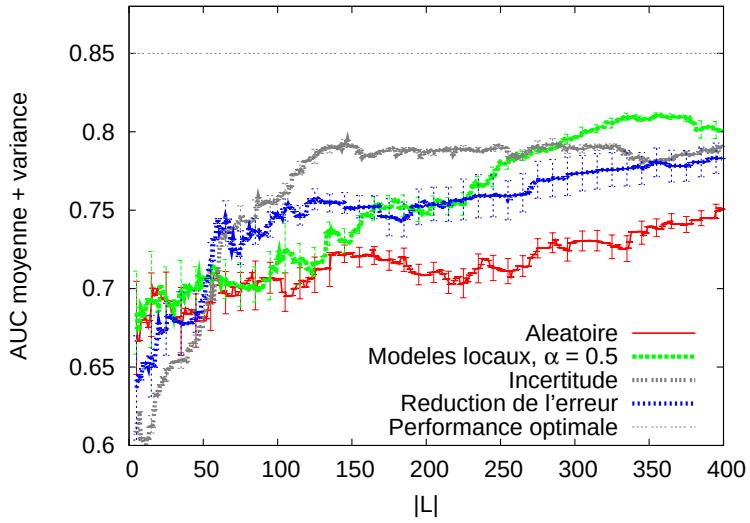

FIG. 3.16 – Évaluation des stratégies actives, dans le cadre de la détection d'émotions exprimées dans la parole. Cette figure représente l'AUC moyenne en fonction du nombre d'exemples étiquetés. Les moustaches sur les courbes représentent la variance de l'AUC observée sur 5 expériences ($\pm 2\sigma$).

Pour être valides les stratégies actives doivent être plus performantes que l'échantillonnage aléatoire, ce qui se vérifie lors de notre expérience. L'échantillonnage aléatoire (courbe rouge) est globalement dominé par les trois autres stratégies. Lorsque $|L| > 50$, la performance la plus faible est toujours réalisée par l'échantillonnage aléatoire. Pour $|L| \in [0, 50]$, seule notre stratégie offre une performance supérieure ou égale à l'échantillonnage aléatoire.

L'échantillonnage par incertitude est plus performant que la réduction de l'erreur de généralisation. Sur la Figure 3.16, la courbe grise domine la bleue quelque soit la valeur de $|L| \in [60, 400]$. La complexité temporelle de l'échantillonnage par incertitude est pourtant beaucoup plus faible que la stratégie basée sur la réduction de l'erreur de généralisation.

Lorsque $|L| < 80$, notre stratégie d'apprentissage actif par modèles locaux (courbe verte) adopte un comportement similaire à l'échantillonnage aléatoire. Cela est dû au fait que le seuil de partitionnement est fixé à 80 exemples étiquetés par zone. Lorsque $|L|$ appartient à l'intervalle $[0, 80]$, le partitionnement récursif de l'espace $\mathbb{X}$ n'est pas encore initié. Ensuite, notre stratégie progresse rapidement lorsque le nombre d'exemples étiquetés augmente. Pour $|L| > 200$, notre stratégie est plus performante que la réduction de l'erreur de généralisation. Et pour $|L| > 280$, notre stratégie est la plus performante. Pour $|L|$ variant de 0 à 400, la performance de notre stratégie croît de manière monotome et est toujours supérieure ou égale à l'échantillonnage aléatoire. Notre stratégie reste la plus performante lorsque le nombre d'exemples étiquetés dépasse 400. L'apprentissage actif par modèles locaux est donc une stratégie compétitive pour la détection d'émotions dans la parole. Les résultats obtenus lors de cette expérience sont très encourageants, même si le paramétrage de notre stratégie est difficile à réaliser en pratique.

3.5 Conclusion

Dans ce chapitre, la curiosité adaptative a été adaptée avec succès à l'échantillonnage sélectif et est utilisée en tant que stratégie active pour résoudre des problèmes de classification. Nous avons amélioré cette stratégie issue de la robotique en proposant un nouveau critère de sélection de zones plus performant et plus simple à mettre œuvre que l'original. Notre stratégie d'apprentissage actif par modèles locaux a été comparée à d'autres stratégies de la littérature, sur un problème jouet (Section 3.3.3) et sur un problème industriel (Section 3.4). Dans les deux cas, notre stratégie offre de très bonnes performances.

Les travaux réalisés dans ce chapitre répondent en partie aux questions soulevées en conclusion du Chapitre 2 (Section 2.4, page 27) :

Compromis exploitation / exploration :
Une stratégie d'apprentissage actif idéale assure un bon compromis entre l'exploitation des exemples étiquetés et l'exploration de l'espace d'entrée $\mathbb{X}$. L'apport principal de ce chapitre est de montrer l'intérêt du partitionnement dichotomique récursif de l'espace d'entré $\mathbb{X}$ pour gérer ce compromis. Notre stratégie active définit les zones de l'espace $\mathbb{X}$ où les modèles locaux progressent le plus, et évite les régions de l'espace où il y a moins à découvrir. Le critère de sélection de zones que nous avons proposé à la Section 3.3 ajuste ce compromis de manière explicite, grâce au paramètre α.

Exploitation des exemples non-étiquetés :
Les exemples non-étiquetés sont abondants lors d'un échantillonnage sélectif et ces exemples sont représentatifs de la distribution des données. Une stratégie d'apprentissage actif idéale exploite les exemples non-étiquetés, soit pour affiner l'apprentissage du modèle prédictif, soit pour améliorer la sélection des exemples à étiqueter. Le critère présenté à la Section 3.3 prend en compte les exemples non-étiquetés grâce à la densité relative, les exemples de l'ensemble U ont une influence sur la sélection des zones d'intérêt. Notre stratégie favorise les régions de l'espace $\mathbb{X}$ densément peuplées. Ainsi, une zone sélectionnée comportant peu d'exemples perd rapidement de son intérêt entre deux itérations ; l'étiquetage d'un nouvel exemple dans cette zone provoque une forte augmentation de la densité relative.

Implication d'un minimum de paramètres :
Disposant de peu d'exemples étiquetés, il est difficile en pratique d'ajuster des paramètres avant un échantillonnage sélectif. Le principal inconvénient de la curiosité adaptative, utilisée en tant que stratégie active, est l'implication d'un grand nombre de paramètres. Les travaux réalisés dans ce chapitre constituent un premier pas vers la réduction du nombre de paramètre à ajuster. Notre critère de sélection de zone (Section 3.3) implique un seul paramètre α, qui se substitue à plusieurs paramètres de la curiosité adaptative difficile à ajuster : i) la mesure de performance utilisée pour évaluer les modèles locaux grâce aux exemples étiquetés de chaque zone ; ii) la fenêtre temporelle utilisée pour évaluer les progrès réalisés par les modèles locaux.

Évaluation du biais entre les ensembles L et Φ :
Les stratégies d'apprentissage actif induisent un biais sur la distribution de l'ensemble L. Les exemples étiquetés, sélectionnés par la stratégie active, ne sont pas représentatifs de l'ensemble Φ. Ce biais peut influencer l'apprentissage et l'évaluation du modèle prédictif. Dans le cas de notre méthode d'apprentissage actif par modèles locaux, les exemples sont tirés aléatoirement dans les zones d'intérêt sélectionnées. Les exemples étiquetés sont donc *iid* localement à chacune des zones, ce qui permet aux modèles locaux d'apprendre sur des données représentatives de l'ensemble Φ. Lors de l'évaluation de notre stratégie un biais entre les ensembles L et Φ apparait, puisqu'un modèle global à l'espace $\mathbb{X}$ est employé. Notre stratégie d'apprentissage actif par modèles locaux offre également la possibilité d'être hybridée avec les stratégies actives de l'état de l'art. Cette hybridation permettrait de sélectionner localement à la meilleure zone, l'exemple le plus utile à l'apprentissage du modèle local. Dans ce cas, le biais entre les ensembles L et Φ aurait une influence sur l'apprentissage des modèles locaux.

Les travaux réalisés dans la suite de ce manuscrit ont pour objectif d'améliorer notre stratégie d'apprentissage actif par modèles locaux. L'objectif du Chapitre 4 est de rendre automatiques les décisions relatives au partitionnement d'une zone : i) *quand* couper une zone ; ii) *ou* couper la zone. Nous choisissons d'exploiter une méthode de discrétisation supervisée, basée sur un formalisme Bayésien. La méthode MODL (*Minimal Optimized Description Length*) [Boullé, 2006b] est particulièrement intéressante pour l'apprentissage actif par modèles locaux, puisqu'elle n'est pas assujettie au sur-apprentissage [Boullé, 2007a], qu'elle est très robuste et non-paramétrique. Le principal inconvénient de cette méthode est qu'elle n'exploite pas les exemples non-étiquetés, qui sont pourtant abondants lors d'un échantillonnage sélectif. L'extension de l'approche MODL au cas de l'apprentissage semi-supervisé est réalisé au Chapitre 4 et constitue un des apports majeurs de cette thèse.

Une stratégie d'apprentissage actif originale est proposée dans le Chapitre 5. Cette stratégie exploite les modèles de discrétisation semi-supervisés définis au Chapitre 4. Le cadre théorique de cette étude se limite aux modèles à un ou deux intervalles, et au cas de données unidimensionnelles. Cette stratégie sélectionne les exemples les plus utiles à l'apprentissage d'un modèle de discrétisation qui cherche à couper l'espace $\mathbb{X}$ en deux intervalles. Notre stratégie est plus performante que la dichotomie probabiliste [Castro and Nowak, 2008], dans la mesure où elle n'a pas besoin d'être renseignée du niveau de bruit présent dans les données et qu'elle ne coupe pas l'espace $\mathbb{X}$ si cela n'est pas justifié. Finalement, cette stratégie peut être exploitée dans le cadre d'un apprentissage actif par modèles locaux pour sélectionner dans une zone les exemples les plus utiles à l'apprentissage du modèle local. Notre stratégie peut également être utilisée dans de nombreuses heuristiques de la littérature, qui exploitent habituellement la dichotomie.

Chapitre 4

Discrétisation Bayésienne semi-supervisée

Sommaire

Ce chapitre a fait l'objet d'une publication : [Bondu *et al.*, 2008]

Au vu des résultats obtenus au Chapitre 3, notre objectif est d'améliorer l'apprentissage actif par modèles locaux en rendant automatiques les décisions relatives au partitionnement récursif de l'espace $\mathbb{X}$. L'idée développée dans ce chapitre est d'exploiter une méthode de discrétisation pour décider *où* et *quand* partitionner une zone. Parmi les méthodes de discrétisation de l'état-de-l'art, nous choisissons d'utiliser l'approche MODL (*Minimal Optimized Description Length*) [Boullé, 2006b]. Cette méthode de discrétisation est exploitable pour le partitionnement récursif de l'espace $\mathbb{X}$, mais n'exploite pas les exemples non-étiquetés. De nombreux travaux montrent que les méthodes d'apprentissage semi-supervisées améliorent significativement les performances d'un apprentissage actif [Chappelle, 2005; Zhu *et al.*, 2003; Muslea *et al.*, 2002]. Ce chapitre est consacré à l'extension de l'approche MODL au cas de l'apprentissage semi-supervisé.

Ce chapitre s'organise de la manière suivante. La Section 4.1 présente deux états-de-l'art synthétiques respectivement dédiés aux méthodes de classification semi-supervisée et aux méthodes de discrétisation supervisée. La Section 4.2 définit une nouvelle méthode de discrétisation semi-supervisée qui s'apparente, dans sa démarche, à l'approche MODL. La Section 4.3 présente les résultats théoriques et empiriques obtenus lors d'une étude approfondie de notre méthode de discrétisation semi-supervisée. Cette section montre notamment la convergence asymptotique des approches semi-supervisée et supervisée, lorsque le nombre d'exemples tend vers l'infini, l'approche supervisée étant munie d'un post-traitement sur la position des bornes de la discrétisation optimale. La Section 4.4 applique notre approche à des problèmes jouet et met en évidence les apports de notre méthode de discrétisation semi-supervisée par rapport à l'approche MODL. Enfin, la Section 4.5 conclut le chapitre et présente les perspectives ouvertes par nos travaux.

4.1 État-de-l'art

L'objectif de ce chapitre est l'élaboration d'une méthode de discrétisation semi-supervisée. La section qui suit présente deux états-de-l'art respectivement dédiés aux méthodes de classification semi-supervisée et aux méthodes de discrétisation.

4.1.1 Apprentissage semi-supervisé

L'apprentissage semi-supervisé manipule conjointement des données étiquetées et non-étiquetées, l'objectif étant d'obtenir un modèle prédictif plus performant qu'en utilisant seulement les exemples étiquetés ou non-étiquetés. Les méthodes d'apprentissage semi-supervisé sont particulièrement intéressantes lors d'un échantillonnage sélectif, et peuvent être utilisées à chaque itération pour améliorer l'apprentissage du modèle prédictif. Dans ce contexte, les exemples non-étiquetés sont abondants et leur étiquetage est coûteux. Le coût d'étiquetage peut être prohibitif pour plusieurs raisons : l'utilisation d'un instrument de mesure ; un temps de traitement trop élevé ; l'implication d'un expert humain... etc. Lors d'un échantillonnage sélectif, les méthodes d'apprentissage semi-supervisé permettent l'apprentissage d'un modèle prédictif exploitant tous les exemples disponibles.

Cet état-de-l'art s'intéresse uniquement aux méthodes d'apprentissage semi-supervisé résolvant des problèmes de classification. Notons cependant qu'il existe des méthodes de clustering semi-supervisé exploitant les étiquettes lors de constitution des groupes. Deux exemples qui ont des étiquettes différentes ne doivent pas appartenir au même groupe à l'issue de l'apprentissage. Il existe également des approches qui enrichissent les données. Par exemple, des contraintes indiquant si deux exemples doivent ou non appartenir au même groupe peuvent être ajoutées à l'ensemble d'apprentissage, indépendamment des étiquettes [Cohn *et al.*, 2003].

L'intérêt des méthodes de classification semi-supervisée pour l'échantillonnage sélectif étant mis en évidence, cette section présente un état-de-l'art synthétique sur ces méthodes.

L'auto-apprentissage :
Le "self-training", aussi appelé auto-apprentissage, est l'une des premières heuristiques utilisées en apprentissage semi-supervisé [Fralick, 1967] (voir Algorithme 5). Cette approche consiste à utiliser les prédictions d'un modèle pour étiqueter de nouveaux exemples. Tout d'abord, le modèle $\mathcal{M}$ est entraîné grâce aux exemples étiquetés (étape A). Le modèle est ensuite utilisé pour prédire les étiquettes des exemples non-étiquetés (étape B). Une mesure d'incertitude sur les prédictions du modèle est définie. Les exemples pour lesquels le modèle a le plus confiance en ses prédictions sont ajoutés à l'ensemble d'apprentissage (étapes C et D). L'étiquetage des exemples sélectionnés est réalisé grâce aux prédictions du modèle.

Notations :

- $\mathcal{M}$ un modèle prédictif muni d'un algorithme d'apprentissage $\mathcal{L}$
- Les ensembles U et L d'exemples non-étiquetés et étiquetés
- n le nombre d'exemples d'apprentissage souhaités.
- L'ensemble d'apprentissage T constitué de couples (x, y), tel que $x \in L$ et $y \in \mathbb{Y}$ ($|T| < n$).

Répéter
(A) Entraîner le modèle $\mathcal{M}$ grâce à T.
(B) Utiliser $\mathcal{M}$ pour prédire les étiquettes $f(x) \forall x \in U$.
(C) Sélectionner l'instance $q \in U$ dont la prédiction $\hat{f}(q)$ est la plus sûre.
(D) Ajouter $(q, \hat{f}(q))$ à T, ajouter q à L, et retirer q de U.
Tant que $|T| < n$

Algorithme 5: Self-Training

Dans la pratique, le "self-training" présente un défaut majeur. Le modèle étiquette de nouveaux exemples grâce à ses prédictions, sans avoir la garantie que ces étiquettes soient correctes. Les éventuelles erreurs de classification se propagent au cours des itérations de l'Algorithme 5 et peuvent se renforcer. Pour limiter cet effet néfaste, une heuristique consiste à retirer les étiquettes des exemples dont la prédiction chute en dessous d'un certain seuil de confiance. La technique du "self-training" a été appliquée au traitement automatique des langues et notamment à la classification d'émotions dans les dialogues [Maeireizo *et al.*, 2004]. Le "self-training" a l'avantage d'être simple à mettre en œuvre, mais dans le cas général, la convergence de cette approche n'est pas garantie. Selon notre point de vue, les hypothèses sous-jacentes sont les suivantes : i) le modèle prédictif est supposé être assez performant pour étiqueter correctement les exemples sélectionnés ; ii) les données sont supposées être séparables.

Le co-apprentissage :
Le "co-training" repose sur l'idée que les exemples d'apprentissage peuvent être considérés selon plusieurs vues [Blum and Mitchell, 1998]. Un exemple $x \in L$ est un vecteur $\{x_1, x_2 ... x_K\}$ appartenant à l'espace $\mathbb{X}$. Le "co-training" exploite deux sous-ensembles de variables disjoints tel que $\mathbb{X} = X_1 \oplus X_2$. L'exemple $x \in L$ peut être considéré selon deux vues : $x = \{x_1, x_2 ... x_n\} \in X_1$ et $x = \{x_{n+1}, x_{n+2} ... x_K\} \in X_2$. Pour illustrer le "co-training", X. Zhu [Zhu, 2005] emploie l'exemple de la classification de pages Internet. Les variables descriptives sont soit relatives au texte (X_1), soit relatives aux images (X_2). Les deux modèles $\mathcal{M}_1$ et $\mathcal{M}_2$ sont entraînés indépendamment sur les deux sous-ensembles de variables X_1 et X_2 (étape A de l'Algorithme 6). Seuls les exemples étiquetés sont utilisés lors de cette étape. Les deux modèles prédisent les étiquettes des exemples non-étiquetés (étape B). Comme pour le "self-training", chaque modèle sélectionne l'exemple dont la prédiction est la plus sûre et lui attribue une étiquette (étape C). Un échange entre les deux vues est réalisé lors de l'étiquetage des nouveaux exemples. Chaque modèle ajoute

un exemple étiqueté à l'ensemble d'apprentissage de son homologue (étape D).

Notations :

- $X_1 \oplus X_2 = \mathbb{X}$ deux ensembles de variables disjoints ;
- $\mathcal{M}_1$ et $\mathcal{M}_2$ deux modèles prédictifs ;
- Les ensembles U et L d'exemples non étiquetés et étiquetés ;
- n le nombre d'exemples d'apprentissage souhaité ;
- Les ensembles d'apprentissage des deux modèles T_1 et T_2 composés de couples (x, y), tel que $x \in L$ est défini soit sur X_1 soit sur X_2 et $y \in \mathbb{Y}$ ($|T_1 \cup T_2| < n$).

Répéter

(A) Entraîner les modèle $\mathcal{M}_1$ et $\mathcal{M}_2$ grâce à T_1 et T_2.
(B) Utiliser $\mathcal{M}_1$ et $\mathcal{M}_2$ pour prédire les étiquettes $f(x) \forall x \in U$.
(C) Sélectionner l'instance $q_1 \in U$ [*respectivement* q_2] dont la prédiction par le modèle $\mathcal{M}_1$ [*respectivement* $\mathcal{M}_2$] est la plus sûre.
(D) Ajouter $(q_1, \hat{f}_1(q))$ à T_2 et $(q_2, \hat{f}_2(q))$ à T_1. Ajouter q_1 et q_2 à L, et retirer q_1 et q_2 de U.

Tant que $|T_1 \cup T_2| < n$

Algorithme 6: Co-Training

Le "co-training" fait l'hypothèse que l'espace $\mathbb{X}$ peut être séparé en deux sous-ensembles de variables disjoints, sur lesquels l'apprentissage des modèles prédictifs est possible. Les modèles prédictifs $\mathcal{M}_1$ et $\mathcal{M}_2$ sont supposés être suffisamment performants pour que l'étiquetage des exemples sélectionnés soit correct. Il existe de nombreuses variantes du co-training dans la littérature [Zhu, 2005]. Par exemple, Nigam et Ghani [Nigam and Rayid, 2000] proposent le "Co-EM" qui probabilise entièrement l'ensemble U, en utilisant les prédictions des modèles. Les exemples d'apprentissage sont pondérés par la probabilité qu'ils soient bien étiquetés. La convergence du "co-training" n'est pas démontrée dans le cas général, il s'agit avant tout d'une heuristique.

La pondération covariative :
Le "covariate shift" [Sugiyama *et al.*, 2007; Sugiyama and Müller, 2005] traite le cas où les exemples d'apprentissage et les exemples de test n'ont pas la même densité de probabilité. L'objectif est d'ajuster les paramètres d'un modèle prédictif. Le vecteur de réels $\theta \in \Theta$ représente ces paramètres et Θ est l'ensemble des paramétrages possibles. L'erreur de généralisation est estimée grâce à l'erreur empirique, calculée sur l'ensemble d'apprentissage T. Étant donnée une fonction de coût, $\mathcal{L}oss : \Theta \times \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{R}$, le meilleur paramétrage, noté θ^*, minimise l'erreur empirique :

$$\theta^* = \underset{\theta \in \Theta}{ArgMin} \left[\frac{1}{|T|} \sum_{i=1}^{|T|} \mathcal{L}oss(\theta, x_i, y_i) \right] \quad avec\, x_i \in \mathbb{X}\,,\; y_i \in \mathbb{Y}\, et\, (x_i, y_i) \in T \tag{4.1}$$

Lorsque les distributions des exemples d'apprentissage et de test sont différentes, l'estimation de l'erreur de généralisation par l'erreur empirique est biaisée. L'optimisation de l'Équation 4.1 n'aboutit pas nécessairement aux paramètres optimaux, même si un grand nombre d'exemples étiquetés est disponible ($|T| \rightarrow +\infty$). L'approche covariative corrige ce biais en pondérant les exemples d'apprentissage lors de l'estimation de l'erreur de généralisation. Tout d'abord, un estimateur de densité est utilisé pour évaluer P_{test} [*respectivement* P_T], la distribution de l'ensemble de test [*respectivement* d'apprentissage]. L'importance d'un exemple $(x, y) \in T$ est donnée par le ratio $\frac{P_{test}(x)}{P_T(x)}$. Ainsi, les exemples d'apprentissage sous-représentatifs de l'ensemble de test se voient attribuer un coefficient élevé. M. Sugiyama [Sugiyama *et al.*, 2007] propose ainsi une estimation non-biaisée de l'erreur de généralisation telle que :

$$\theta^* = \underset{\theta \in \Theta}{ArgMin} \left[\frac{1}{|T|} \sum_{i=1}^{|T|} \frac{P_{test}(x_i)}{P_T(x_i)} \times \mathcal{L}oss(\theta, x_i, y_i) \right] avec x_i \in \mathbb{X}, y_i \in \mathbb{Y} et (x_i, y_i) \in T \tag{4.2}$$

Le biais existant entre les distributions des exemples étiquetés et non-étiquetés est un problème récurrent en échantillonnage sélectif. Notons qu'il serait envisageable d'appliquer la pondération covariative dans ce cadre, et de corriger ainsi le biais entre les distributions des ensembles L et U. L'estimation de la densité des exemples étiquetés reste néanmoins une étape critique, en raison du faible nombre d'exemples étiquetés disponibles.

L'approche par modèle génératif
Un modèle génératif cherche à estimer la distribution jointe des exemples et des classes, $P(x, y) = P(y)P(x|y)$ avec $x \in \mathbb{X}$ et $y \in \mathbb{Y}$. La distribution estimée est supposée appartenir à une famille $\{P(x, y)_\theta\}$. Le vecteur $\theta \in \Theta$ correspond aux paramètres de modélisation de $P(x, y)$, et Θ est l'ensemble des paramétrages possibles. La décomposition suivante est adoptée : $P(x, y)_\theta = P(y)_\theta P(x|y)_\theta$. La distribution $P(y)_\theta$ est définie a priori, selon notre connaissance de la répartition des classes et éventuellement grâce à θ. $P(x|y)_\theta$ est identifiable au sein d'une famille de distributions, par exemple un mélange de gaussiennes.

Dans ce cas, $P(x|y)_\theta$ est défini par le vecteur θ qui inclut la moyenne et la variance de chaque gaussienne.

Lors d'un apprentissage supervisé, les approches génératives cherchent les paramètres $\theta \in \Theta$ qui maximisent $P(x,y)_\theta$ sur les exemples étiquetés. Le maximum de vraisemblance (MLE) est généralement utilisé pour maximiser $p(T)_\theta$:

$$\underset{\theta\in\Theta}{ArgMax} \log\left[p(T)_\theta\right] =$$

$$\underset{\theta\in\Theta}{ArgMax} \sum_{i=1}^{|T|} \log\left[p(y_i)_\theta p(x_i|y_i)_\theta\right] \qquad (x_i,y_i)\in T \tag{4.3}$$

Les approches semi-supervisées considèrent également les exemples non-étiquetés de l'ensemble U. Dans ce cas, la quantité à maximiser est $p(T,U)_\theta = p(T)_\theta p(U|T)_\theta$. La mesure de vraisemblance s'écrit de la manière suivante, avec $(x_i,y_i)\in T \quad x_{i'}\in U$:

$$\underset{\theta\in\Theta}{ArgMax} \left[\sum_{i=1}^{|T|} \log\left[p(y_i)_\theta p(x_i|y_i)_\theta\right] + \sum_{i'=1}^{|U|} \log\left[\underset{y_{j'}\in\mathbb{Y}}{\mathrm{E}_{P(.)_\theta}}\, p(x_{i'},y_{j'})\right]\right]$$

$$\underset{\theta\in\Theta}{ArgMax} \left[\sum_{i=1}^{|T|} \log\left[p(y_i)_\theta p(x_i|y_i)_\theta\right] + \sum_{i'=1}^{|U|} \log\left[\sum_{j'=1}^{|\mathbb{Y}|} p(y_{j'})_\theta p(x_{i'}|y_{j'})_\theta\right]\right] \tag{4.4}$$

Illustration : X. Zhu illustre les approches génératives grâce à un problème de classification binaire à deux dimensions [Zhu, 2005]. Un a priori uniforme est adopté sur la distribution des classes, et la distribution conditionnelle aux classes est supposée être gaussienne :

$$p(y)_\theta = \frac{1}{|\mathbb{Y}|} \forall y \in \mathbb{Y}$$

$$p(x|y)_\theta = \frac{1}{\sigma_y\sqrt{2\pi}} \times e^{\frac{-(x-\mu_y)^2}{2\sigma_y^2}}$$

Les paramètres de modélisation sont les suivants : $\theta = \{\mu_1,\mu_2,\sigma_1,\sigma_2\}$, avec μ_1 et σ_1 [*respectivement* μ_2 et σ_2] la moyenne et la variance de la distribution $P(x|y_1)$ [*respectivement* $P(x|y_2)$].

L'approche supervisée (Équation 4.3) et l'approche semi-supervisée (Équation 4.4) sont comparées grâce aux données présentées par la Figure 4.1. La partie gauche de la Figure 4.1 représente les exemples étiquetés par des points dans un espace à deux dimensions, $\mathbb{X} = \mathbb{R}^2$. Les exemples de la classe "1" sont symbolisés par des cercles et les exemples de la classe "2" par des croix. La partie droite de la Figure 4.1 représente également les exemples non-étiquetés.

$P(x,y)_\theta$ est estimée de deux façons, selon l'approche supervisée (Équation 4.3) et selon l'approche semi-supervisée (Équation 4.4). La Figure 4.2 représente les lignes de niveaux de $P(x,y)_\theta$ dans les deux cas. La frontière de décision apparaît également sur ces figures et

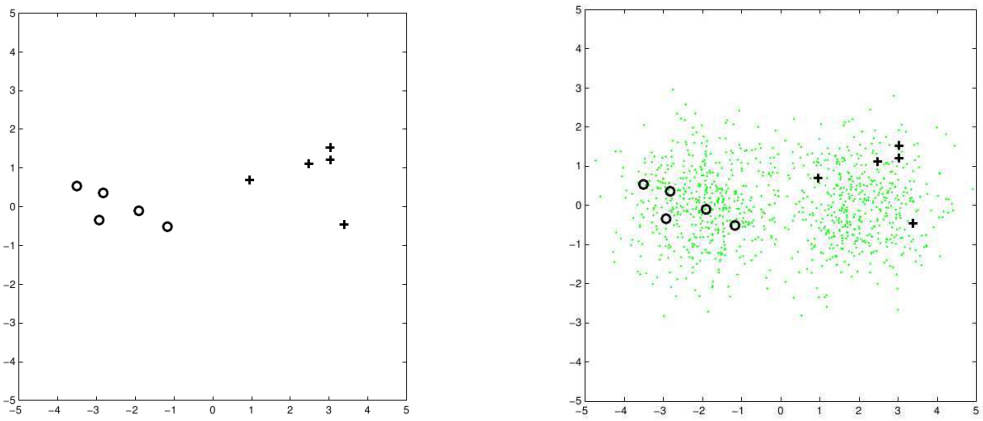

FIG. 4.1 – Exemples étiquetés (à gauche) et non-étiquetés (à droite).

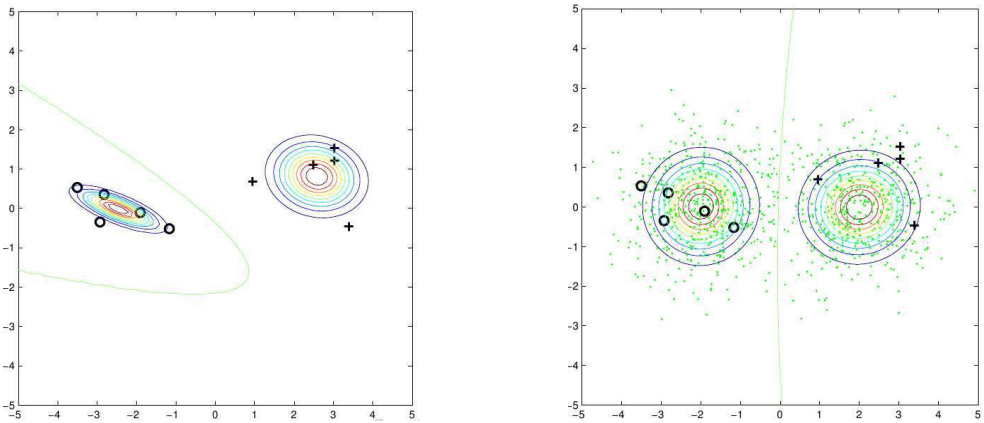

FIG. 4.2 – Apprentissage supervisé (à gauche) et semi-supervisé (à droite), en utilisant un modèle génératif.

est symbolisée par un trait vert continu. Dans le cas de ce problème jouet, l'utilisation de l'ensemble U lors de l'apprentissage du modèle prédictif améliore l'estimation de $P(x, y)$. L'approche semi-supervisée offre de bons résultats car les données sont conformes aux hypothèses de modélisation. En l'occurrence, la distribution conditionnelle des classes est corrélée à la distribution des exemples (étiquetés et non-étiquetés) et est une gaussienne. Dans le cas où les hypothèses de modélisation sont erronées, l'utilisation des exemples non-étiquetés n'améliore pas forcément la qualité du modèle prédictif, et peut même la dégrader.

4.1.2 Méthodes de discrétisation

Cette section présente succinctement les principales familles de méthodes de discrétisation. Cet état-de-l'art est inspiré des travaux de M. Boullé [Boullé, 2007b] et se place du point de vue de la classification supervisée. Les méthodes de discrétisation exploitent un échantillon d'exemples étiquetés, noté T, et estiment $P(y|x)$ la distribution des classes $y \in \mathbb{Y}$ conditionnellement à une variable explicative x.

Tout d'abord, nous introduisons la discrétisation d'une variable numérique grâce à un exemple illustratif. Le jeu de données Iris [D.J. Newman and Merz, 1998] est composé de 150 exemples représentant des fleurs de la famille des Iris. Chaque exemple est caractérisé par quatre variables numériques correspondant à la largeur et la longueur des pétales et sépales. La variable à expliquer peut prendre trois valeurs : $\mathbb{Y} = \{Versicolor, Virginica, Setosa\}$. Dans le cadre de ce problème illustratif, une seule variable explicative est considérée : la largeur de sépale. Le but est de déterminer les corrélations entre cette variable et la classe à prédire. La Figure 4.3 reporte pour chaque valeur de largeur de sépale le nombre d'exemples de l'échantillon T par variété d'iris. Par exemple, parmi les 26 iris dont la largeur de sépale est égale à 3.0 cm, 12 sont de la variété Virginica, 8 de la variété Versicolor et 6 de la variété Setosa.

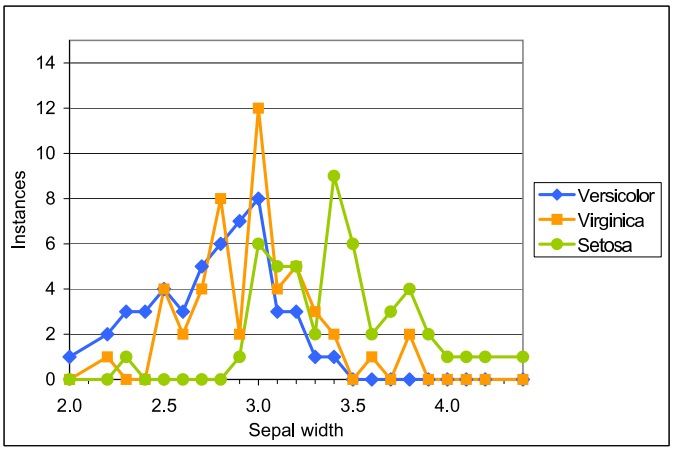

FIG. 4.3 – Nombre d'exemples par variété d'Iris pour chaque valeur de largeur de sépale.

L'objectif de ce problème d'apprentissage est de prédire la probabilité qu'une fleur

appartienne à une classe, connaissant la largeur de ses sépales. La variable explicative est discrétisée en une série d'intervalles et la distribution $P(y|x)$ est estimée empiriquement dans chaque intervalle, grâce au comptage des exemples de chaque classe. Quels sont les intervalles qui conservent au mieux l'information prédictive de la largeur des sépales ?

Une décision prudente suppose que la largeur de sépale n'est pas une variable informative, les variations observées sur la Figure 4.3 étant probablement dues au hasard. Dans ce cas, la variable explicative est discrétisée en un unique intervalle. La prédiction réalisée correspond à la répartition moyenne des classes, soit $\hat{P}(y|x) = \frac{1}{3}$ quelque soit la classe $y \in \mathbb{Y}$ et quelque soit l'exemple $x \in \mathbb{X}$. À l'inverse, une décision risquée suppose que les valeurs observées sur la Figure 4.3 sont toutes significatives. La variable explicative est alors discrétisée en autant d'intervalles qu'il y a de valeurs de largeur de sépale distinctes. Par exemple, la probabilité qu'une fleur soit de la classe "$Virginia$" lorsque la largeur de ses sépales est comprise entre 2.9 et 3.1 cm est de $\frac{12}{26}$. Cette prédiction est très précise mais peu robuste, étant donné le faible nombre d'exemples étiquetés ($|T| = 150$).

Les méthodes de discrétisation ont pour objectif de définir les intervalles qui conservent au mieux l'information prédictive d'une variable explicative. Ces méthodes doivent assurer un bon compromis entre la qualité informationnelle (en définissant des intervalles homogènes vis-à-vis de la variable à prédire), et la qualité statistique (en définissant des intervalles dont les effectifs sont suffisants pour assurer une bonne généralisation).

Les méthodes de discrétisation sont largement utilisées dans le domaine de l'apprentissage artificiel. Ces méthodes permettent d'exploiter des algorithmes d'apprentissage qui traitent uniquement des variables catégorielles, lorsque les exemples de l'ensemble T sont définis par des variables numériques. Par exemple, les arbres de décisions discrétisent les variables explicatives numériques avant de sélectionner la variable de décision à chaque nœud de l'arbre. Un classifieur Bayésien naïf peut exploiter des variables discrétisées, pour estimer les probabilités conditionnelles.

Les méthodes de discrétisation sont généralement définies par trois éléments : i) un critère d'évaluation, qui mesure la qualité d'une discrétisation ; ii) un algorithme d'optimisation, qui cherche une discrétisation performante ; iii) un critère d'arrêt qui stoppe l'algorithme d'optimisation. La discrétisation de variables numériques est abondamment traitée dans la littérature. Quelques méthodes représentatives de l'état-de-l'art sont présentées ici (les algorithmes d'optimisation employés ne sont pas présentés).

Les méthodes non-supervisées :
La plupart des méthodes de discrétisation de la littérature sont supervisées, c'est-à-dire qu'elles exploitent la variable explicative et la variable à expliquer pour évaluer une discrétisation. Les méthodes de discrétisation non-supervisées, comme par exemple *"EqualWidth"* et *"EqualFrequency"* [Sturges, 1926; Scott, 1979], n'exploitent que la variable explicative. La méthode "EqualWidth" divise le domaine numérique d'une variable explicative en intervalles de largeur égale. La méthode "EqualFrequency" divise l'ensemble des exemples étiquetés en intervalles d'effectif constant. Pour ces deux méthodes, le nombre d'intervalles considérés lors de la discrétisation est un paramètre à ajuster.

Les méthodes basées sur le taux d'erreur :
Ces méthodes de discrétisation utilisent un critère d'évaluation basé sur le taux d'erreur observé lors des prédictions des étiquettes des éléments de l'ensemble T. L'objectif est de rechercher la discrétisation qui minimise ce taux d'erreur. Ces approches utilisent généralement deux paramètres pour contrôler le sur-apprentissage : i) le nombre maximum d'intervalles ; ii) l'effectif minimum par intervalle. L'ajustement de ces deux paramètres fait l'objet de nombreux travaux [Holte, 1993; Maass, 1994]. Les méthodes de discrétisation basées sur la minimisation du taux d'erreur ne sont pas adaptées à la modélisation probabiliste. Le taux d'erreur est uniquement basé sur les étiquettes des exemples de l'ensemble T, c'est-à-dire les valeurs majoritaires de la variable à expliquer. L'ensemble de la distribution des valeurs à expliquer est ignoré, ce qui limite les performances prédictives de ces approches [Kohavi and Sahami, 1996].

Les méthodes basées sur le critère du χ^2 :
De nombreuses méthodes de discrétisation sont basées sur le test du χ^2. Ce critère est souvent utilisé comme mesure de dépendance des distributions estimées entre deux intervalles. Deux intervalles adjacents présentant des différences de distribution statistiquement significatives sont séparés, ils sont fusionnés dans le cas contraire. De cette façon, le χ^2 est exploité dans [Kerber, 1991] en tant que critère de discrétisation binaire. Dans [Lechevalier, 1990], une version normalisée du χ^2 est optimisée sur l'ensemble de tous les intervalles, ce qui permet des comparaisons équitables d'un point de vue numérique entre des discrétisations comportant un nombre d'intervalles différent. Dans [Boullé, 2004], le critère d'optimisation évaluant les discrétisations est le niveau de confiance associé au test du χ^2, ce qui permet des comparaisons équitables d'un point de vue statistique.

Les méthodes basées sur l'entropie :
La théorie de l'information de Shannon [Shannon, 1948] est à la base de nombreuses méthodes de discrétisation [Catlett, 1991; Kononenko *et al.*, 1984; Zighed *et al.*, 1998; Fayyad and Irani, 1992]. L'entropie conditionnelle est utilisée pour évaluer la pureté des intervalles d'une discrétisation, vis-à-vis des valeurs de la variable à expliquer. L'optimisation de ce critère sur l'ensemble des intervalles conduit à une discrétisation qui comporte autant d'intervalles que l'ensemble T comporte d'exemples étiquetés. Cette discrétisation est optimale pour l'entropie, mais est très mauvaise en généralisation. Il existe plusieurs possibilités pour remédier à ce problème. Dans les arbres de décisions ID3 [Quinlan, 1986] et C4.5 [Quinlan, 1993], l'entropie est utilisée comme critère d'optimisation lors d'une discrétisation binaire. Le sur-apprentissage est évité dans ce cas en limitant le nombre d'intervalles à deux. L'entropie est difficile à évaluer de façon fiable. En pratique, les probabilités empiriques obtenues par le comptage des exemples étiquetés dans les intervalles sont exploitées.

L'approche MODL *(Minimal Optimized Description Length)* :
Parmi les méthodes de l'état-de-l'art, nous choisissons d'utiliser l'approche supervisée MODL pour améliorer notre stratégie d'apprentissage actif par modèles locaux (Chapitre 3). Cette approche [Boullé, 2006b] transpose le problème de la discrétisation d'une variable numérique en un problème de sélection de modèles. Une famille de modèles de discrétisation basée sur la statistique d'ordre est définie. La probabilité de la classe $y \in \mathbb{Y}$ est estimée conditionnellement au rang de l'exemple $x \in \mathbb{X}$, et non pas grâce à la valeur de la variable explicative caractérisant x. L'estimation de $P(y|x)$ est donc invariante par toute transformation monotone de la variable explicative et est peu sensible aux données atypiques (outliers). Un modèle de discrétisation $M(I, \{N_i\}, \{N_{ij}\})$ est défini par les paramètres suivants : i) I est le nombre d'intervalles ; ii) $\{N_i\}$ est le nombre d'exemples dans chaque intervalle ; iii) $\{N_{ij}\}$ est le nombre d'exemples de chaque classe dans chaque intervalle. Les paramètres $\{N_i\}$ spécifient les bornes des intervalles grâce aux rangs des valeurs explicatives. Et les paramètres $\{N_{ij}\}$ caractérisent la distribution $P(y|x)$ grâce aux effectifs de chaque valeur à expliquer localement à l'intervalle i. Une démarche Bayésienne est appliquée pour sélectionner le meilleur modèle de discrétisation, noté $\mathcal{M}_{map}$ *(Maximum a posteriori)*. Le meilleur modèle de discrétisation maximise $P(M|T)$, la probabilité du modèle M connaissant les données T. En exploitant la formule de Bayes et en considérant que le terme $P(T)$ est constant quelque soit le modèle, cela revient à maximiser $P(M)P(T|M)$. La distribution a priori des modèles $P(M)$ et la vraisemblance des données $P(T|M)$ sont calculées analytiquement en exploitant le caractère discret de la famille de modèles. La démarche Bayésienne employée adopte des hypothèses faiblement informatives sur les données (détaillées à la Section 4.2.1), cette démarche est dite objective. Finalement, le $\mathcal{M}_{map}$ minimise le critère suivant :

$$\mathcal{C} = \underbrace{\log(N) + \log C_{N+I-1}^{I-1} + \sum_{i=1}^{I} \log C_{N_i+J-1}^{J-1}}_{-\log P(M)} + \underbrace{\sum_{i=1}^{I} \log \left(\frac{N_i!}{\prod_{j=1}^{J} N_{ij}!} \right)}_{-\log P(T|M)} \tag{4.5}$$

Le critère $\mathcal{C}$ utilise le log négatif d'une probabilité, ce qui correspond à une quantité d'information [Shannon, 1948]. Le premier terme correspond au choix du nombre d'intervalles, le deuxième terme au choix des bornes des intervalles. Le troisième terme représente le choix des distributions de la variable à expliquer dans chaque intervalle. Le dernier terme représente la probabilité d'observer les valeurs de la variable à expliquer connaissant le modèle de discrétisation. Seuls les exemples étiquetés sont exploités par le critère $\mathcal{C}$. Pourtant, l'échantillonnage sélectif manipule conjointement des exemples étiquetés et non-étiquetés. Ce chapitre est consacré à l'extension de l'approche MODL au cas de l'apprentissage semi-supervisé. Cette méthode de discrétisation est particulièrement intéressante pour l'apprentissage actif par modèles locaux (Chapitre 3) :

- MODL est une approche non-paramétrique au sens de [Robert, 2006], c'est-à-dire que le nombre de paramètres de modélisation est fini mais croît vers l'infini avec le nombre d'exemples d'apprentissage. Cette méthode de discrétisation peut asymptotiquement approximer n'importe quelle distribution conditionnelle.

- Cette méthode de discrétisation n'est pas sujette au sur-apprentissage.
- MODL n'implique pas de paramètres utilisateur, ce qui permet de décider *ou* et *quand* couper une zone sans avoir à ajuster de paramètres.
- Les étiquettes de l'ensemble T sont prises en compte lors de la discrétisation de l'espace $\mathbb{X}$. Pour un nombre fixé d'exemples, la discrétisation optimale désignée par MODL comporte un seul intervalle [*respectivement* plusieurs intervalles], dans le cas où les étiquettes sont complètement mélangées [*respectivement* forment plusieurs groupes homogènes distincts].
- MODL fait des hypothèses faiblement informatives sur les données. Il s'agit d'une approche Bayésienne *"objective"* [Berger, 2006], qui s'adapte à une grande variété de jeux de données.
- MODL est une méthode robuste, qui estime son erreur de généralisation de manière fiable. Cette approche s'est distinguée à ce sujet lors de challenges internationaux [Boullé, 2007b].

4.2 Une nouvelle méthode de discrétisation semi-supervisée

Cette section présente une nouvelle méthode de discrétisation semi-supervisée qui s'apparente, dans sa démarche, à l'approche MODL. Les choix de modélisation ainsi que les notations employées sont définis à la Section 4.2.1. La distribution a priori des modèles est présentée à la Section 4.2.2. La vraisemblance des données conditionnellement aux modèles est définie à la Section 4.2.3. Notre démarche aboutit à un critère d'évaluation semi-supervisé (Section 4.2.4), dont l'optimisation désigne le modèle le plus probable connaissant les données.

4.2.1 Modélisation

Nos choix de modélisation sont issus des travaux de M. Boullé [Boullé, 2007b]. La seule différence avec l'approche MODL est que, dans notre cas, seule une partie des données est étiquetée.

Famille de modèles
La famille de modèles de discrétisation employée est la même que dans l'approche supervisée. Un modèle de discrétisation est défini par :

- un nombre d'intervalles ;
- une partition de la variable explicative en intervalles, spécifiée sur les rangs des valeurs explicatives ;
- la distribution des valeurs de la variable à expliquer par intervalle, spécifiée par les effectifs de chaque valeur à expliquer localement à l'intervalle.

Critère Bayésien de sélection de modèles
Le critère d'évaluation présenté dans ce chapitre est une expression analytique de la probabilité qu'un modèle explique les données d'apprentissage. La distribution a priori des modèles fait l'objet des mêmes hypothèses que dans l'approche supervisée :

- *Hiérarchie du paramétrage des modèles :* le nombre d'intervalles est compris entre 1 et $|\Phi|$ de façon équiprobable. Pour un nombre d'intervalles donné, toutes les partitions en intervalles des rangs de la variable explicative sont équiprobables. Pour un intervalle donné, toutes les distributions conditionnelles des valeurs de la variable à expliquer sont équiprobables. À chaque niveau de la hiérarchie, une distribution uniforme est adoptée.

- *Indépendance des distributions conditionnelles sur les intervalles des modèles de discrétisation.*

Notations :
Les notations utilisées dans la suite de ce chapitre sont présentées ici.

Les données
Les données D sont composées de deux sous-ensembles T et U qui correspondent respectivement aux données étiquetées et non-étiquetées, avec $D = T \cup U$. L'ensemble T contient des couples (x, y), où $x \in \mathbb{R}$ et $y \in \mathbb{Y}$ est une valeur discrète représentant la classe de l'exemple x. L'ensemble U contient des réels. Les notations suivantes sont adoptées :

- N, le nombre d'exemples observables ($N = |D|$) ;
- N^l, le nombre d'exemples étiquetés ($N^l = |T|$) ;
- J, le nombre de classes observées dans les données ($J = |\mathbb{Y}|$).

La famille de modèles
La famille $\mathbb{M}$ contient des modèles $M(I, \{N_i\}, \{N_{ij}\})$ dont les paramètres sont :

- I, le nombre d'intervalles du modèle ;
- $\{N_i\}$, le nombre d'exemples dans chaque intervalle ;
- $\{N_{ij}\}$, le nombre d'exemples de chaque classe dans chaque intervalle ;

Étant donnés un modèle M et des données D, on définit également :

- N_i^l, le nombre d'exemples étiquetés dans l'intervalle i
- N_{ij}^l, le nombre d'exemples étiquetés de la classe j dans l'intervalle i.

Une approche Bayésienne :
L'objectif de notre méthode de discrétisation est de sélectionner le modèle le plus probable connaissant les données. À la différence de l'approche MODL, notre méthode manipule conjointement des exemples étiquetés et des exemples non-étiquetés. Une approche Bayésienne est employée, la distribution a posteriori des modèles connaissant les données est définie par l'Équation 4.6.

$$P(M|D) = \frac{P(M)\ P(D|M)}{P(D)} \tag{4.6}$$

4.2.2 Distribution a priori des modèles

La distribution a priori des modèles est de la même forme que dans l'approche supervisée MODL [Boullé, 2006b]. $P(M)$ est définie grâce aux paramètres de modélisation de $\mathbb{M}$. Cette loi a priori exploite la hiérarchie du paramétrage des modèles. Tout d'abord le nombre d'intervalles I est choisi, les bornes des intervalles $\{N_i\}$ puis les effectifs de chaque classe $\{N_{ij}\}$ sont ensuite définis. A chaque niveau de cette hiérarchie, un prior uniforme est adopté. La distribution jointe $P(I, \{N_i\}, \{N_{ij}\})$ peut être décomposée comme suit :

$$P(M) = P(I) \times P(\{N_i\}|I) \times P(\{N_{ij}\}|\{N_i\}, I)$$

Le nombre d'intervalles I est supposé être uniformément distribué entre 1 et N :

$$P(I) = \frac{1}{N}$$

Toutes les subdivisions des données en I intervalles sont supposées être équiprobables, pour un nombre d'intervalles donné. Le calcul de la probabilité d'un ensemble de bornes $P(\{N_i\}|I)$ se pose comme un problème de dénombrement. Le nombre de discrétisations possibles de N instances en I intervalles est égale à C_{N+I-1}^{I-1} :

$$P(\{N_i\}|I) = \frac{1}{C_{N+I-1}^{I-1}}$$

Le dernier terme $P(\{N_{ij}\}, \{N_i\}, I)$ peut être écrit sous la forme d'un produit grâce à l'hypothèse d'indépendance des distributions des classes entre les intervalles. Pour un intervalle donné i contenant N_i instances, toutes les distributions possibles des classes sont considérées comme étant équiprobables. Le nombre d'affectations possibles de j classes à N_i instances se dénombre par $C_{N_i+J-1}^{J-1}$:

$$P(\{N_{ij}\}|\{N_i\}, I) = \prod_{i=1}^{I} \frac{1}{C_{N_i+J-1}^{J-1}}$$

Finalement, la distribution a priori des modèles s'écrit comme suit :

$$P(M) = \frac{1}{N} \times \frac{1}{C_{N+I-1}^{I-1}} \times \prod_{i=1}^{I} \frac{1}{C_{N_i+J-1}^{J-1}} \tag{4.7}$$

4.2.3 Vraisemblance des données conditionnelle au modèle

Nous présentons ici $P(D|M)$, la vraisemblance des données étant donné le modèle. Tout d'abord, Λ la famille de modèles d'étiquetage est définie. Notre méthode de discrétisation semi-supervisée manipule des données étiquetées et non-étiquetées, Λ représente tous les étiquetages possibles. Chaque modèle d'étiquetage $\lambda(N^l, \{N_i^l\}, \{N_{ij}^l\}) \in \Lambda$ est caractérisé par les paramètres suivant :

- N^l le nombre d'exemples étiquetés
- N_i^l le nombre d'exemples étiquetés dans l'intervalle i
- N_{ij}^l le nombre d'exemples étant de la classe j, dans l'intervalle i

Selon la formule de la probabilité totale, $\sum_{\lambda\in\Lambda} P(\lambda|M) = 1$. La vraisemblance peut s'écrire de la manière suivante :

$$P(D|M) = \sum_{\lambda\in\Lambda} P(\lambda|M) \times P(D|M,\lambda)$$

$P(D|M)$ se décompose en une somme sur tous les étiquetages possibles. Cette somme implique $P(\lambda|M)$ la probabilité du modèle d'étiquetage connaissant le modèle de discrétisation. Ce calcul peut être simplifié en considérant que $P(D|M,\lambda)$ est nulle pour tous les modèles d'étiquetage incompatibles avec les données observées et avec le modèle de discrétisation M. Un seul modèle d'étiquetage λ^* est considéré. L'expression précédente peut s'écrire de la manière suivante :

$$P(D|M) = P(\lambda^*|M) \times P(D|M,\lambda^*)$$

Le premier terme $P(\lambda^*|M)$ peut s'écrire sous la forme d'un produit, grâce à l'hypothèse d'indépendance des distributions entre les intervalles du modèle de discrétisation. Dans un intervalle i contenant N_{ij} instances de chaque classe, le calcul de $P(\lambda^*|M)$ revient à chercher la probabilité d'obtenir N_{ij}^l instances de chaque classe en tirant N_i^l instances. Le nombre de tirages conduisant aux $\{N_{ij}^l\}$ peut être dénombré. Pour calculer $P(\lambda^*|M)$, on fait l'hypothèse que les N_i^l instances étiquetées dans un intervalle sont choisies aléatoirement, selon une distribution uniforme :

$$P(\lambda^*|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} C_{N_{ij}}^{N_{ij}^l}}{C_{N_i}^{N_i^l}} \tag{4.8}$$

Considérons un exemple très simple et intuitif pour expliquer l'Équation 4.8. Un intervalle i peut être comparé à un "sac" contenant des boules noires et des boules blanches. Le nombre de boules de chaque sorte est connu et correspond aux paramètres $\{N_{ij}\}$. Dans le cadre de notre exemple illustratif, le sac contient $N_{i1} = 6$ boules noires et $N_{i2} = 20$

boules blanches. Quelle est la probabilité de tirer simultanément $N_{i1}^l = 2$ boules noires et $N_{i2}^l = 3$ boules blanches ? Pour répondre à cette question, on utilise C_{26}^5 le nombre de tirages possibles et $C_6^2 \times C_{20}^3$ le nombre de tirages composés de 2 boules noires et 3 boules blanches. En considérant que tous les tirages sont équiprobables, la probabilité à calculer est donnée par : $\frac{C_6^2 \times C_{20}^3}{C_{26}^5}$.

Le deuxième terme $P(D|M, \lambda^*)$ est estimé en adoptant un a priori uniforme sur l'ensemble des permutations possibles de N_{ij}^l exemples de chaque classe parmi N_i^l exemples. L'hypothèse d'indépendance des distributions entre les intervalles du modèle de discrétisation est une nouvelle fois exploitée :

$$P(D|M, \lambda^*) = \prod_{i=1}^{I} \frac{1}{\frac{N_i^l!}{N_{i1}^l! N_{i2}^l! \ldots N_{iJ}^l!}} = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} N_{ij}^l!}{N_i^l!}$$

Finalement, la vraisemblance des données connaissant le modèle est donnée par :

$$P(D|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} C_{N_{ij}}^{N_{ij}^l} \times N_{ij}^l!}{C_{N_i}^{N_i^l} \times N_i^l!}$$

Dans chaque intervalle, le nombre d'exemples non-étiquetés est dénoté par : $N_{ij}^u = N_{ij} - N_{ij}^l$ et $N_i^u = N_i - N_i^l$. L'expression précédente se simplifie :

$$P(D|M) = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} \frac{N_{ij}!}{N_{ij}^u!}}{\frac{N_i!}{N_i^u!}}$$

$$P(D|M) = \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i^u!}{\prod_{j=1}^{J} N_{ij}^u!} \right] \tag{4.9}$$

4.2.4 Critère d'évaluation des modèles

Le meilleur modèle de discrétisation $M \in \mathbb{M}$ maximise $P(M|D)$ la probabilité a posteriori du modèle connaissant les données. Le critère d'évaluation présenté dans cette partie est issu de l'Équation 4.7 et de l'Équation 4.9. Le modèle maximum a posteriori ($\mathcal{M}_{map} \in \mathbb{M}$) est défini de la manière suivante :

$$\mathcal{M}_{map} = \underset{M \in \mathbb{M}}{ArgMax} \left[\frac{1}{N} \times \frac{1}{C_{N+I-1}^{I-1}} \times \prod_{i=1}^{I} \frac{1}{C_{N_i+J-1}^{J-1}} \times \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i^u!}{\prod_{j=1}^{J} N_{ij}^u!} \right] \right] \tag{4.10}$$

$I(M|D) = -\log(P(M|D))$ est l'information d'un modèle conditionnelle aux données. Maximiser $P(M|D)$ revient à minimiser $I(M|D)$. Le $\mathcal{M}_{map}$ peut être défini grâce au critère $\mathcal{C}_{semi\ super}$:

$$\begin{aligned}\mathcal{M}_{map} = \underset{M\in\mathbb{M}}{ArgMin}\, \mathcal{C}_{semi\ super}(M) = \underset{M\in\mathbb{M}}{ArgMin}\, \log(N) + \log C_{N+I-1}^{I-1} \\ + \sum_{i=1}^{I} \log C_{N_i+J-1}^{J-1} + \sum_{i=1}^{I} \log\left(\frac{N_i!}{\sum_{j=1}^{J} N_{ij}!}\right) - \sum_{i=1}^{I} \log\left(\frac{N_i^u!}{\sum_{j=1}^{J} N_{ij}^u!}\right)\end{aligned} \tag{4.11}$$

Le premier terme du critère $\mathcal{C}_{semi\ super}$ correspond au choix du nombre d'intervalles et le second terme correspond au choix des bornes des intervalles. Le troisième terme représente le choix de la distribution de chaque classe dans chaque intervalle. Les deux derniers termes codent la vraisemblance des données connaissant le modèle.

4.3 Résultats théoriques et empiriques

Nous présentons ici les résultats théoriques et empiriques obtenus lors d'une étude approfondie de notre méthode de discrétisation semi-supervisée.

4.3.1 Comportement asymptotique du critère

Cette section compare les critères d'évaluation des approches de discrétisation supervisée et semi-supervisée. Tout d'abord, le critère $\mathcal{C}_{semi\ super}$ est réécrit pour le cas où les données sont entièrement étiquetées. Le même travail est effectué lorsqu'aucun exemple d'apprentissage n'est étiqueté. Cette étude montre que le critère d'évaluation est consistant avec l'approche de discrétisation supervisée MODL [Boullé, 2006b]. Enfin, le comportement de notre critère est étudié dans le cas semi-supervisé, lorsque le nombre d'exemples étiquetés varie.

Le critère d'évaluation de l'approche supervisée ne considère que les exemples étiquetés. Le critère $\mathcal{C}$ (Équation 4.5) peut être réécrit en employant les notations de la Section 4.2.1 :

$$\mathcal{C}_{super} = \underbrace{\log(N^l) + \log C_{N^l+I-1}^{I-1} + \sum_{i=1}^{I} \log C_{N_i^l+J-1}^{J-1}}_{-\log P(M)} + \underbrace{\sum_{i=1}^{I} \log\left(\frac{N_i^l!}{\sum_{j=1}^{J} N_{ij}^l!}\right)}_{-\log P(T|M)} \tag{4.12}$$

Données entièrement étiquetées :
Les données sont supposées être entièrement étiquetées. Nous considérons le cas particulier où $D = L$ et $U = \emptyset$. Le dernier terme de l'Équation 4.11 est nulle car pour chaque intervalle pour chaque classe, $N_i^u = 0$ et $N_{ij}^u = 0$. Le critère $\mathcal{C}_{semi\ super}$ peut se réécrire comme suit :

$$\mathcal{C}_{semi\ super}(M) = \log(N) + \log(C_{N+I-1}^{I-1}) + \sum_{i=1}^{I} \log(C_{N_i+J-1}^{J-1}) + \sum_{i=1}^{I} \log\left(\frac{N_i!}{\sum_{j=1}^{J} N_{ij}!}\right)$$

Lorsque tous les exemples sont étiquetés, $N = N^l$, $N_i = N_i^l$ pour chaque intervalle et $N_{ij} = N_{ij}^l$. Dans ce cas, le critère semi-supervisé $\mathcal{C}_{semi\ super}$ et le critère supervisé $\mathcal{C}_{super}$ sont identiques ; cela montre la cohérence des deux approches.

Données entièrement non-étiquetées :
Nous nous plaçons dans le cas contraire, où aucun exemple d'apprentissage n'est étiqueté ($D = U$ et $L = \emptyset$). Pour chaque intervalle et pour chaque classe $N_i^u = N_i$ et $N_{ij}^u = N_{ij}$. La vraisemblance des données $P(D|M)$ est égale à 1 quelque soit le modèle envisagé. La vraisemblance (Équation 4.9) peut se réécrire comme suit :

$$P(D|M) = \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i!}{\prod_{j=1}^{J} N_{ij}!}\right]$$
$$P(D|M) = 1$$

La distribution a posteriori du modèle est égale à la distribution a priori du modèle, $P(M|D) = P(M)$. Le modèle le plus probable ($\mathcal{M}_{map}$) contient un seul intervalle. Les deux approches donnent la même discrétisation car tous les paramètres du critère $\mathcal{C}_{super}$ sont nuls dans ce cas. Le critère d'évaluation $\mathcal{C}_{semi\ super}$ s'exprime de la manière suivante :

$$\mathcal{C}_{semi\ super}(M) = \log(N) + \log(C_{N+I-1}^{I-1}) + \sum_{i=1}^{I} \log(C_{N_i+J-1}^{J-1})$$

Mélange de données étiquetées et non-étiquetées :
Les méthodes de discrétisation supervisée et semi-supervisée se différencient par le nombre de modèles codés par la distribution a priori $P(M)$. Dans le cas semi-supervisé, $|\mathbb{M}|$ le nombre de modèles possibles est plus important que dans le cas supervisé. Les exemples non-étiquetés sont autant d'emplacements supplémentaires où les bornes séparant les intervalles du modèle optimal peuvent se positionner. Le nombre de choix possibles sur l'emplacement des bornes étant plus important, le coût de modélisation de la distribution $P(M)$ augmente. Lorsque le nombre d'exemples non-étiquetés augmente, l'optimisation du critère $\mathcal{C}_{semi\ super}$ induit un modèle optimal $\mathcal{M}_{map}$ incluant potentiellement un plus petit nombre d'intervalles.

Ce comportement est illustré grâce à une expérience très simple. Nous considérons un problème de classification binaire pour lequel tous les exemples de la classe "0" [*respectivement* "1"] sont localisés en $x = 0$ [*respectivement* $x = 1$]. Le nombre total d'exemples

d'apprentissage N varie durant l'expérience. Pour chaque valeur de N, on cherche N^l_{min} le nombre minimum d'exemples à étiqueter pour que le critère $\mathcal{C}_{semi\ super}$ préfère un modèle à deux intervalles, plutôt qu'un modèle à un intervalle. Durant toute l'expérience, il y a autant d'exemples étiquetés dans chacune des classes.

La Figure 4.4 trace N^l_{min} en fonction de $N = N^l + N^u$, pour les approches supervisée et semi-supervisée (le graphique de droite utilise une échelle logarithmique). Pour le critère $\mathcal{C}_{super}$, le nombre minimal d'exemples étiquetés nécessaire pour discrétiser correctement les données ne dépend pas de N. Dans le cas supervisé, $N^l_{min} = 6$ quelle que soit la valeur de N. Un comportement différent est observé pour le critère $\mathcal{C}_{semi\ super}$. La figure 4.4 quantifie l'influence de N sur la sélection du modèle optimal $\mathcal{M}_{map}$. Lorsque le nombre total d'exemples N augmente, N^l_{min} croît de l'ordre de $\log(N)$.

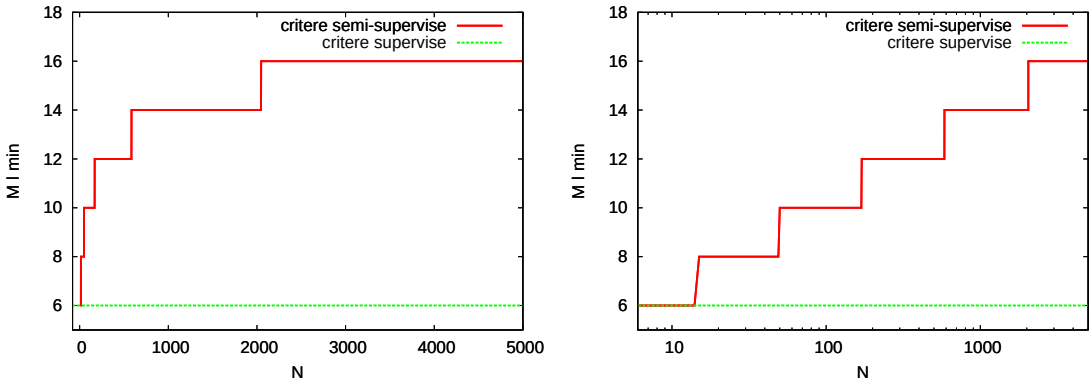

FIG. 4.4 – Mélange d'exemples étiquetés et non-étiquetés : l'axe vertical représente le nombre minimum d'exemples étiquetés pour préférer un modèle à deux intervalles ; l'axe horizontal représente le nombre total d'exemples (échelle logarithmique à droite)

Synthèse des résultats :

La Figure 4.5 synthétise les résultats de la Section 4.3.1 :

- Les critères $\mathcal{C}_{semi\ super}$ et $\mathcal{C}_{super}$ sont analytiquement équivalents lorsque $U = \emptyset$;
- Le critère $\mathcal{C}_{semi\ super}$ est égal à $-\log P(M)$ lorsque $L = \emptyset$, dans ce cas, les approches supervisée et semi-supervisée donnent la même discrétisation ;
- L'approche semi-supervisée est pénalisée par un coût de modélisation élevé lorsque les données comportent des exemples étiquetés et non-étiquetés, dans ce cas, l'optimisation du critère $\mathcal{C}_{semi\ super}$ définit un modèle optimal avec moins d'intervalles que pour l'approche supervisée.

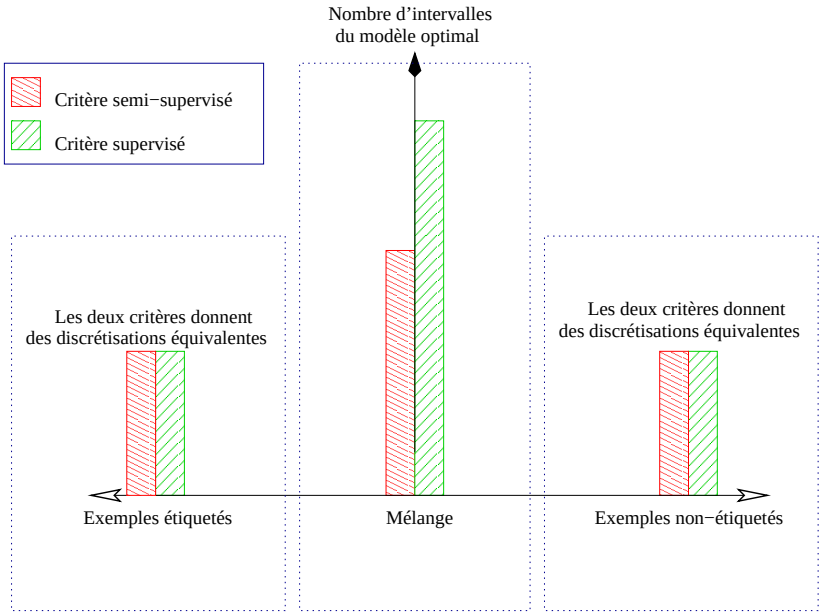

FIG. 4.5 – Résumé de la Section 4.3.1 : L'axe horizontal représente le type de données et l'axe vertical symbolise le nombre d'intervalles du modèle optimal $\mathcal{M}_{map}$

4.3.2 Optimisation du critère

Le modèle de discrétisation le plus probable connaissant les données (le $\mathcal{M}_{map}$) est défini par l'optimisation du critère d'évaluation $\mathcal{C}_{semi\ super}$. Cette section présente l'optimisation de notre critère et montre comment les algorithmes de l'approche MODL peuvent être utilisés dans le cas semi-supervisé. Tout d'abord, une approche gloutonne est présentée. Des heuristiques de post-optimisation sont ensuite utilisées pour améliorer les performances de l'approche gloutonne. La description de ces algorithmes est empruntée à [Boullé, 2007b] et traite uniquement du cas de la discrétisation d'une variable numérique.

4.3.2.1 Propriété de la solution optimale

Dans l'approche supervisée MODL [Boullé, 2006b] l'optimisation des paramètres $\{N_{ij}\}$ se déduit facilement des données, lorsque les $\{N_i\}$ sont fixés. Les bornes des intervalles $\{N_i\}$ sont utilisées pour déterminer $\{N_{ij}\}$, les effectifs de chaque classe dans chaque intervalle. Dans le cas de la discrétisation semi-supervisée, seuls les $\{N_{ij}^l\}$ peuvent être déduits de cette façon. Les effectifs des données étiquetées sont connus pour chaque intervalle, mais les paramètres N_{ij} restent indéterminés à cause des exemples non-étiquetés.

Les $\{N_{ij}\}$ qui maximisent le critère $\mathcal{C}_{semi\ super}$ sont notés $\{N_{ij}^\diamond\}$. Nous montrons que dans chaque intervalle, les $\{N_{ij}^\diamond\}$ respectent la proportion des classes observées sur les exemples étiquetés. Les $\{N_{ij}^\diamond\}$ sont définis comme suit :

$$N_{ij}^{\diamond} = \left\lceil (N_i + 1) \times \frac{N_{ij}^l}{N_i^l} - 1 \right\rceil \tag{4.13}$$

Démonstration.
Cette démonstration se restreint au cas d'un modèle de discrétisation incluant un seul intervalle. Ces calculs peuvent être répétés indépendamment sur I intervalles, puisque la distribution des données est supposée être indépendante entre les intervalles du modèle. Nous considérons ici un problème de classification binaire ($J = 2$). Soit la fonction $f(N_{i1}, N_{i2})$ correspondant au critère $\mathcal{C}_{semi\,sup}$, pour lequel tous les paramètres sont fixés exceptés N_{i1} et N_{i2}. L'Équation 4.9 peut se réécrire comme suit :

$$P(D|M) = \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{N_{ij}!(N_i - N_i^l)!}{(N_{ij} - N_{ij}^l)!N_i!}$$

Le seul terme du critère $\mathcal{C}_{semi\,sup}$ qui dépent des paramètres $\{N_{ij}\}$ est $-\log(P(D|M))$:

$$-\log(P(D|M)) = \sum_{i=1}^{I} \sum_{j=1}^{J} \log\left(\frac{(N_{ij} - N_{ij}^l)!N_i!}{N_{ij}!(N_i - N_i^l)!}\right)$$

Dans notre cas, les paramètres N_i et N_i^l sont constants. Notre objectif est de caractérisé le minimum de la fonction $f(N_{i1}, N_{i2})$ par une expression analytique :

$$f(N_{i1}, N_{i2}) = \log\left(\frac{(N_{i1} - N_{i1}^l)!}{N_{i1}!}\right) + \log\left(\frac{(N_{i2} - N_{i2}^l)!}{N_{i2}!}\right)$$

Les termes N_{i1}^l et N_{i2}^l sont constants, et $N_{i2} = N_i - N_{i1}$. f peut être réécrite comme une fonction impliquant une seule variable :

$$\begin{aligned} f(N_{i1}) &= \log\left(\frac{(N_{i1} - N_{i1}^l)!}{N_{i1}!}\right) + \log\left(\frac{(N_i - N_{i1} - N_{i2}^l)!}{(N_i - N_{i1})!}\right) \\ &= \sum_{k=1}^{N_{i1} - N_{i1}^l} \log k - \sum_{k=1}^{N_{i1}} \log k + \sum_{k=1}^{N_i - N_{i1} - N_{i2}^l} \log k - \sum_{k=1}^{N_i - N_{i1}} \log k \\ &= -\sum_{k=N_{i1} - N_{i1}^l + 1}^{N_{i1}} \log k - \sum_{k=N_i - N_{i1} - N_{i2}^l + 1}^{N_i - N_{i1}} \log k \end{aligned}$$

Et

$$f(N_{i1} + 1) = -\sum_{k=N_{i1} - N_{i1}^l + 2}^{N_{i1}+1} \log k - \sum_{k=N_i - N_{i1} - N_{i2}^l}^{N_i - N_{i1} - 1} \log k$$

Par conséquent :

$$\begin{aligned} f(N_{i1}) - f(N_{i1}+1) &= \log(N_{i1}+1) - \log(N_{i1}+1-N_{i1}^l) - \log(N_i - N_{i1}) + \log(N_i - N_{i2}^l - N_{i1}) \\ &= \log\left(\frac{(N_{i1}+1)(N_i - N_{i2}^l - N_{i1})}{(N_{i1}+1-N_{i1}^l)(N_i - N_{i1})}\right) \end{aligned}$$

$f(N_{i1})$ croît si :

$f(N_{i1}) - f(N_{i1}+1) > 0$

$\Leftrightarrow \quad \frac{(N_{i1}+1)(N_i - N_{i2}^l - N_{i1})}{(N_{i1}+1-N_{i1}^l)(N_i - N_{i1})} > 1$

$\Leftrightarrow \quad -N_{i2}^l \times N_{i1} - N_{i2}^l > -N_{i1}^l \times N_i + N_{i1}^l \times N_{i1}$

$\Leftrightarrow \quad N_{i1} < \frac{-N_{i2}^l + N_{i1}^l \times N_i}{N_{i1}^l + N_{i2}^l}$

De la même façon, $f(N_{i1})$ décroit si :

$f(N_{i1}) - f(N_{i1}+1) < 0 \Leftrightarrow N_{i1} > \frac{-N_{i2}^l + N_{i1}^l \times N_i}{N_{i1}^l + N_{i2}^l}$

$f(N_{i1})$ étant une fonction discrète, son maximum est atteint pour $N_{i1} = \lceil \frac{-N_{i2}^l + N_{i1}^l \times N_i}{N_{i1}^l + N_{i2}^l} \rceil$.

Cette expression peut être généralisée pour des problèmes à J classes[1] :

$$N_{ij}^{\diamond} = \left\lceil (N_i + 1) \times \frac{N_{ij}^l}{N_i^l} - 1 \right\rceil$$

□

La partie entière "$\lceil . \rceil$" utilisée dans l'expression 4.13 peut provoquer des problèmes d'arrondi. Dans le cas où $\sum_{j=1}^{J} N_{ij}^{\diamond} = N_i - 1$, il suffit de choisir un des $\{N_{ij}^{\diamond}\}$ et de lui ajouter 1. Tous les choix possibles sont équivalents et optimaux au sens du critère $\mathcal{C}_{semi\ super}$. Ce phénomène a été observé sur des cas pratiques, et est démontré dans le cas d'un problème binaire et d'un modèle à un intervalle (voir Annexe B).

L'Équation 4.13 est une expression analytique permettant de déduire les $\{N_{ij}^{\diamond}\}$ à partir des données. De cette façon, les paramètres à optimiser sont les mêmes que dans l'approche MODL et les mêmes algorithmes d'optimisation peuvent être exploités.

4.3.2.2 Heuristique gloutonne

L'heuristique gloutonne ascendante est décrite par l'Algorithme 7. Il s'agit d'un algorithme générique utilisé pour l'optimisation d'une partition univariée [Zighed and Rakotomalala, 2000]. Dans notre cas, l'objectif est de trouver le modèle de discrétisation

[1] L'expression généralisée de $N_{ij}^{\diamond}$ a été vérifiée empiriquement sur des problèmes de classification multi-classes.

$M \in \mathbb{M}$ qui minimise la fonction $\mathcal{C}out(M)$, c'est-à-dire la valeur du critère $\mathcal{C}_{semi\ super}$. Le modèle initial $M_{initial}$ comporte autant d'intervalles qu'il y a d'exemples d'apprentissage. Cette heuristique consiste à évaluer toutes les fusions possibles m entre deux intervalles adjacents. La meilleure fusion est effectuée si elle améliore le coût du modèle courant. L'Algorithme 7 est itératif, il se répète tant que le modèle courant est amélioré.

Notations :

- La fonction $\mathcal{C}out : \mathbb{M} \rightarrow \mathbb{R}$, correspondant à la valeur du critère $\mathcal{C}_{semi\ super}$
- $M_{initial}$, le modèle de discrétisation initial, tel que $I = N$ et $N_i = 1, \quad \forall i \in [1, I]$
- M', le modèle de discrétisation optimal

```
/* Initialisation des variables */
M' ← M_initial
amelioration = vrai

Répéter
    /* Recherche de la meilleure amélioration */
    M_courant ← M'

    Pour toutes les fusions m de deux intervalles adjacents faire
        /* Évaluation de la fusion m */
        M_fusion ← M' + m

        Si Cout(M_fusion) < Cout(M_courant) Alors
            M_courant ← M_fusion
        Fin Si

    Fin Pour

    /* Test de l'amélioration */
    Si Cout(M_courant) < Cout(M') Alors
        M' ← M_courant  amelioration = vrai
    Sinon
        amelioration = faux
    Fin Si

Tant que amelioration = vrai
```

Algorithme 7: Heuristique gloutonne ascendante

Le modèle de discrétisation retourné par cet algorithme est une solution sous-optimale, car l'espace des modèles $\mathbb{M}$ n'est que partiellement parcouru. Les fusions successives peuvent conduire à un optimum local. Implémenté naïvement, cet algorithme a une com-

plexité temporelle en $\mathcal{O}(N^3)$. L'additivité du Critère 4.11 peut être exploitée pour mémoriser des résultats intermédiaires et réduire l'impact de chaque fusion aux deux intervalles impliqués. Finalement, l'heuristique gloutonne ascendante est implémentée en $\mathcal{O}(N \log N)$.

4.3.2.3 Heuristiques de post-optimisation

L'heuristique gloutonne ascendante est suivie de deux étapes de post-optimisation qui ont pour objectif d'améliorer le modèle retourné par l'Algorithme 7, noté $M'(I', \{N'_i\}, \{N'_{ij}\})$. Le nombre d'intervalles I' correspond éventuellement à un optimum local, car l'heuristique gloutonne s'arrête lorsqu'aucune fusion d'intervalles adjacents n'améliore le modèle courant. La première étape de post-optimisation agit sur le nombre d'intervalles du modèle de discrétisation. L'Algorithme 7 est répété jusqu'à l'obtention d'un modèle comportant un seul intervalle, noté $M''(I'', \{N''_i\}, \{N''_{ij}\})$. À chaque itération, la meilleure fusion est effectuée même si elle n'améliore pas immédiatement le modèle courant, et la meilleure discrétisation rencontrée est retenue.

La deuxième étape de post-optimisation concerne les bornes des intervalles, $\{N''_i\}$. Ces bornes sont issues de fusions successives qui ne sont jamais remises en question. Les paramètres $\{N''_i\}$ correspondent éventuellement à un optimum local, car toutes les bornes possibles n'ont pas été envisagées. Pour résoudre ce problème, un voisinage local de M'' basé sur des opérations élémentaires entre deux intervalles adjacents est défini :

- suppression d'un intervalle, par fusion de trois intervalles adjacents existant suivie d'un découpage de l'intervalle fusionné ;
- ajustement de frontières entre deux intervalles, par fusion de deux intervalles adjacents existants suivie d'un découpage de l'intervalle fusionné ;
- ajout d'un nouvel intervalle, par découpage d'un intervalle existant.

L'additivité du critère $\mathcal{C}_{semi\ super}$ permet le parcours exhaustif du voisinage d'un modèle M'' en $\mathcal{O}(N)$. L'exploration systématique de ce voisinage permet d'échapper à une gamme importante d'optima locaux et d'améliorer la qualité du modèle de discrétisation [Boullé, 2007b].

4.3.3 Biais de discrétisation

Les approches de discrétisation supervisée et semi-supervisée sont basées sur les statistiques de rang. Les emplacements des bornes séparant les intervalles du modèle optimal sont définis dans un espace discret, grâce au nombre d'exemples appartenant à chaque intervalles, $\{N_i\}$. Le biais de discrétisation présenté dans cette section a pour objectif de positionner les bornes du modèle dans un espace continu : le domaine de définition de la variable explicative. Dès lors, deux problèmes élémentaires se posent : i) le positionnement d'une borne entre deux exemples d'apprentissage ; ii) le positionnement d'une borne parmi des exemples non-étiquetés contigus.

Positionnement d'une borne entre deux exemples d'apprentissage :
Les paramètres $\{N_i\}$ [*respectivement* $\{N_i^l\}$] issus de l'optimisation du critère $\mathcal{C}_{semi\ super}$ [*respectivement* $\mathcal{C}_{super}$] ne sont pas suffisants pour définir l'emplacement des bornes dans un domaine continu. Il existe une infinité de positions possibles pour une borne entre deux exemples d'apprentissage. Un prior est adopté dans [Boullé, 2006b] qui considère la meilleure position d'une borne comme étant la médiane de la distribution de l'emplacement de la vraie borne, notée P_{tb}. Cette médiane minimise l'erreur quadratique moyenne de généralisation (MSE), quelle que soit la distribution P_{tb}. L'objectif est de placer une borne entre deux exemples d'apprentissage, sans disposer d'informations relatives à la distribution P_{tb}. Dans ce cas, P_{tb} est supposée être uniforme. Une borne est placée à mi-chemin des deux exemples d'apprentissage concernés.

Positionnement d'une borne parmi des exemples non-étiquetés contigus :
Lorsque les paramètres $\{N_i^l\}$ restent constants, l'optimisation du critère semi-supervisé ne définit pas de manière unique l'emplacement des bornes du $\mathcal{M}_{map}$. Ce phénomène est illustré dans la suite de cette section par un problème jouet. Considérons une zone de $\mathbb{X}$ où aucun exemple n'est étiqueté ; toutes les bornes possibles de cette zone ont le même coût au sens du critère $\mathcal{C}_{semi\ super}$. Notre approche de discrétisation semi-supervisée n'est pas capable de déterminer le meilleur emplacement d'une borne dans une zone non-étiquetée de l'espace $\mathbb{X}$. Le même prior que précédemment, minimisant l'erreur quadratique moyenne de généralisation [Boullé, 2006b], est adopté pour définir l'emplacement des bornes du modèle optimal. Les exemples non-étiquetés sont supposés être tirés selon P_{tb} et cette distribution est supposée être constante entre chaque exemple d'apprentissage. Les exemples non-étiquetés sont utilisés pour estimer la médiane de P_{tb}. Une borne est finalement placée au milieu d'une zone non-étiquetée[2], à mi-chemin des deux exemples non-étiquetés concernés.

La Figure 4.9 illustre les résultats présentés dans la Section 4.3.3. Les approches de discrétisation supervisée et semi-supervisée sont incapables de positionner une borne entre deux exemples étiquetés. Dans les deux cas, le même prior définit de manière unique le meilleur emplacement des bornes. Le principal apport de l'approche semi-supervisée est d'exploiter les exemples non-étiquetés pour affiner l'estimation de la médiane de P_{tp}.

[2]Le nombre d'exemples non-étiquetés contigus doit être le même de part est d'autre d'une borne.

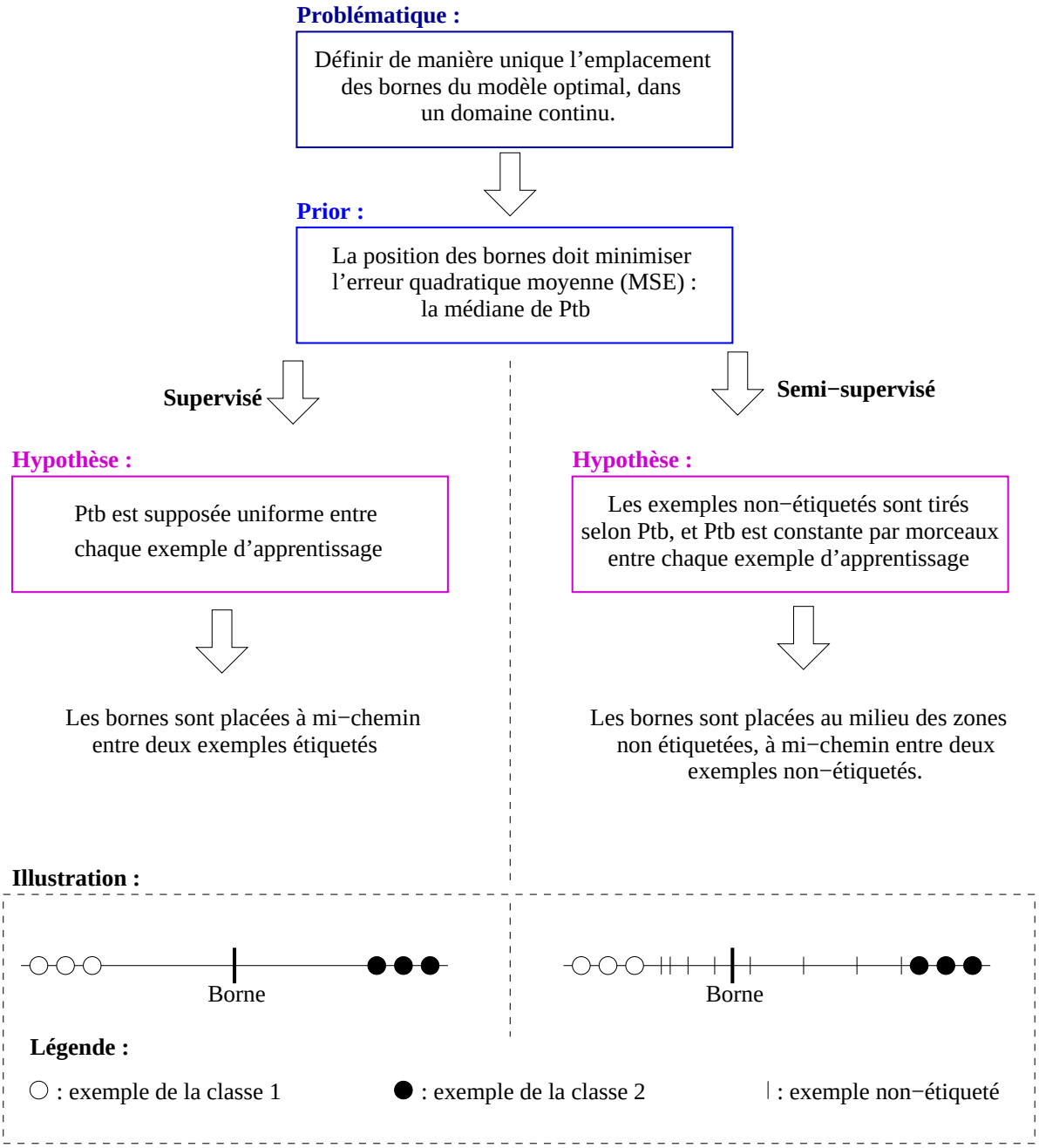

FIG. 4.6 – Résumé des résultats de la Section 4.3.3

Preuve empirique :
Considérons un problème de classification binaire uni-varié, dont les exemples d'apprentissage sont caractérisés par une variable explicative comprise entre 0 et 1. Ce jeu de données comporte trois parties distinctes. La partie "A" de la Figure 4.7 contient 40

exemples étiquetés par la classe "1". Ces exemples sont répartis régulièrement dans l'intervalle $[0, 0.4]$. De la même façon, la partie "B" contient 20 exemples non-étiquetés et la partie "C" contient 40 exemples étiquetés par la classe "2".

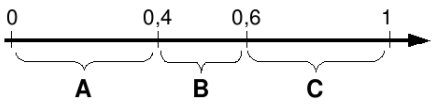

FIG. 4.7 – Problème jouet : les données

Dans le cadre de cette expérience, la famille de modèles $\mathbb{M}$ est restreinte aux modèles à deux intervalles. L'objectif de cette expérience est de trouver la meilleure position de la borne $b \in [0, 1]$ séparant les deux intervalles du modèle. La position d'une borne est déduite grâce au nombre d'exemples appartenant à chaque intervalle, $\{N_1, N_2\}$.

Il existe plusieurs modèles compatibles avec une borne, ces modèles étant définis par les paramètres $\{N_{ij}\}$. La probabilité d'une borne b est estimée par un moyennage Bayésien qui considère tous les modèles compatibles avec la borne b. Cette évaluation n'est pas biaisée par le choix d'un modèle particulier parmi tous les modèles compatibles avec une borne. Pour une borne donnée b, I et $\{N_{ij}\}$l sont fixés et les paramètres $\{N_{ij}\}$ restent indéfinis. $P(b|D)$ est estimée de la manière suivante :

$$P(b|D) = \sum_{\{N_{ij}\}} P(\underbrace{b, \{N_{ij}\}}_{M \in \mathbb{M}} |D)$$

La formule de Bayes est utilisée pour décomposer cette expression :

$$P(b|D) \times P(D) = \sum_{\{N_{ij}\}} P(D|b, \{N_{ij}\}) \times P(b, \{N_{ij}\}) \tag{4.14}$$

Les paramètres $\{N_i\}$ et $\{N_{ij}\}$ définissent un modèle $M \in \mathbb{M}$. Les termes de l'Expression 4.14 correspondent à $P(D|M)P(M)$ et sont calculés grâce au critère $\mathcal{C}_{semi\ super}$.

La Figure 4.8 trace $-\log P(b|D)$ en fonction de la position de la borne b. Les valeurs minimales de cette courbe nous indiquent où doit se positionner la borne pour que les modèles correspondants expliquent au mieux les données observées. Selon la Figure 4.8, il ne faut ni couper dans la partie "A", ni dans la partie "C". En revanche, toutes les coupures de la partie "B" sont équivalentes et optimales au sens du critère $\mathcal{C}_{semi\ super}$.

Le graphique droit de la Figure 4.8 fait un zoom sur la partie "B" du jeu de données et montre que la probabilité des bornes est constante[3]. Cette expérience montre empiriquement que l'optimisation du critère $\mathcal{C}_{semi\ super}$ ne détermine pas de manière unique la position des bornes, dans une zone exclusivement composée d'exemples non-étiquetés.

[3]L'expérience menée montre que $-\log P(b|D)$ est constant dans la partie "B" du jeu de données, à la douzième décimale près. On observe du bruit numérique en augmentant la précision, aucune variation significative ne peut être mise en évidence.

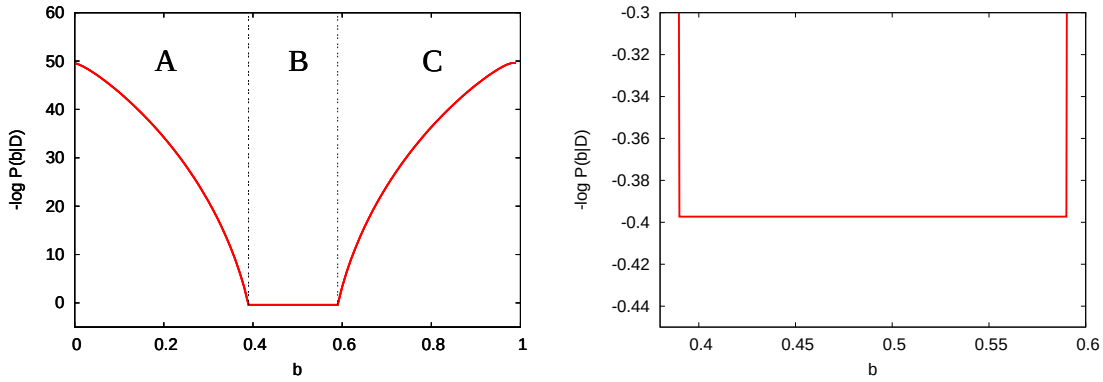

FIG. 4.8 – Quantité d'information de la coupure *vs* position de la coupure. Le graphique de droite fait un zoom sur la partie "B" du jeu de données.

Ce résultat est inattendu et difficile à démontrer formellement en raison du moyennage Bayésien sur les modèles. Le même résultat est obtenu lorsque les parties "A" est "C" comportent un nombre différent d'exemples étiquetés. Intuitivement, ce phénomène peut s'expliquer par le fait qu'aucune préférence n'a été exprimée sur l'emplacement des bornes lors de l'élaboration du critère $\mathcal{C}_{semi\ super}$; cela est cohérent avec une démarche Bayésienne objective [Berger, 2006].

4.3.4 Convergence asymptotique

Cette section montre que notre approche de discrétisation semi-supervisée est asymptotiquement équivalente à l'approche supervisée MODL [Boullé, 2006b], munie d'un post-traitement qui place les bornes du modèle optimal au milieu des zones non-étiquetées. Comme l'illustre la Figure 4.9, la démarche employée pour montrer l'équivalence des deux approches s'appuie sur trois résultats intermédiaires : le biais de discrétisation établi à la section 4.3.3 ; l'optimisation des paramètres $\{N_{ij}\}$ présentée à la section 4.3.2 ; et l'interprétation de la vraisemblance $P(D|M)$ par l'entropie. En exploitant le biais de discrétisation comme une connaissance a priori, nous montrons que le prior $P(M)$ est identique pour les deux approches. Selon le modèle optimal défini par $\mathcal{M}_{map} = min_{M \in \mathbb{M}} \mathcal{C}_{semi\ super}(M)$, l'entropie des ensembles D, U et L est la même. Ce résultat permet d'établir la convergence de la vraisemblance $P(D|M)$ dans le cas supervisé et dans le cas semi-supervisé, lorsque $N^l \rightarrow \infty$, $N^u \rightarrow \infty$ et $\frac{N^l}{N^u} = Cst$.

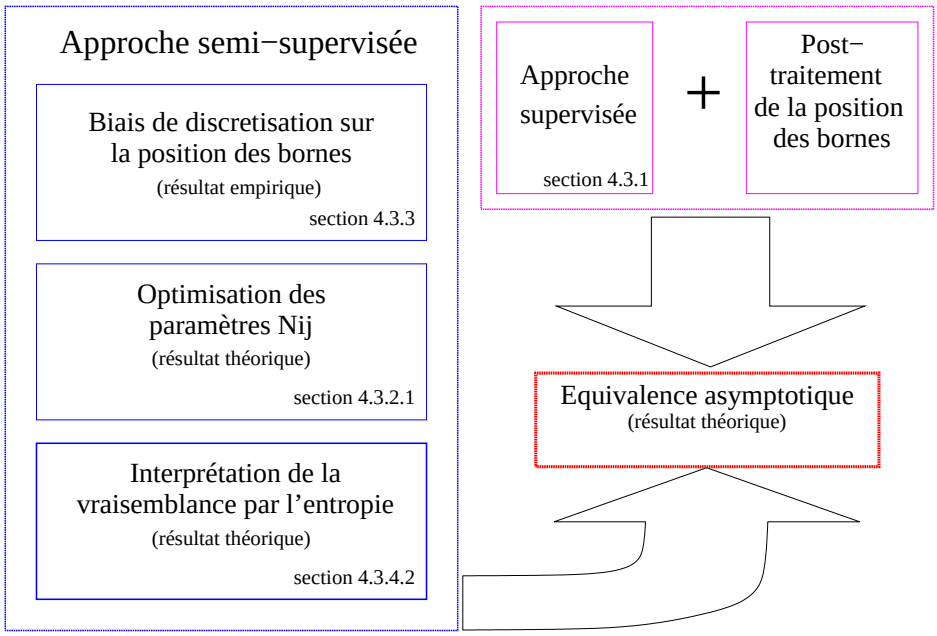

FIG. 4.9 – Convergence des approches supervisée et semi-supervisée.

4.3.4.1 Unification du prior

Le biais de discrétisation défini à la Section 4.3.3 modifie notre connaissance a priori de la distribution des modèles. Les bornes du $\mathcal{M}_{map}$ sont contraintes de se placer au milieu des zones non-étiquetées, ce qui réduit considérablement le nombre de positions envisageables pour chacune des bornes. Le critère $\mathcal{C}_{semi\ super}(M)$ considère N positions possibles pour chaque borne. Nous nous ramenons à N^l positions possibles en exploitant le biais de discrétisation. Dans ces conditions, le prior présenté à la Section 4.2.2 peut être réécrit de manière identique à l'Équation 4.12 :

$$P(M) = \frac{1}{N^l} \times \frac{1}{C^{I-1}_{N^l+I-1}} \times \prod_{i=1}^{I} \frac{1}{C^{J-1}_{N^l_i+J-1}}$$

4.3.4.2 Vraisemblance asymptotiquement identique

Lemme A : *La vraisemblance des données s'exprime selon l'entropie conditionnelle au modèle M (notée H_M) des ensembles Φ, U et L :*

- *Cas supervisé* $-\log P(D|M)_{super} = N^l H_M(L) + \mathcal{O}(\log N)$
- *Cas semi-supervisé* $-\log P(D|M)_{semi\ super} = N H_M(D) - N^u H_M(U) + \mathcal{O}(\log N)$

Démonstration.
• Soit $H_M(D)$ l'entropie de Shannon [Shannon, 1948] des données, connaissant le modèle de discrétisation M. Nous supposons ici que $H_M(D)$ est égale à son estimation empirique :

$$H_M(D) = -\sum_{i=1}^{I} \frac{N_i}{N} \left[\sum_{j=1}^{J} \frac{N_{ij}}{N_i} \log\left(\frac{N_{ij}}{N_i}\right) \right]$$

• Dans le cas semi-supervisé $P(D|M)_{semi\ super} = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} \frac{N_{ij}!}{N_{ij}^u!}}{\frac{N_i!}{N_i^u!}}$. Par conséquent :

$$-\log P(D|M)_{semi\ super} = \sum_{i=1}^{I} \left[\log(N_i!) - \log(N_i^u!) - \sum_{j=1}^{J} \log(N_{ij}!) + \sum_{j=1}^{J} \log(N_{ij}^u!) \right]$$

L'approximation de Stirling donne $\log(n!) = n\log(n) - n + \mathcal{O}(\log n)$:

$$\begin{aligned} -\log P(D|M)_{semi\ super} = \sum_{i=1}^{I} \Big[& N_i \log(N_i) - N_i - N_i^u \log(N_i^u) + N_i^u - \sum_{j=1}^{J} \left[N_{ij}\log(N_{ij}) - N_{ij} \right] \\ & + \sum_{j=1}^{J} \left[N_{ij}^u \log(N_{ij}^u) - N_{ij}^u \right] + \mathcal{O}(\log N_i) - \mathcal{O}(\log N_i^u) \\ & - \sum_{j=1}^{J} \mathcal{O}(\log N_{ij}) + \sum_{j=1}^{J} \mathcal{O}(\log N_{ij}^u) \Big] \end{aligned}$$

Nous exploitons le fait que $\sum_{j=1}^{J} N_{ij} = N_i$ et $\sum_{j=1}^{J} N_{ij}^u = N_i^u$:

$$-\log P(D|M)_{semi\ super} = \sum_{i=1}^{I} \Big[\sum_{j=1}^{J} N_{ij}^u \left(\log N_{ij}^u - \log N_i^u \right) - N_{ij} \left(\log N_{ij} - \log N_i \right) + \mathcal{O}(\log N_i) \Big]$$

$$-\log P(D|M)_{semi\ super} = \sum_{i=1}^{I} \Big[-N_i \sum_{j=1}^{J} \frac{N_{ij}}{N_i} \log\left(\frac{N_{ij}}{N_i}\right) + N_i^u \sum_{j=1}^{J} \frac{N_{ij}^u}{N_i^u} \log\left(\frac{N_{ij}^u}{N_i^u}\right) + \mathcal{O}(\log N_i) \Big]$$

L'entropie étant additive sur des ensembles disjoints, nous obtenons :

$$-\log P(D|M)_{semi\ super} = NH_M(D) - N^u H_M(U) + \mathcal{O}(\log N)$$

• La démonstration pour le cas supervisé est similaire :

$$P(D|M)_{super} = \prod_{i=1}^{I} \frac{\prod_{j=1}^{J} N_{ij}^l!}{N_i^l!}$$

$$-\log P(D|M)_{super} = \sum_{i=1}^{I}\left[\log(N_i^l!) - \sum_{j=1}^{J}\log(N_{ij}^l!)\right]$$

$$-\log P(D|M)_{super} = \sum_{i=1}^{I}\left[N_i^l\log(N_i^l) - N_i^l - \sum_{j=1}^{J}\left[N_{ij}^l\log(N_{ij}^l) - N_{ij}^l\right] + \mathcal{O}(\log N_i^l) - \sum_{j=1}^{J}\mathcal{O}(\log N_{ij}^l)\right]$$

$\sum_{j=1}^{J} N_{ij} = N_i$ et $\sum_{j=1}^{J} N_{ij}^h = N_i^h$ donc on a :

$$-\log P(D|M)_{super} = \sum_{i=1}^{I}\left[-N_i^l\sum_{j=1}^{J}\frac{N_{ij}^l}{N_i^l}\log\left(\frac{N_{ij}^l}{N_i^l}\right) + \mathcal{O}(\log N_i)\right]$$

$$-\log P(D|M)_{super} = N^l H_M(L) + \mathcal{O}(\log N)$$

□

Lemme B : *Les valeurs des paramètres $\{N_{ij}\}$ qui maximisent le critère $\mathcal{C}_{semi\,sup}$ (notés $\{N_{ij}^{\diamond}\}$) correspondent à la proportion des étiquettes observées dans chaque intervalle (voir Section 4.3.2) :*

$$N_{ij}^{\diamond} = \left\lceil (N_i + 1) \times \frac{N_{ij}^l}{N_i^l} - 1 \right\rceil$$

La conséquence du ***lemme B*** est que les ensembles U, L et D ont la même entropie conditionnelle au $\mathcal{M}_{map}$. En effet, les $N_{ij}^{\diamond}$ respectent la proportion des étiquettes observées dans les intervalles du modèle optimal. En exploitant le ***lemme A***, on obtient pour le cas semi-supervisé :

$$-\log P(D|\mathcal{M}_{map})_{semi\,super} = NH_{\mathcal{M}_{map}}(D) - N^u H_{\mathcal{M}_{map}}(U) + \mathcal{O}(\log N)$$

$$-\log P(D|\mathcal{M}_{map})_{semi\,super} = (N - N^u)H_{\mathcal{M}_{map}}(L) + \mathcal{O}(\log N)$$

$$-\log P(D|\mathcal{M}_{map})_{semi\,super} = N^l H_{\mathcal{M}_{map}}(L) + \mathcal{O}(\log N)$$

La vraisemblance du modèle optimal $P(D|\mathcal{M}_{map})$ est asymptotiquement identique dans le cas supervisé et dans le cas semi-supervisé :

$$-\log P(D|\mathcal{M}_{map})_{semi\ super} + \log P(D|\mathcal{M}_{map})_{super} = \mathcal{O}(\log N)$$

$$\lim_{N\to+\infty} \frac{-\log P(D|\mathcal{M}_{map})_{semi\ super} + \log P(D|\mathcal{M}_{map})_{super}}{-\log P(D|\mathcal{M}_{map})_{semi\ super}} = 0$$

L'experience illustrative qui suit montre que l'approche supervisée et l'approche semi-supervisée cherchent à résoudre le même problème d'optimisation lorsque[4] $N^l \to \infty$, $N^u \to \infty$ et $\frac{N^l}{N^u} = Cst$.

Illustration
L'objectif de cette expérience est d'étudier l'écart de la vraisemblance dans le cas semi-supervisé et dans le cas supervisé. Nous définissons la différence des "$-$ log vraisemblances" de la manière suivante :

$$\Delta = -\log P(D|M)_{semi\ super} + \log P(D|M)_{super}$$

Nous définissons également la différence relative des "-log vraisemblances" tel que :

$$\Delta_{\mathcal{R}elatif} = \frac{-\log P(D|M)_{semi\ super} + \log P(D|M)_{super}}{-\log P(D|M)_{semi\ super}}$$

En fixant le modèle de discrétisation M, Δ et $\Delta_{\mathcal{R}elatif}$ dépendent uniquement des données. Durant l'expérience, le nombre total d'exemples N et le nombre d'exemples étiquetés N^l varient. Le même jeu de données qu'à la Section 4.3.3 est utilisé[5]. Le modèle de discrétisation M comporte deux intervalles ($I = 2$) et la borne définie par $\{N_1, N_2\}$ est toujours située à $b = 0.5$. Δ et $\Delta_{\mathcal{R}elatif}$ sont évalués pour chaque valeur du couple (N, N^l), en considérant les $N^{\diamond}_{ij}$ optimaux définis à la Section 4.3.2.1.

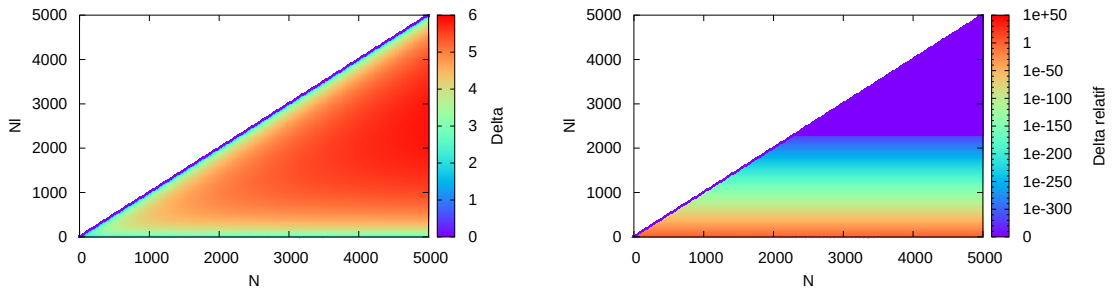

FIG. 4.10 – Δ et $\Delta_{\mathcal{R}elatif}$ en fonction de N et N^l

[4] La convergence des deux approches est également vérifiée pour $N^l \to \infty$ et $N^u = Cst$
[5] Dans le cadre de cette expérience les données sont bruitées, 10% des étiquettes sont erronées.

Le graphique gauche de la Figure 4.10 trace Δ pour $N \in [0, 5000]$ et $N^l \in [0, 5000]$. Un code couleur est utilisé pour indiquer la valeur de Δ en fonction de N et N^l. Ce graphique montre que Δ varie de l'ordre de $\mathcal{O}(\log N)$ lorsque N et N^l augmentent. Par exemple, pour $N = 5000$ et $N^l = 2500$ on observe $\Delta \approx 5$. Cela est cohérent avec l'interprétation basée sur l'entropie qui donne $\Delta = \mathcal{O}(\log N)$.

Lorsque N tend vers l'infini, Δ est négligeable devant la vraisemblance, $\Delta \ll -\log P(D|M)$. Le graphique droit de la Figure 4.10 trace $\Delta_{\mathcal{R}elatif}$ en fonction de N et N^l. Ce graphique montre que l'écart relatif des vraisemblances tend vers 0 lorsque N et N^l tendent vers l'infini. Cette expérience illustre le fait que l'approche supervisée munie d'un post-traitement sur la position des bornes et l'approche semi-supervisée tendent à résoudre le même problème d'optimisation, étant donné le biais de discrétisation défini à la Section 4.3.3.

4.4 Application à des problèmes jouets

Cette section étudie l'influence du post-traitement de la position des bornes du $\mathcal{M}_{map}$ (Section 4.3.4), sur la qualité de la discrétisation optimale. Les expériences de cette section sont menées sur deux problèmes jouets, mettant en jeu une classification binaire. L'objectif est d'estimer la distribution conditionnelle $P(y|x)$, avec $y \in \mathbb{Y}$ et $x \in \mathbb{X}$, en exploitant des données d'apprentissage.

Un classifieur Bayésien naïf est utilisé pour évaluer la méthode de discrétisation munie, ou non, du post-traitement. Ce classifieur exploite la discrétisation optimale pour estimer la distribution conditionnelle $P(y|x)$ [Boullé, 2006b]. L'évaluation du modèle prédictif est effectuée grâce à l'AUC, en fonction du nombre d'exemples étiquetés. Les résultats obtenus sont exprimés sous la forme de courbes d'apprentissage, qui tracent l'AUC moyenne en fonction du nombre d'exemples étiquetés. Les expériences sont répétées 10 fois et les exemples étiquetés sont choisis de manière aléatoire. Les expériences débutent avec 2 exemples étiquetés, à chaque itération, 2 exemples supplémentaires sont étiquetés.

Dans le cadre de nos expériences, la famille $\mathbb{M}$ est restreinte aux modèles incluant un ou deux intervalles. Les deux problèmes jouets considérés impliquent des données d'apprentissage dont la distribution n'est pas uniforme. Cette section présente donc deux cas favorables au post-traitement, pour lesquels les exemples non-étiquetés permettent d'améliorer l'évaluation de la médiane de P_{tb} (Section 4.3.3).

Problèmes jouets :
Nous définissons ici les deux jeux de données exploités lors de nos expériences.

L'échelon
Ce problème jouet consiste à estimer la position du front montant d'un échelon grâce à des données d'apprentissage. Le jeu de données utilisé comporte 100 exemples. La variable explicative x caractérisant les exemples est définie par $x = e^{\alpha}$, α variant de 0 à 10 avec un pas de 0.1. Les 53 exemples tel que $x < 220$ appartiennent à la classe "1" et les 47 autres exemples sont de la classe "2". La Figure 4.11 illustre ce jeu de données.

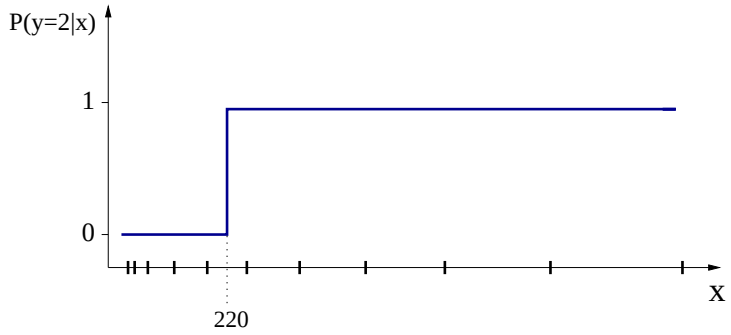

FIG. 4.11 – Jeu de données "échelon"

Les gaussiennes

Ce Jeu de données met en jeu une classification binaire et comporte 1000 exemples. Les exemples appartenant à la classe "1" sont répartis selon une distribution gaussienne ayant pour moyenne 0.25 et pour variance 0.2. De la même façon, les exemples de la classe "2" suivent une distribution gaussienne de moyenne 0.75 et de variance de 0.6. La Figure 4.12 illustre ce jeu de données.

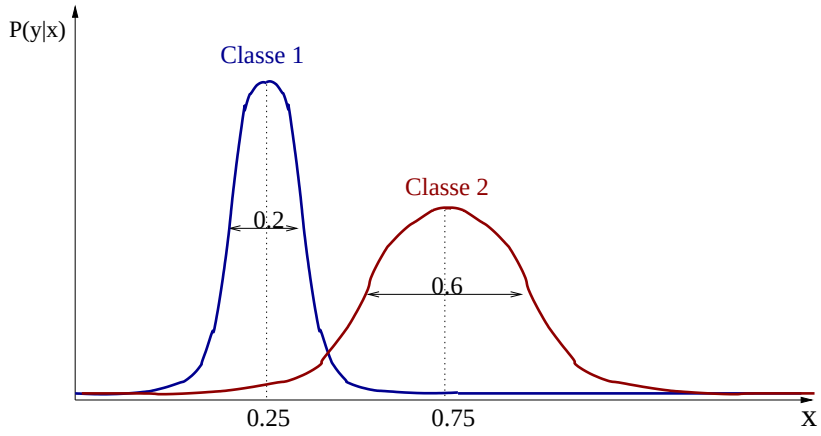

FIG. 4.12 – Jeu de données "gaussiennes"

Résultats :

La Figure 4.13 [*respectivement* 4.14] trace l'AUC moyenne en fonction du nombre d'exemples étiquetés, pour le problème de l'échelon [*respectivement* le problème des gaussiennes]. Les deux courbes de chaque figure correspondent à l'approche supervisée MODL, munie, ou non, du post-traitement sur les bornes du $\mathcal{M}_{map}$. L'analyse des résultats obtenus est la même pour les deux problèmes jouets étudiés.

En considérant moins de 6 exemples étiquetés, les deux approches de discrétisation donnent un $\mathcal{M}_{map}$ comportant un seul intervalle. Dans ce cas, l'AUC observée est égale

à 0.5. Au delà de 6 exemples étiquetés, le $\mathcal{M}_{map}$ comporte deux intervalles. Quelque soit la méthode de discrétisation, la qualité du $\mathcal{M}_{map}$ s'améliore lorsque le nombre d'exemples étiquetés augmente. Dans les deux cas, la post-optimisation de la position de la borne améliore l'approche MODL. Cette amélioration est faible mais toujours présente. L'écart entre les deux courbes a tendance à diminuer lorsque le nombre d'exemples étiquetés augmente. Lors du post-traitement, la borne du $\mathcal{M}_{map}$ se déplace entre deux exemples étiquetés. L'erreur du modèle prédictif peut être réduite, dans le meilleur des cas, de l'ordre de $\frac{1}{N^l}$.

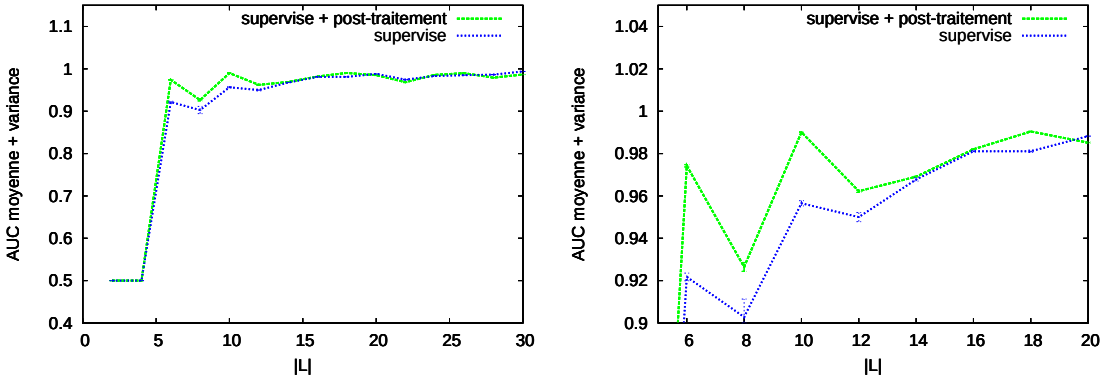

FIG. 4.13 – Échelon : AUC moyenne en fonction du nombre d'exemples étiquetés (le graphique de gauche fait un zoom sur les courbes).

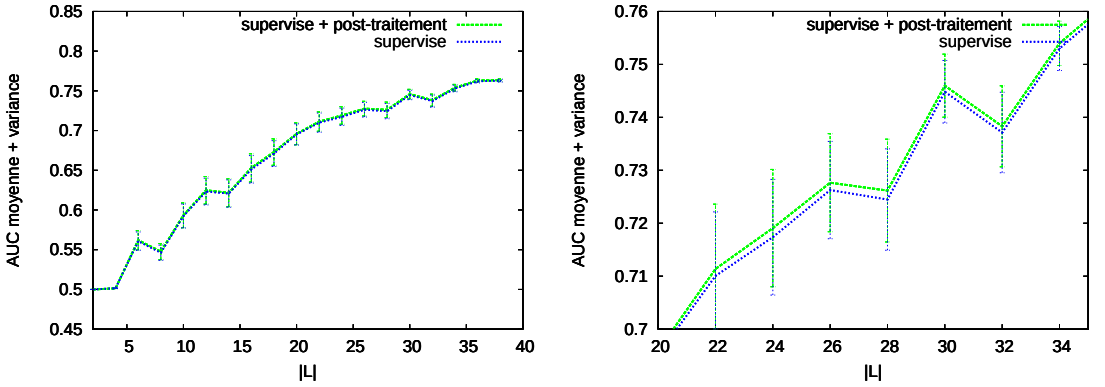

FIG. 4.14 – Gaussiennes : AUC moyenne en fonction du nombre d'exemples étiquetés (le graphique de gauche fait un zoom sur les courbes).

4.5 Discussion

Ce chapitre propose une extension de l'approche de discrétisation MODL au cas de l'apprentissage semi-supervisé. Notre méthode semi-supervisée exploite les mêmes choix de modélisation et les mêmes hypothèses que l'approche supervisée [Boullé, 2006b]. La seule différence est que notre approche manipule conjointement des exemples étiquetés et non-étiquetés. Tout d'abord, une famille de modèles de discrétisation est définie. Une démarche Bayésienne est ensuite appliquée, et conduit à un critère analytique dont l'optimisation désigne le modèle le plus probable connaissant les données.

Plusieurs résultats théoriques importants sont établis lors d'une étude approfondie de notre critère d'évaluation. Nous montrons que la prise en compte des exemples non-étiquetés dans la définition du prior $P(M)$ est inutile. Nous montrons également que notre approche semi-supervisée est asymptotiquement[6] équivalente à l'approche supervisée, munie d'un post-traitement sur la position des bornes du modèle optimal. Dans notre cas, le meilleur moyen d'exploiter les exemples non-étiquetés est d'utiliser l'approche supervisée, puis de placer les bornes du modèle optimal au milieu des zones non-étiquetées. Ce résultat montre que, malgré les hypothèses faiblement informatives adoptées sur les données, les exemples non-étiquetés apportent de l'information utile à la discrétisation.

Notre méthode de discrétisation semi-supervisée a été élaborée dans un cadre théorique restreint : i) des modèles de discrétisation incluant un ou deux intervalles ; ii) des données caractérisées par une seule variable explicative. Des travaux futurs pourraient étendre les démonstrations réalisées dans ce chapitre au cas des modèles multi-intervalles ($I > 2$) et au cas de données caractérisées par plusieurs variables ($dim(\mathbb{X}) > 1$). Notre approche a uniquement été testée sur des problèmes artificiels. Des expériences comparatives pourraient être menées sur des données réelles, pour déterminer si le post-traitement de la position des bornes du modèle optimal améliore toujours l'approche supervisée.

Les travaux réalisés dans ce chapitre sont exploitables pour l'amélioration de la stratégie d'apprentissage actif par modèles locaux définie au Chapitre 3. Les décisions relatives au partitionnement récursif de l'espace $\mathbb{X}$ peuvent être prises par la méthode de discrétisation semi-supervisée présentée dans ce chapitre. Cette méthode peut décider *où* et *quand* couper une zone, sans impliquer de paramètres utilisateur. Cette amélioration n'est pas réalisée dans le cadre de cette thèse et fait partie des futurs travaux envisageables.

Notre stratégie d'apprentissage actif par modèles locaux a pour objectif de définir les zones d'intérêt dans l'espace $\mathbb{X}$, où l'étiquetage de nouveaux exemples est le plus utile à l'apprentissage des modèles locaux. Les exemples étiquetés sont choisis de manière aléatoire dans les zones sélectionnées. Cette stratégie active peut être améliorée en sélectionnant, localement à la meilleure zone, l'exemple qui est le plus utile à l'apprentissage du modèle local. Le chapitre suivant propose une stratégie d'apprentissage actif fondée sur notre méthode de discrétisation semi-supervisée. Cette nouvelle stratégie active est plus performante que la dichotomie probabiliste [Horstein, 1963], lorsque celle-ci n'est pas correctement renseignée du bruit d'étiquetage présent dans les données. Ce résultat est très encourageant et ouvre nos travaux à d'autres heuristiques, qui exploitent elles aussi une méthode dichotomique et des données bruitées.

[6] Les approches supervisée et semi-supervisée convergent lorsque $N^l \to +\infty$ et $N^u \to +\infty$.

Chapitre 5

Apprentissage actif Bayésien

Sommaire

L'apprentissage actif par modèles locaux présenté au Chapitre 3 peut être amélioré en exploitant une stratégie active pour sélectionner, localement à la meilleure zone, l'exemple le plus utile à l'apprentissage du modèle local. Cette hybridation peut potentiellement être réalisée en exploitant n'importe quelle stratégie active de la littérature (Chapitre 2). Dans ce cinquième chapitre, nous proposons une stratégie active originale, adaptée au partitionnement récursif dichotomique de l'espace $\mathbb{X}$. Cette nouvelle stratégie d'apprentissage actif est fondée sur la méthode de discrétisation semi-supervisée présentée au Chapitre 4.

Nous formalisons notre stratégie active à la Section 5.1. Notre démarche aboutit à un critère analytique qui correspond à la probabilité des modèles de discrétisation connaissant les données, et un exemple candidat à l'étiquetage. La Section 5.2 étudie la complexité temporelle de notre approche. Nous proposons dans cette section des optimisations algorithmiques qui permettent d'exploiter notre stratégie en un temps raisonnable, dans le cas où : i) les modèles de discrétisation comportent un ou deux intervalles ; ii) le problème traité est une classification binaire ; iii) les données d'apprentissage sont caractérisées par une seule variable explicative. Ce cadre théorique est cohérent avec l'objectif de ce chapitre : l'utilisation de notre stratégie active dans le cadre du partitionnement récursif de $\mathbb{X}$. La Section 5.3 évalue notre stratégie active sur un problème jouet. Les expériences menées dans cette section caractérisent l'influence du bruit d'étiquetage présent dans les données, sur les performances de notre stratégie. Dans cette section, notre approche est comparée favorablement à une stratégie de référence : la dichotomie probabiliste [Horstein, 1963]. Ce résultat important ouvre nos travaux à d'autres heuristiques qui exploitent, elles aussi, une méthode dichotomique et des données bruitées.

5.1 Une nouvelle méthode d'apprentissage actif

Cette section présente une stratégie originale d'apprentissage actif fondée sur la méthode de discrétisation semi-supervisée définie au Chapitre 4. Les choix de modélisation adoptés à la Section 4.2.1 (page 71), sont conservés. À chaque itération d'un échantillonnage sélectif, notre stratégie sélectionne l'exemple qui, une fois étiqueté, maximisera l'espérance des modèles de $\mathbb{M}$. Nous exploitons le critère semi-supervisé $\mathcal{C}_{semi\ super}$ (Section 4.2.4, page 75) en tant qu'expression analytique de la distribution $P(M|D)$. Finalement, notre démarche aboutit à un critère dont l'optimisation désigne l'exemple $x_{t+1} \in U$ qui maximise l'espérance de $P(M|D, x_{t+1})$.

5.1.1 Formalisation

Soit $P_{(.|D)}(M) = P(M|D)$ la distribution a posteriori des modèles de discrétisation, connaissant les données. Notre stratégie active sélectionne l'exemple $x_{t+1} \in U$ qui maximise l'espérance de $P(M|D, x_{t+1})$ sur l'ensemble des modèles $M \in \mathbb{M}$:

$$\underset{x_{t+1}\in U}{ArgMax} \ \underset{M\in\mathbb{M}}{\mathrm{E}}_{P_{(.|D)}} [P(M|D, x_{t+1})]$$

$$\underset{x_{t+1}\in U}{ArgMax} \sum_{M\in\mathbb{M}} P(M|D) \times P(M|D, x_{t+1})$$

L'étiquette y_{t+1} est inconnue, mais la distribution $P(y|M, D, x_{t+1})$ de la classe $y \in \mathbb{Y}$ connaissant le modèle et les données, peut être évaluée. Grâce à la formule de la probabilité totale, nous écrivons :

1. $\sum_{y\in\mathbb{Y}} P(y|M, D, x_{t+1}) = 1$
2. $P(M|D, x_{t+1}) = \sum_{y\in\mathbb{Y}} P(y|M, D, x_{t+1}) P(M|D, x_{t+1}, y)$

Finalement,

$$\underset{x_{t+1}\in U}{ArgMax} \sum_{M\in\mathbb{M}} P(M|D) \times \left[\sum_{y\in\mathbb{Y}} P(y|M, D, x_{t+1}) \times P(M|D, x_{t+1}, y) \right]$$

Cette expression est ensuite développée en exploitant la formule de Bayes :

$$\underset{x_{t+1}\in U}{ArgMax} \sum_{M\in\mathbb{M}} \left[\frac{P(M) \times P(D|M)}{P(D)} \times \sum_{y\in\mathbb{Y}} \left[P\left(y|M, D, x_{t+1}\right) \times \frac{P(M) \times P(D, x_{t+1}, y|M)}{P(D, x_{t+1}, y)} \right] \right] \tag{5.1}$$

La distribution jointe $P(D, x_{t+1}, y)$ est décomposée comme suit :

$$P(D, x_{t+1}, y) = P(D) \times P(x_{t+1}|D) \times P(y|D, x_{t+1}) \tag{5.2}$$

L'exemple x_{t+1} est considéré comme étant tiré uniformément dans l'ensemble U, puisqu'a priori, tous les exemples non-étiquetés ont la même probabilité d'être sélectionnés.

Les termes $P(D)$ et $P(x_{t+1}|D)$ de l'Équation 5.1 restent donc constants lorsque que le modèle M varie. Cette équation se réécrit comme suit :

$$\underset{x_{t+1}\in U}{ArgMax} \sum_{M\in\mathbb{M}} \left[\overbrace{P(M)\times P(D|M)}^{A} \times \sum_{y\in\mathbb{Y}} \left[\underbrace{P(y|M,D,x_{t+1})}_{C} \times \frac{\overbrace{P(M)\times P(D,x_{t+1},y|M)}^{B}}{\underbrace{P(y|D,x_{t+1})}_{D}} \right] \right] \times Cste \tag{5.3}$$

Le terme "A" de l'Équation 5.3 est déduit du critère $\mathcal{C}_{semi\ super}$:

$$P(M)P(D|M) = \frac{1}{N} \times \frac{1}{C_{N+I-1}^{I-1}} \times \prod_{i=1}^{I} \frac{1}{C_{N_i+J-1}^{J-1}} \times \prod_{i=1}^{I} \left[\frac{\prod_{j=1}^{J} N_{ij}!}{N_i!} \times \frac{N_i^u!}{\prod_{j=1}^{J} N_{ij}^u!} \right]$$

Le terme "B" est calculé de la même façon, en ajoutant le couple (x_{t+1}, y) à l'ensemble d'apprentissage T.

Le terme "C" est évalué par la prédiction du modèle M. Le modèle courant estime la probabilité d'observer la classe y, connaissant l'exemple x_{t+1}. Cette prédiction se base sur la proportion des exemples étiquetés par la classe y, dans l'intervalle contenant l'exemple x_{t+1}.

Le terme "D" représente la probabilité d'observer la classe y, connaissant l'exemple x_{t+1} et les données. Ce terme est difficile à évaluer car il n'implique aucun modèle particulier. Pour évaluer ce terme, nous choisissons d'intégrer sur l'ensemble des modèles $M \in \mathbb{M}$. En exploitant la formule de la probabilité totale $[\sum_{M'\in\mathbb{M}} P(M'|D) = 1]$, nous pouvons écrire :

$$P(y|D,x_{t+1}) = \sum_{M'\in\mathbb{M}} P(M'|D) \times P(y|D,M',x_{t+1})$$

$$P(y|D,x_{t+1}) = \sum_{M'\in\mathbb{M}} \frac{P(D|M')P(M')}{P(D)} \times P(y|D,M',x_{t+1})$$

La probabilité des données $P(D)$ reste constante lorsque le modèle M varie. Finalement, l'espérance de $P(M|D,x_{t+1})$ est évaluée par le critère $\mathcal{C}_{actif}(x_{t+1})$:

$$\mathcal{C}_{actif}(x_{t+1}) = \sum_{M\in\mathbb{M}} \left[P(M)P(D|M) \times \sum_{y\in\mathbb{Y}} \left[\frac{P(y|M,D,x_{t+1}) \times P(M) \times P(D,x_{t+1},y|M)}{\sum_{M'\in\mathbb{M}} P(M') \times P(D|M') \times P(y|D,M',x_{t+1})} \right] \right] \tag{5.4}$$

Cette expression est normalisée grâce au maximum a posteriori, de manière à limiter les problèmes d'arrondis machine lors de son implémentation (voir annexe C). Notre stratégie sélectionne l'exemple $x_{t+1} \in U$ maximisant $\mathcal{C}_{actif}(x_{t+1})$.

5.1.2 Illustration par un problème jouet

Cette section étudie le comportement de notre stratégie active sur un problème jouet. L'expérience réalisée ici montre comment le critère $\mathcal{C}_{actif}(x_{t+1})$ sélectionne un exemple à étiqueter parmi les éléments de l'ensemble U.

Considérons un problème de classification binaire uni-varié, dont les exemples d'apprentissage sont caractérisés par une variable explicative comprise entre 0 et 1. Ce jeu de données comporte trois parties distinctes. La partie "A" de la Figure 5.1 contient 40 exemples régulièrement répartis dans l'intervalle $[0, 0.4]$. Parmi ces 40 exemples, 5 sont étiquetés par la classe "1" et sont régulièrement répartis dans l'intervalle $[0, 0.4]$. De la même façon, la partie "C" contient 35 exemples non-étiquetés et 5 exemples de la classe "2". La partie "B" contient 20 exemples non-étiquetés. La Figure 5.1 représente ce jeu de données.

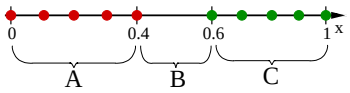

FIG. 5.1 – Problème jouet : ● exemple de classe "1", ● exemple de classe "2", les 90 exemples non-étiquetés ne sont pas représentés.

Dans le cadre de cette expérience, la famille de modèles $\mathbb{M}$ est restreinte aux modèles comportant un ou deux intervalles. L'espérance de $P(M|D, x_{t+1})$ est calculée pour chaque exemple non-étiqueté $x_{t+1} \in U$. La Figure 5.2 trace la valeur du critère $\mathcal{C}_{actif}(x_{t+1})$ (axe vertical) en fonction de la position de l'exemple x_{t+1} (axe horizontal). La valeur maximale de la courbe de la Figure 5.2 désigne l'exemple à étiqueter. Dans le cadre de ce problème jouet, notre stratégie sélectionne l'exemple qui se situe au milieu de la partie "B". L'expérience réalisée dans cette section montre que notre stratégie est valide et qu'elle est capable d'exprimer une préférence sur l'exemple à étiqueter.

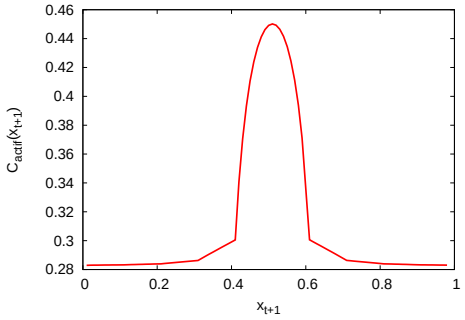

FIG. 5.2 – Espérance de $P(M|D, x_{t+1})$ en fonction de la position de x_{t+1}.

5.2 Complexité et optimisation

Cette section présente une étude sur la complexité temporelle de notre stratégie d'apprentissage actif. L'implémentation naïve de notre stratégie étant coûteuse en temps de calcul, nous proposons ensuite des optimisations algorithmiques, permettant d'exploiter notre stratégie en un temps raisonnable. Cette étude est restreinte au cas où : i) les modèles de discrétisation incluent au plus deux intervalles ; ii) les exemples d'apprentissage sont caractérisés par une seule variable explicative ; iii) le nombre de classes est fixé à deux. Ce cadre théorique est cohérent avec l'objectif de ce chapitre : l'utilisation de notre stratégie active dans le cadre du partitionnement dichotomique récursif de l'espace $\mathbb{X}$.

5.2.1 Complexité temporelle initiale

Nous calculons ici l'ordre de grandeur maximal de la complexité de notre stratégie active, lors de la sélection d'un seul exemple à étiqueter. Pour calculer l'espérance de $P(M|D, x_{t+1})$, la famille de modèles est parcourue deux fois. L'équation 5.4 comporte deux sommes sur $\mathbb{M}$. L'Algorithme 8 détaille le parcours exhaustif de $\mathbb{M}$, et met en jeu trois boucles imbriquées :

- parcours de toutes les coupures (N_1, N_2) possibles ;
- parcours de tous les effectifs (N_{11}, N_{12}) possibles, dans le premier intervalle ;
- parcours de tous les effectifs (N_{21}, N_{22}) possibles, dans le deuxième intervalle.

Notations :

- $M(I, \{N_i\}, \{N_{ij}\})$, le modèle courant
- $I = 2$
- $\{N_i\} = N_1, N_2$
- $\{N_{ij}\} = N_{11}, N_{12}, N_{21}, N_{22}$

Le parcours de $\mathbb{M}$ est défini comme suit :

Pour N_1 **de** 0 **à** N **faire**
  /*On a $N_2 = N - N_1$*/
  Pour N_{11} **de** N_{11}^l **à** $N_1 - N_{12}^l$ **faire**
    /*On a $N_{12} = N_1 - N_{11}$*/
    Pour N_{21} **de** N_{21}^l **à** $N_2 - N_{22}^l$ **faire**
      /*On a $N_{12} = N_1 - N_{11}$*/
      Parcours de tous les modèles à deux intervalles, grâce au modèle courant $M(I, \{N_i\}, \{N_{ij}\})$.
    Fin Pour
  Fin Pour
Fin Pour

Algorithme 8: Parcours des modèles en $\mathcal{O}(N^3)$

Dans le pire des cas, chaque boucle comporte N itérations. La complexité du parcours de la famille de modèles est donc de $\mathcal{O}(N^3)$. L'équation 5.5 illustre la complexité totale de

notre stratégie active. D'une part, les deux parcours de $\mathbb{M}$ sont imbriqués. D'autre part, le parcours des exemples non-étiquetés a une complexité de l'ordre de $\mathcal{O}(N)$. Finalement, la sélection de l'exemple $x_{t+1} \in U$ maximisant $\mathcal{C}_{actif}(x_{t+1})$ a une complexité de l'ordre de $\mathcal{O}(N^7)$. Malgré le cas très simple considéré ici ($I \leq 2$ et $J = 2$), la complexité temporelle de notre stratégie active est très élevée, lorsque celle-ci est implémentée naïvement.

$$\underbrace{\underset{x_{t+1} \in U}{ArgMax}}_{\mathcal{O}(N)} \underbrace{\sum_{M \in \mathbb{M}} \left[P(M)P(D|M) \times \sum_{y \in \mathbb{Y}} \left[\frac{P(y|M, D, x_{t+1}) \times P(M) \times P(D, x_{t+1}, y|M)}{\underbrace{\sum_{M' \in \mathbb{M}} P(M') \times P(D|M') \times P(y|D, M', x_{t+1})}_{\mathcal{O}(N^3)}} \right] \right]}_{\mathcal{O}(N^3)} \tag{5.5}$$

La suite de cette section présente plusieurs optimisations permettant de réduire la complexité de notre stratégie active. Les optimisations proposées ne dégradent pas le calcul de $\mathcal{C}_{actif}(x_{t+1})$, le parcours de $\mathbb{M}$ étant toujours exhaustif.

5.2.2 Optimisation par parallélisation

Nous exploitons ici l'additivité de l'Équation 5.4 et l'indépendance de certains termes, pour paralléliser les calculs nécessaires à l'évaluation du critère $\mathcal{C}_{actif}(x_{t+1})$.

Factorisation des calculs lors du parcours de $\mathbb{M}$:
Ce paragraphe montre comment le parcours exhaustif de la famille de modèles $\mathbb{M}$ peut être effectué une seule fois.

$$\mathcal{C}_{actif}(x_{t+1}) = \sum_{M \in \mathbb{M}} \left[P(M)P(D|M) \times \sum_{y \in \mathbb{Y}} \left[\underbrace{\frac{P(y|M, D, x_{t+1}) \times P(M) \times P(D, x_{t+1}, y|M)}{\sum_{M' \in \mathbb{M}} P(M') \times P(D|M') \times P(y|D, M', x_{t+1})}}_{A} \right] \right] \tag{5.6}$$

La partie "A" de l'équation 5.6 dépend uniquement de la classe $y \in \mathbb{Y}$ et de l'instance $x_{t+1} \in U$. Ce terme est indépendant du modèle M impliqué dans la première somme sur $\mathbb{M}$. L'Équation 5.6 peut être réorganisée de la manière suivante :

$$\mathcal{C}_{actif}(x_{t+1}) = \Bigg[\frac{\overbrace{\sum_{M \in \mathbb{M}} \boldsymbol{P(M)P(D|M)} \times P(y_1|M, D, x_{t+1}) \times P(M)P(D, x_{t+1}, y_1|M)}^{A}}{\underbrace{\sum_{M' \in \mathbb{M}} \boldsymbol{P(M')P(D|M')} \times P(y_1|D, M', x_{t+1})}_{B}} + \frac{\overbrace{\sum_{M \in \mathbb{M}} \boldsymbol{P(M)P(D|M)} \times P(y_2|M, D, x_{t+1}) \times P(M)P(D, x_{t+1}, y_2|M)}^{C}}{\underbrace{\sum_{M' \in \mathbb{M}} \boldsymbol{P(M')P(D|M')} \times P(y_2|D, M', x_{t+1})}_{D}} \Bigg] \tag{5.7}$$

Les termes "A", "B", "C" et "D" de l'équation 5.7 peuvent être calculés indépendamment les uns des autres. Étant donné leur caractère additif, ces quatre termes sont calculés en parallèle lors d'un unique parcours de la famille de modèles $\mathbb{M}$. La complexité temporelle de notre stratégie active se réduit alors à $\mathcal{O}(N^4)$.

Factorisation des calculs lors du parcours de l'ensemble U :
Ce paragraphe montre comment les termes du type "$P(M)P(D|M)$", représentés en **gras** dans l'équation 5.7, peuvent être calculés une seule fois. Ces termes sont indépendants de x_{t+1}, et sont pourtant recalculés pour chaque exemple non-étiqueté. Ces étapes inutiles sont évitées en parallélisant le calcul des termes "A", "B", "C" et "D", lors d'un unique parcours de l'ensemble U. Encore une fois, le caractère additif de l'équation 5.7 est exploité.

L'Algorithme (9) sélectionne l'exemple non-étiqueté maximisant $\mathcal{C}_{actif}(x_{t+1})$ en une complexité de l'ordre de $\mathcal{O}(N^4)$. Cet algorithme met en jeu quatre vecteurs de réels correspondant aux termes "A", "B", "C" et "D" de l'Équation 5.7. Pour chaque vecteur, chaque élément correspond à un exemple non-étiqueté particulier, caractérisé par l'indice $x_{ID} \in [1, N^u]$. Les éléments de ces quatre vecteurs sont initialisés à 0 et sont ensuite calculés en parallèle, lors du parcours de l'ensemble U. Les termes du type "$P(M)$" et "$P(M)P(D|M)$" ne sont calculés qu'une seule fois. Une étape finale est nécessaire pour calculer le critère $\mathcal{C}_{actif}(x_{t+1})$. Les parties "$A$", "$B$", "$C$" et "$D$" de l'équation 5.7 sont assemblées, pour chaque exemple $x_{t+1} \in U$.

Notations :

- $M(I, \{N_i\}, \{N_{ij}\})$, le modèle courant
- $I = 2$
- $\{N_i\} = N_1, N_2$
- $\{N_{ij}\} = N_{11}, N_{12}, N_{21}, N_{22}$
- 5 vecteurs de N^u réels initialisés à 0, dont chaque élément correspond à un exemple non-étiqueté particulier caractérisé par l'indice $x_{ID} \in [1, N^u]$. À l'issue de cet algorithme, on a :
  1. $A[x_{ID}] = P(M)P(D|M) \times P(y_1|M, D, x_{ID}) \times P(M)P(D, x_{ID}, y_1|M)$
  2. $B[x_{ID}] = P(M')P(D|M') \times P(y_1|D, M', x_{ID})$
  3. $C[x_{ID}] = P(M)P(D|M) \times P(y_2|M, D, x_{ID}) \times P(M)P(D, x_{ID}, y_2|M)$
  4. $D[x_{ID}] = P(M')P(D|M') \times P(y_2|D, M', x_{ID})$
  5. $\mathcal{C}_{actif}[x_{ID}] = \mathrm{E}_{P_{(.|D)}\, M \in \mathbb{M}}\ [P(M|D, x_{ID})]$

Pour N_1 **de** 0 **à** N **faire**
  */*On a $N_2 = N - N_1$ */*
  Pour N_{11} **de** N_{11}^l **à** $N_1 - N_{12}^l$ **faire**
    */*On a $N_{12} = N_1 - N_{11}$ */*
    Pour N_{21} **de** N_{21}^l **à** $N_2 - N_{22}^l$ **faire**
      */*On a $N_{12} = N_1 - N_{11}$ */*

      */*Calcul de $P(M)$ et $P(M)P(D|M)$ */*
      $ProbModel = P(M)$
      $ProbModelData = P(M)P(D|M)$

      Pour x_{ID} **de** 0 **à** N **faire**
        */*Pour chaque instance $x_{t+1} \in U$ */*
        $A[x_{ID}] \leftarrow^+ ProbModelData.P(y_1|M, D, x_{t+1}).ProbModel.P(D, x_{t+1}, y_1|M)$
        $B[x_{ID}] \leftarrow^+ ProbModelData.P(y_1|D, M, x_{t+1})$
        $C[x_{ID}] \leftarrow^+ ProbModelData.P(y_2|M, D, x_{t+1}).ProbModel.P(D, x_{t+1}, y_2|M)$
        $D[x_{ID}] \leftarrow^+ ProbModelData.P(y_2|D, M, x_{t+1})$
      Fin Pour
    Fin Pour
  Fin Pour
Fin Pour

*/*Sélection de l'exemple $q \in U$ qui maximise $\mathcal{C}_{actif}$ */*
$Best\mathcal{C}_{actif} \leftarrow 0$
$q \leftarrow \emptyset$

Pour x_{ID} **de** 0 **à** N **faire**
  */*Calcul final de l'espérance, pour chaque instance $x_{t+1} \in U$ */*
  $\mathcal{C}_{actif}[x_{ID}] \leftarrow \frac{A[x_{ID}]}{B[x_{ID}]} + \frac{C[x_{ID}]}{D[x_{ID}]}$

  */*Test de l'exemple courrant*/*
  Si $\mathcal{C}_{actif}[x_{ID}] > Best\mathcal{C}_{actif}$ **Alors**
    $Best\mathcal{C}_{actif} \leftarrow \mathcal{C}_{actif}[x_{ID}]$
    $q \leftarrow x_{ID}$
  Fin Si
Fin Pour

Algorithme 9: Calcul de $\mathcal{C}_{actif}(x_{t+1}) \quad \forall x_{t+1} \in U$ en $\mathcal{O}(N^4)$

5.2.3 Optimisation du calcul de $P(M)P(D, x_{t+1}, y|M)$

Le critère $\mathcal{C}_{semi\ super}$ exprime la quantité d'information d'un modèle conditionnellement aux données, notée $I(M|D)$ (Section 4.2.4, page 75). Ce critère est exploité par l'Algorithme 9 pour calculer $P(M)P(D|M) = e^{-I(M|D)}$. Les termes "$P(M)P(D, x_{t+1}, y_1|M)$" et "$P(M)P(D, x_{t+1}, y_2|M)$" sont calculés pour chaque modèle $M \in \mathbb{M}$ et pour chaque exemple supplémentaire $x_{t+1} \in U$, muni de l'étiquette y_1 ou y_2. Nous montrons que les termes "$P(M)P(D, x_{t+1}, y_1|M)$" et "$P(M)P(D, x_{t+1}, y_2|M)$" peuvent se déduire des calculs effectués pour le terme "$P(M)P(D|M)$". Nous démontrons que $I(M|D, x_{t+1}, y) = \Delta + I(M|D)$. Selon l'exemple supplémentaire x_{t+1}, le terme Δ peut prendre quatre valeurs différentes :

- L'exemple x_{t+1} appartient au premier intervalle du modèle M et $y = y_1$:

$$I(M|D, x_{t+1}, y_1) = \log\left(\frac{N_1^h}{N_{11}^h}\right) + I(M|D)$$

- L'exemple x_{t+1} appartient au premier intervalle du modèle M et $y = y_2$:

$$I(M|D, x_{t+1}, y_2) = \log\left(\frac{N_1^h}{N_{12}^h}\right) + I(M|D)$$

- L'exemple x_{t+1} appartient au deuxième intervalle du modèle M et $y = y_1$:

$$I(M|D, x_{t+1}, y_1) = \log\left(\frac{N_2^h}{N_{21}^h}\right) + I(M|D)$$

- L'exemple x_{t+1} appartient au deuxième intervalle du modèle M et $y = y_2$:

$$I(M|D, x_{t+1}, y_1) = \log\left(\frac{N_2^h}{N_{22}^h}\right) + I(M|D)$$

Démonstration.

Cette démonstration se restreint au cas de modèles à deux intervalles ($I = 2$), et à des problèmes de classification binaire ($J = 2$).

$$I(M|D) = \log(N) + \log(C_{N+I-1}^{I-1}) + \log\left(\frac{(N_1+1)!}{N_1^h!}\right) + \log\left(\frac{(N_2+1)!}{N_2^h!}\right)$$
$$+ \log\left(\frac{N_{11}^h!}{N_{11}!}\right) + \log\left(\frac{N_{12}^h!}{N_{12}!}\right) + \log\left(\frac{N_{21}^h!}{N_{21}!}\right) + \log\left(\frac{N_{22}^h!}{N_{22}!}\right)$$

Dans un premier temps, on suppose que l'exemple supplémentaire x_{t+1} est dans le **premier intervalle** du modèle M et que cette instance est **étiquetée par la classe y_1**. Trois paramètres varient par rapport au calcul de $I(M|D)$:

$N_1^h \leftarrow N_1^h - 1$
$N_{11}^l \leftarrow N_{11}^l + 1$
$N_{11}^h \leftarrow N_{11}^h - 1$

Dans ces conditions :

$$I(M|D, x_{t+1}, y_1) = \log(N) + \log(C_{N+I-1}^{I-1}) + \log\left(\frac{(N_1+1)!}{\boldsymbol{(N_1^h - 1)!}}\right) + \log\left(\frac{(N_2+1)!}{N_2^h!}\right)$$
$$+ \log\left(\frac{\boldsymbol{(N_{11}^h - 1)!}}{N_{11}!}\right) + \log\left(\frac{N_{12}^h!}{N_{12}!}\right) + \log\left(\frac{N_{21}^h!}{N_{21}!}\right) + \log\left(\frac{N_{22}^h!}{N_{22}!}\right)$$

Soit :
$\Delta = I(M|D, x_{t+1}, y_1) - I(M|D)$

$\Delta = \log\left(\frac{N_1^h! \times (N_{11}^h - 1)!}{(N_1^h - 1)! \times N_{11}^h!}\right)$

$\Delta = \log\left(\frac{N_1^h}{N_{11}^h}\right)$

La démonstration est similaire pour les trois autres cas.

□

Les quatres situations possibles sont pré-calculées avant de parcourir l'ensemble des exemples non-étiquetés. Pour chaque situation, le calcul de Δ se substitue à l'évaluation du critère $\mathcal{C}_{semi\ super}$, ce qui est très avantageux d'un point de vue algorithmique. Le calcul des termes "$P(M)P(D, x_{t+1}, y_1|M)$" et de "$P(M)P(D, x_{t+1}, y_2|M)$" se réduit au parcours de l'ensemble U et au test déterminant l'intervalle auquel appartient l'exemple courant $x_{t+1} \in U$. Cette optimisation réduit la complexité de notre stratégie active d'un terme constant.

5.2.4 Optimisation du parcours des modèles

L'Algorithme (9) calcule $P(M)P(D|M) = e^{-I(M|D)}$ pour chaque modèle $M \in \mathbb{M}$. L'évaluation du terme "$I(M|D)$" est relativement coûteuse en temps de calcul, le critère $\mathcal{C}_{semi\ super}$ impliquant des factoriels. Nous montrons ici que le terme "$I(M|D)$" peut être déduit des calculs effectués pour le modèle précédent.

Boucle sur les (N_{21}, N_{22}) :
Trois boucles imbriquées sont mises en jeu dans l'Algorithme (9). Ici, nous considérons la boucle faisant varier les paramètres (N_{21}, N_{22}). Le modèle courant est noté M_c et le prochain modèle de la boucle est noté M_n. Nous démontrons la relation suivante :

$$I(M_n|D) = \log\left(\frac{(N_{21}^h + 1) \times N_{22}}{(N_{21} + 1) \times N_{22}^h}\right) + I(M_c|D)$$

Boucle sur les (N_{11}, N_{12}) :
Ici, nous considérons la boucle faisant varier les paramètres (N_{11}, N_{12}). Comme précédemment, nous démontrons la relation suivante :

$$I(M_n|D) = \log\left(\frac{(N_{11}^h + 1) \times N_{12}}{(N_{11} + 1) \times N_{12}^h}\right) + I(M_c|D)$$

Démonstration.

Soient M_c le modèle courant et M_n le prochain modèle de la boucle faisant varier les paramètres (N_{21}, N_{22}) :

$$I(M_c|D) = \log(N) + \log(C_{N+I-1}^{I-1}) + \log\left(\frac{(N_1+1)!}{N_1^h!}\right) + \log\left(\frac{(N_2+1)!}{N_2^h!}\right)$$
$$+ \log\left(\frac{N_{11}^h!}{N_{11}!}\right) + \log\left(\frac{N_{12}^h!}{N_{12}!}\right) + \log\left(\frac{N_{21}^h!}{N_{21}!}\right) + \log\left(\frac{N_{22}^h!}{N_{22}!}\right)$$

Entre deux itérations de la boucle, le modèle courant subit les modifications suivantes :

$N_{21} \leftarrow N_{21} + 1$
$N_{21}^h \leftarrow N_{21}^h + 1$
$N_{22} \leftarrow N_{22} - 1$
$N_{22}^h \leftarrow N_{22}^h - 1$

Nous pouvons calculer le terme $I(M_n|D)$:

$$I(M_n|D) = \log(N) + \log(C_{N+I-1}^{I-1}) + \log\left(\frac{(N_1+1)!}{N_1^h!}\right) + \log\left(\frac{(N_2+1)!}{N_2^h!}\right)$$
$$+ \log\left(\frac{N_{11}^h!}{N_{11}!}\right) + \log\left(\frac{N_{12}^h!}{N_{12}!}\right) + \log\left(\frac{(N_{21}^h+1)!}{(N_{21}+1)!}\right) + \log\left(\frac{(N_{22}^h-1)!}{(N_{22}-1)!}\right)$$

Soit :
$\Delta = I(M_n|D) - I(M_c|D)$

$\Delta = \log\left(\frac{(N_{21}^h+1)!\times N_{21}!\times(N_{22}^h-1)!\times N_{22}!}{(N_{21}+1)!\times N_{21}^h!\times(N_{22}-1)!\times N_{22}^h!}\right)$

$\Delta = \log\left(\frac{(N_{21}^h+1)\times N_{22}}{(N_{21}+1)\times N_{22}^h}\right)$

La démonstration pour la boucle sur les (N_{11}, N_{12}) est similaire.

□

L'optimisation du parcours des modèles réduit les calculs nécessaires à l'évaluation du terme "$I(M|D)$", pour l'ensemble des modèles $M \in \mathbb{M}$. Finalement, la complexité de notre stratégie d'apprentissage actif reste en $\mathcal{O}(N^4)$, mais est réduite d'un facteur constant. En prenant en compte les optimisations présentées dans les Sections 5.2.3 et 5.2.4, le facteur constant mesuré en pratique est supérieur à 5000.

5.3 Évaluation

Disposant d'un algorithme efficace pour l'évaluation de $\mathcal{C}_{actif}(x_{t+1})$, nous pouvons désormais mettre en œuvre notre stratégie active. Nous choisissons d'appliquer notre stratégie active à l'estimation d'une fonction échelon à partir de données bruitées [Castro and Nowak, 2008], en raison du cadre théorique adopté dans ce chapitre ($I \leq 2$ et $J = 2$). Tout d'abord, nous définissons le problème d'apprentissage utilisé lors de nos expériences. Une stratégie concurrente est ensuite présentée : la dichotomie probabiliste. Cette approche est une stratégie de référence pour l'estimation d'une fonction échelon à partir de données bruitées [Horstein, 1963]. Les expériences réalisées dans cette section produisent deux types de résultats. Les résultats illustratifs montrent le comportement de notre stratégie active, au cours des itérations successives d'un échantillonnage sélectif. Les résultats comparatifs évaluent la performance de chaque stratégie, en fonction du nombre d'exemples étiquetés. Les expériences menées dans cette section cherchent à caractériser l'influence du bruit présent dans les données, sur les performances des stratégies actives.

5.3.1 Jeu de données

Le jeu de données utilisé lors de nos expériences comporte 100 exemples uniformément distribués dans l'intervalle $[0, 1]$. Le domaine de définition de la variable explicative x est divisé en deux intervalles $[0, \theta[$ et $[\theta, 1]$, avec θ la position de l'échelon. L'objectif de ce problème d'apprentissage est d'évaluer θ à partir de données bruitées. Les exemples d'apprentissage de l'intervalle $[0, \theta]$ [*respectivement* $]\theta, 1]$] sont majoritairement étiquetés par la classe "1" [*respectivement* par la classe "2"]. La probabilitié que l'étiquette d'un exemple soit erronée est constante sur l'intervalle $[0, 1]$ et est notée $p \in [0, 0.5]$. Lorsque $p = 0$, les exemples d'apprentissage forment un échelon pur. Lorsque $p = 0.5$, les deux classes sont complètement mélangées, l'estimation de θ est alors impossible. La Figure 5.3 illustre ce jeu de données et trace la probabilité d'observer la classe "1", connaissant la valeur de la variable explicative x.

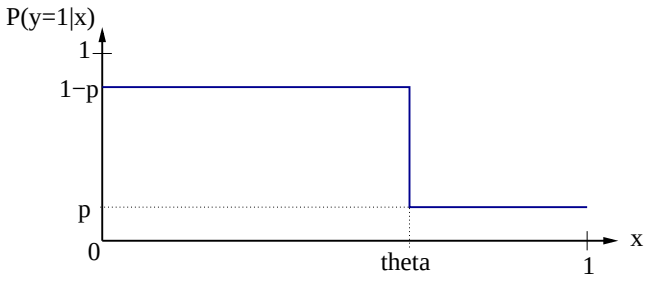

FIG. 5.3 – Échelon bruité : cette figure représente la probabilité d'observer la classe "1", connaissant la valeur de la variable explicative x.

5.3.2 Stratégies concurrentes

Notre stratégie active est comparée à deux autres stratégies de littérature : i)l'échantillonnage aléatoire ; ii) la dichotomie probabiliste.

Échantillonnage aléatoire

L'échantillonnage aléatoire sélectionne les exemples uniformément, selon leur distribution de probabilité. Cette stratégie joue un rôle de référence et est utilisée pour mesurer l'apport de notre approche basée sur la maximisation de l'espérance des modèles de discrétisation.

Dichotomie probabiliste

Notations :

- $P_\theta(x)$ la distribution de probabilité a priori de θ.
- p la probabilité qu'une étiquette soit erronée.
- Un échelon bruité, tel que l'intervalle $[0, \theta]$ [*respectivement* $[\theta, 1]$] soit majoritairement constitué d'exemples de la classe "1" [*respectivement* de la classe "2"].
- n le nombre d'exemples d'apprentissage souhaité.
- Les ensembles U et L d'exemples non étiquetés et étiquetés, avec $L \cup U = \Phi$
- L'ensemble d'apprentissage T avec $|T| < n$

*/*initialisation des ensembles L et U*/*
$L = \emptyset$ et $U = \Phi$

*/*initialisation de $P_\theta(x)$*/*
(I) $P_\theta(x) \leftarrow \frac{1}{N} \quad \forall x \in \Phi$

Répéter
  (A) Trouver l'exemple non-étiqueté $x^* \in U$ tel que : $\sum_{x \in [0, x^*]} P_\theta(x) = \frac{1}{2}$
  (B) Demander l'étiquette $f(x^*)$, ajouter $(x^*, f(x^*))$ à T et x^* à L, retirer x^* de U.
  (C) Mettre à jour $P_\theta(x)$ tel que :
  Si $f(x^*) = 1$ **Alors**
    (i) $P_\theta(x) \leftarrow 2.p.P_\theta(x) \qquad \forall x \in [0, x^*]$
    (ii) $P_\theta(x) \leftarrow 2.(1-p).P_\theta(x) \quad \forall x \in]x^*, 1]$
  Fin Si
  Si $f(x^*) = 2$ **Alors**
    (iii) $P_\theta(x) \leftarrow 2.(1-p).P_\theta(x) \quad \forall x \in [0, x^*]$
    (iv) $P_\theta(x) \leftarrow 2.p.P_\theta(x) \qquad \forall x \in]x^*, 1]$
  Fin Si
Tant que $|T| < n$

Algorithme 10: Dichotomie probabiliste

Dans le cas où $p = 0$, la dichotomie "classique" peut être utilisée pour trouver la position θ en étiquetant un minimum d'exemples. À chaque itération d'un échantillonnage sélectif, l'exemple qui se situe à mi-chemin des étiquettes "1" et "2" les plus proches est sélectionné. Selon la valeur de la nouvelle étiquette, θ se situe à droite ou à gauche du dernier exemple étiqueté. L'espace de recherche de θ est réduit environ de moitié à chaque itération. Dans le cas où $p > 0$, la dichotomie classique ne converge pas sur la position θ, certains exemples étant étiquetés par la mauvaise classe.

M. Horstein généralise la stratégie dichotomique au cas de données bruitées [Horstein, 1963]. Nous appelons cette stratégie : la dichotomie probabiliste. Cette approche suppose que le niveau de bruit p est connu. La distribution de probabilité a priori de θ est définie, et est notée $P_\theta(x)$.

L'Algorithme 10 présente la dichotomie probabiliste. Au début de l'algorithme, aucune étiquette n'est disponible. La première étape (I) initialise la distribution $P_\theta(x)$ selon un a priori uniforme. À chaque itération de l'Algorithme 10, l'étape A désigne l'exemple à étiqueter. L'exemple non-étiqueté $x^* \in U$ qui "coupe" la distribution $P_\theta(x)$ en deux parties égales est sélectionné. x^* est l'exemple le plus proche de la médiane de $P_\theta(x)$, ce qui s'écrit : $\sum_{x \in [0,x^*]} P_\theta(x) = \frac{1}{2}$. L'étiquette $f(x^*)$ est ensuite demandée à un expert (étape B). Le couple $(x^*, f(x^*))$ est ajouté à l'ensemble d'apprentissage T. La distribution $P_\theta(x)$ est mise à jour à chaque itération, selon la valeur de la nouvelle étiquette $f(x^*)$ (étape C). Cette étape prend également en considération le niveau de bruit p.

5.3.3 Résultats illustratifs

Cette section illustre le comportement de notre stratégie active sur le problème d'apprentissage défini à la Section 5.3.1. Les expériences réalisées montrent, pour chaque exemple $x_{t+1} \in U$, l'évolution de $\mathcal{C}_{actif}(x_{t+1})$ au cours des itérations successives d'un échantillonnage sélectif. Les exemples sélectionnés, ainsi que leurs étiquettes, sont représentés. Deux cas sont traités ici : i) l'échelon pur ($p = 0$) ; ii) l'échelon bruité ($p > 0$). Dans les deux cas la position de l'échelon est fixé à $\theta = 0.5$.

Échelon pur :
La Figure 5.4 détaille la sélection d'exemples, lors des premières itérations d'un échantillonnage sélectif. Les courbes représentent[1] $\mathcal{C}_{actif}(x_{t+1})$ (axe vertical), en fonction de la position de l'exemple supplémentaire x_{t+1} (axe horizontal). La valeur maximum de chaque courbe est symbolisée par un petit triangle rouge, et correspond à la position de l'exemple sélectionné à chaque itération. Les exemples de la classe "1" [*respectivement* "2"] sont symbolisés par des points verts [*respectivement* bleus]. Au début de l'échantillonnage sélectif, aucun exemple n'est étiqueté. Lors de la première itération (graphique "A" de la Figure 5.4), le critère $\mathcal{C}_{actif}(x_{t+1})$ est maximal pour $x_{t+1} = 0$ et $x_{t+1} = 1$. En cas d'égalité, un des exemples est choisi de manière aléatoire. Finalement, l'exemple $x_{t+1} = 1$ est étiqueté par la classe "2". Lors de la deuxième itération (graphique "B" de la Figure 5.4), l'exemple $x_{t+1} = 0$ est étiqueté par la classe "1". La courbe de la troisième itération semble très plate (graphique "C" de la Figure 5.4). Le critère $\mathcal{C}_{actif}(x_{t+1})$ est pourtant maximal en deux points, la Figure 5.5 fait un zoom sur cette courbe et montre l'emplacement des deux exemples candidats à l'étiquetage. Finalement, l'exemple $x_{t+1} = 0.28$ est sélectionné et étiqueté par la classe "1". À partir de 2 exemples étiquetés dans chaque classe, notre stratégie adopte un comportement proche de la dichotomie. Sur les graphiques "E, F, G, H" de la Figure 5.4, les exemples sélectionnés se situent quasiment à mi-chemin des étiquettes "1" et "2" les plus proches. Notre stratégie converge à 0.5, après avoir étiqueté 9 exemples.

[1] L'espérance est normalisée de telle sorte que son maximum soit égale à 1.

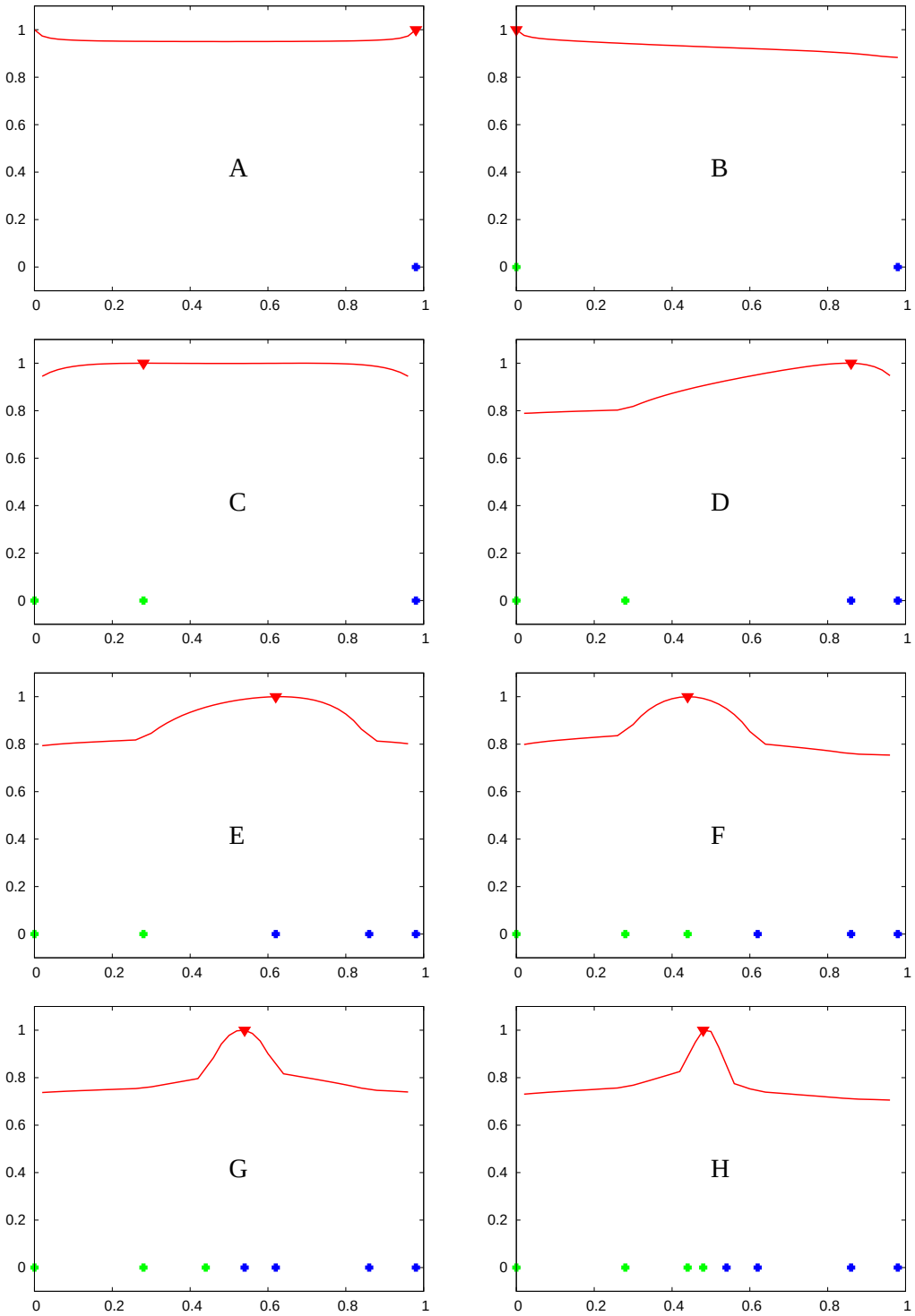

FIG. 5.4 – Visualisation des exemples étiquetés pour l'échelon pur. L'axe vertical représente $\mathcal{C}_{actif}(x_{t+1})$ et l'axe horizontal correspond à position de x_{t+1}.

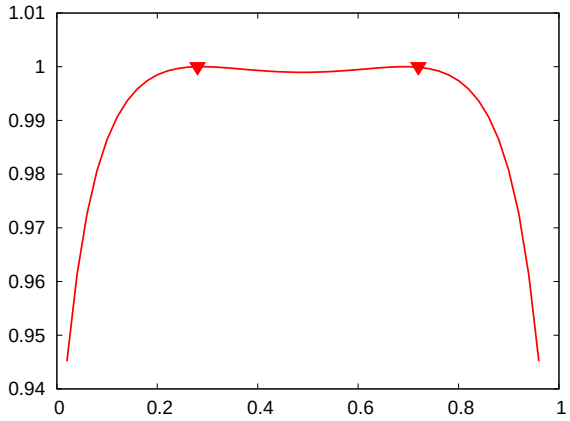

FIG. 5.5 – Zoom sur l'itération "C" de la Figure 5.4.

Échelon bruité :
La Figure 5.6 présente un scénario d'étiquetage impliquant un exemple dont l'étiquette est erronée. Comme précédemment, les courbes tracent les valeurs du critère $\mathcal{C}_{actif}(x_{t+1})$, en fonction de la position de l'exemple supplémentaire x_{t+1}. Le graphique "A" de la Figure 5.6 correspond à la quatrième itération de l'échantillonnage sélectif. L'exemple maximisant l'espérance de $P(M|D, x_{t+1})$ est étiqueté de manière incorrecte. Cet exemple bruité est symbolisé par un carré vert ($x_{t+1} = 0.86$). Lors des trois itérations suivantes (graphiques "B, C, D" de la Figure 5.6), les exemples sélectionnés se situent à droite de l'exemple bruité. Ce comportement est correct car pour le moment, rien ne laisse présager du bruit présent dans les données. Notre stratégie cherche à affiner ce qu'elle croit être la position de l'échelon. Au cours de ces trois itérations, les exemples sélectionnés se rapprochent de l'exemple bruité. Sur le graphique "D" de la Figure 5.6, deux exemples contigus sont étiquetés par des classes différentes ($x_{t+1} = 0.86$ et $x_{t+1} = 0.88$). À la 8ème itération (graphique "E" de la Figure 5.6), un exemple de la classe "2" est étiqueté juste à gauche de l'exemple bruité. À partir de cette itération, l'exemple bruité est détecté. Lors des itérations suivantes (graphiques "F, G, H" de la Figure 5.6), notre stratégie continue efficacement sa recherche dans le reste de l'espace. Le même type de comportement que précédemment est observé (Figure 5.4). Cette expérience montre que notre stratégie se comporte correctement face à des données bruitées. La position exacte du front montant de l'échelon est déterminée en étiquetant 12 exemples.

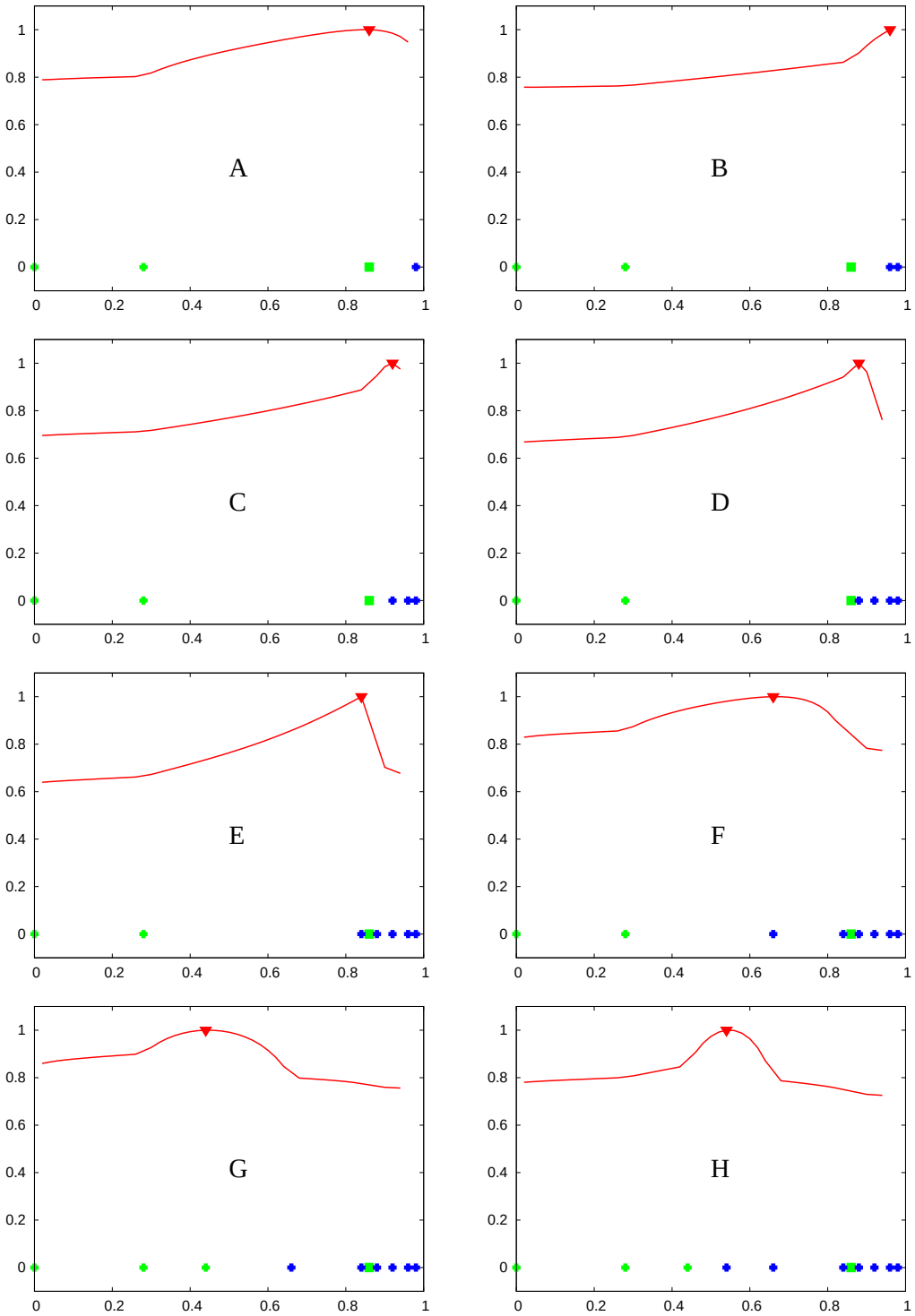

FIG. 5.6 – Visualisation des exemples étiquetés pour l'échelon bruité. L'axe vertical représente $\mathcal{C}_{actif}(x_{t+1})$ et l'axe horizontal correspond à position de x_{t+1}.

5.3.4 Résultats comparatifs

Cette section présente plusieurs séries d'expériences comparatives, réalisées sur le problème d'apprentissage présenté à la Section 5.3.1. La position de l'échelon est fixée à $\theta = 0.675$. L'objectif de ces expériences est de caractériser l'influence du bruit d'étiquetage p sur les performances des stratégies actives. Lors de nos expériences, p varie entre 0 et 0.20. Les stratégies actives évaluées ici sont présentées à la Section 5.3.2.

Évaluation : deux modèles prédictifs possibles
L'évaluation des stratégies actives doit dépendre uniquement de la qualité des exemples étiquetés. Toutes les stratégies sont évaluées en utilisant le même modèle prédictif. Seule la sélection des exemples étiquetés a une influence sur la performance du modèle prédictif. Les expériences présentées dans cette section impliquent deux modèles prédictifs :

A. le maximum a posteriori défini à la Section 4.2.4, qui inclut un ou deux intervalles ;
B. un modèle *"dichotomique"*, qui comporte toujours deux intervalles et dont la frontière sépare la distribution $P_\theta(x)$ en deux parties équiprobables.

Dans le premier cas, le motif à découvrir n'est pas supposé être un échelon ($I \leq 2$). Le modèle prédictif a le choix de couper, ou non, le domaine de définition de la variable explicative x. Le modèle prédictif évalue la position de la coupure, s'il y en a une. Dans le deuxième cas, les données sont supposées être organisées selon deux intervalles ($I = 2$). Le modèle prédictif cherche uniquement à évaluer la position de l'échelon.

La qualité des stratégies actives est évaluée grâce à l'AUC théorique du modèle prédictif, en fonction du nombre d'exemples étiquetés. Nous calculons l'AUC théorique en expoitant la position de l'échelon et la position de la borne prédite par le modèle de discrétisation (Annexe A.d). Il s'agit d'un calcul exact qui ne dépend pas d'un ensemble de test.

Protocole expérimental :
Le protocole expérimental adopté est le suivant : i) les expériences sont répétées 100 fois, de manière à obtenir une AUC moyenne et une mesure de variance pour chaque valeur de $|L|$; ii) au début de chaque expérience aucun exemple n'est étiqueté, $L = \emptyset$; iii) à chaque itération de l'échantillonnage sélectif, un exemple est sélectionné puis étiqueté ; iv) les expériences s'arrêtent lorsque $|L| = 20$.

Trois cas possibles pour la dichotomie probabiliste :
À la différence des autres stratégies évaluées, la dichotomie probabiliste est renseignée du niveau de bruit p. Lors de nos expériences, trois cas sont envisagés :

a. la dichotomie probabiliste est correctement renseignée du niveau bruit p, ce dernier variant dans l'intervalle $[0, 0.20]$;
b. la dichotomie probabiliste est renseignée par un niveau de bruit nul, noté $p_{false} = 0$, alors que le bruit d'étiquetage p varie dans l'intervalle $[0, 0.20]$;
c. la dichotomie probabiliste est renseignée par une valeur erronée du bruit, tel que $p_{false} \in [0, 0.20]$, le bruit d'étiquetage p étant nul.

A Évaluation : le modèle prédictif employé est le $\mathcal{M}_{map}$

Ici, les stratégies sont évaluées grâce au maximum a posteriori défini à la Section 4.2.4. La famille de modèles $\mathbb{M}$ est restreinte au cas où $I \leq 2$.

A.a Cas où la dichotomie probabiliste est correctement renseignée du niveau bruit :

La Figure 5.7 trace l'AUC moyenne observée sur 100 expériences (axe vertical) en fonction du nombre d'exemples étiquetés (axe horizontal). Les moustaches sur les courbes représentent la variance des résultats ($\pm 2\sigma$). Les graphiques de la Figure 5.7 correspondent à quatre niveaux de bruit différents.

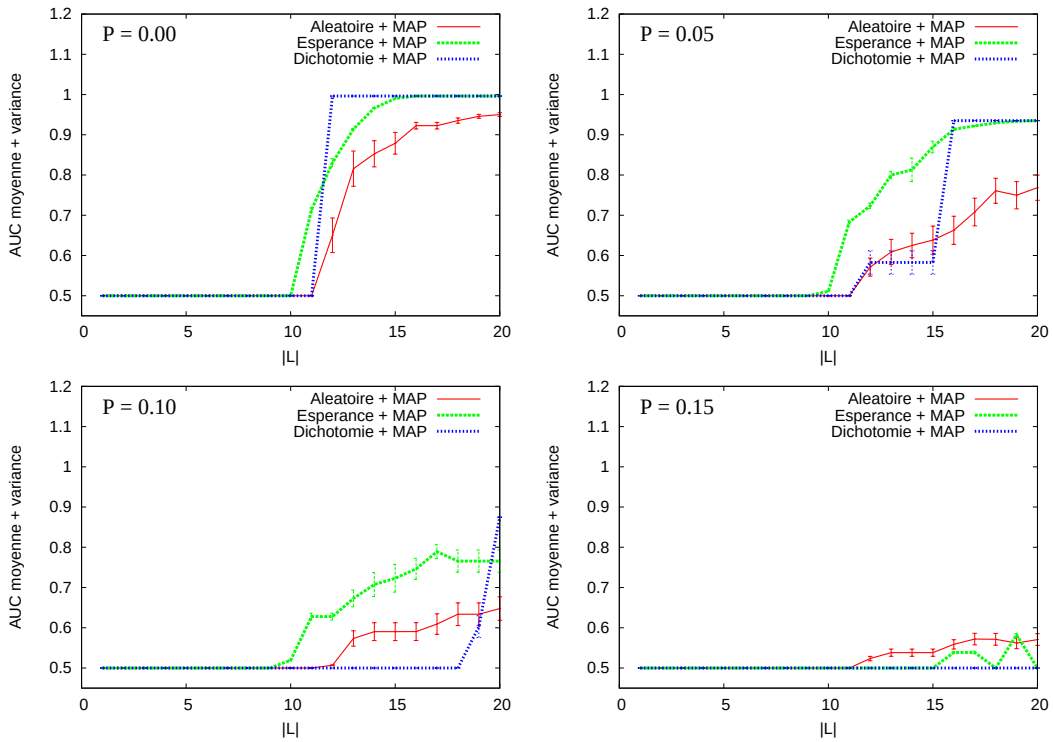

FIG. 5.7 – Performance moyenne des stratégies actives *vs.* le nombre d'exemples étiquetés. Cas où la dichotomie probabiliste est correctement renseignée du niveau de bruit

Considérons dans un premier temps le cas où le bruit d'étiquetage est nul ($p = 0$). La stratégie aléatoire présente une AUC constante, égale à 0.5, lorsque le nombre d'exemples étiquetés est inférieur à 12 (courbe rouge). Dans ce cas, le maximum a posteriori comporte un seul intervalle et estime les distributions conditionnelles aux classes de manière uniforme. Lorsque suffisamment d'exemples sont étiquetés pour produire un $\mathcal{M}_{map}$ à deux intervalles ($|L| > 12$), la stratégie aléatoire progresse et atteint une AUC de 0.95 à la fin de l'expérience.

Notre stratégie basée sur la maximisation de l'espérance de $P(M|D, x_{t+1})$ présente deux caractéristiques intéressantes (courbe verte) : i) cette stratégie est plus performante

que l'aléatoire quelque soit le nombre d'exemples étiquetés ; ii) cette stratégie requiert moins d'exemples étiquetés que l'aléatoire pour produire un modèle optimal à deux intervalles (l'AUC est supérieure à 0.5, pour $|L| > 10$). Finalement, notre stratégie se comporte honorablement et atteint la performance optimale (AUC=1) après avoir étiqueté 15 exemples.

Comme la stratégie aléatoire, la dichotomie probabiliste présente une AUC égale à 0.5 lorsque le nombre d'exemples étiquetés est inférieur à 12 (courbe bleue). Cette stratégie progresse ensuite très rapidement, et atteint la performance optimale dès que le $\mathcal{M}_{map}$ inclue deux intervalles ($|L| = 13$). Dans le cas d'un bruit d'étiquetage nul, la dichothomie probabiliste est la stratégie la plus performante.

Lorsque le niveau de bruit augmente ($p = 0.05$, $p = 0.10$, $p = 0.15$), les performances des trois stratégies diminuent. Notre stratégie reste plus performante que la stratégie aléatoire, pour $p = 0.05$ et $p = 0.10$. La dichotomie probabiliste se dégrade très nettement lorsque p croît. Plus le bruit d'étiquetage augmente, plus le nombre d'exemples étiquetés nécessaires à l'obtention d'un $\mathcal{M}_{map}$ à deux intervalles est important.

Les expériences réalisées dans ce paragraphe montrent que notre stratégie active est plus performante que la dichotomie probabiliste, lorsque le $\mathcal{M}_{map}$ est employé en tant que modèle prédictif. Les expériences présentées ici sont pourtant favorables à la dichotomie probabiliste, car le niveau de bruit p est connu.

A.b Cas où la dichotomie probabiliste est renseignée par un niveau de bruit nul :

Comme précédemment, la Figure 5.8 compare les performances des stratégies actives en fonction du nombre d'exemples étiquetés. Dans la pratique, le niveau de bruit p n'est pas nécessairement connu. Dans ce paragraphe la dichotomie probabiliste est renseignée d'un niveau de bruit nul, alors que p varie.

Lorsque $p = 0.05$, la dichotomie probabiliste se comporte comme précédement (courbes bleues) : i) l'AUC est égale à 0.5 pour $|L| \leq 12$; ii) l'AUC est égale à 1 pour $|L| > 12$. Pour $p = 0.10$, la dichotomie probabiliste adopte un comportement différent. Au cours de l'échantillonnage sélectif, les performances du $\mathcal{M}_{map}$ augmentent dans un premier temps, puis chutent de manière drastique. Il s'agit d'un comportement pathologique dû au fait que les étiquettes découvertes sont particulièrement mélangées. Aprés avoir formé deux intervalles pour $|L| = 12$, le $\mathcal{M}_{map}$ comporte de nouveau un unique intervalle pour $|L| > 15$. Lorsque le niveau de bruit est supérieur à 0.10, les exemples sélectionnés par la dichothomie probabiliste ne permettent pas la construction d'un $\mathcal{M}_{map}$ incluant deux intervalles.

Les expériences présentées dans ce paragraphe montrent que les exemples sélectionnés par la dichotomie probabiliste ne sont pas adaptés à la construction du $\mathcal{M}_{map}$, cette stratégie étant renseignée par un niveau de bruit nul. Une fois encore, notre stratégie offre de meilleures performances que la dichotomie probabiliste.

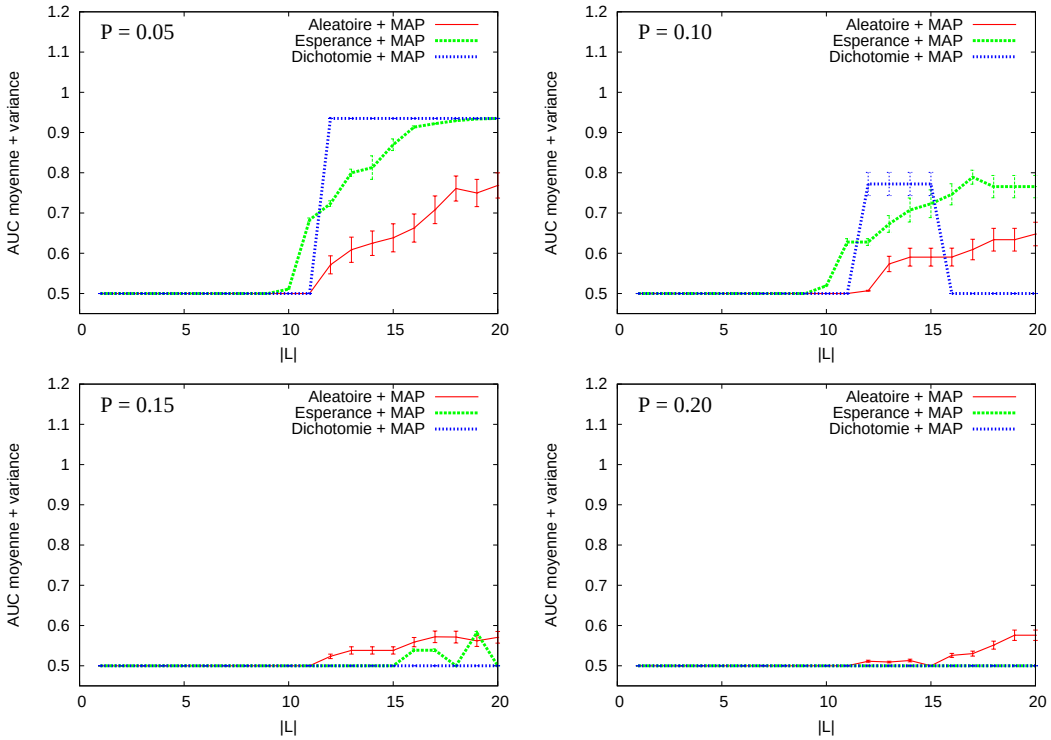

FIG. 5.8 – Performance moyenne des stratégies actives *vs.* le nombre d'exemples étiquetés. Cas où la dichotomie probabiliste est renseignée d'un niveau de bruit nul

A.c Cas où la dichotomie est renseignée par $p_{false} \in [0, 0.20]$, alors que $p = 0$:
Ici, la dichotomie probabiliste est renseignée par un niveau de bruit erroné, noté p_{false}, le véritable niveau de bruit p étant nul. La Figure 5.9 compare les performances des stratégies actives en fonction du nombre d'exemples étiquetés. Les expériences réalisées dans ce paragraphe montrent que les performances de la dichotomie probabiliste diminuent fortement lorsque la valeur de p_{false} augmente. Pour une valeur de p_{false} supérieure à 0.10, le modèle optimal comporte un seul intervalle quel que soit le nombre d'exemples étiquetés compris entre 1 et 20.

Notre stratégie active basée sur la maximisation de l'espérance de $P(M|D, x_{t+1})$ est performante, sans être renseignée du niveau de bruit présent dans les données. Cela constitue un avantage majeur sur la dichotomie probabiliste. De plus, notre stratégie ne fait pas l'hypothèse que les données sont organisées selon deux intervalles et considère l'éventualité de données complétement mélangées.

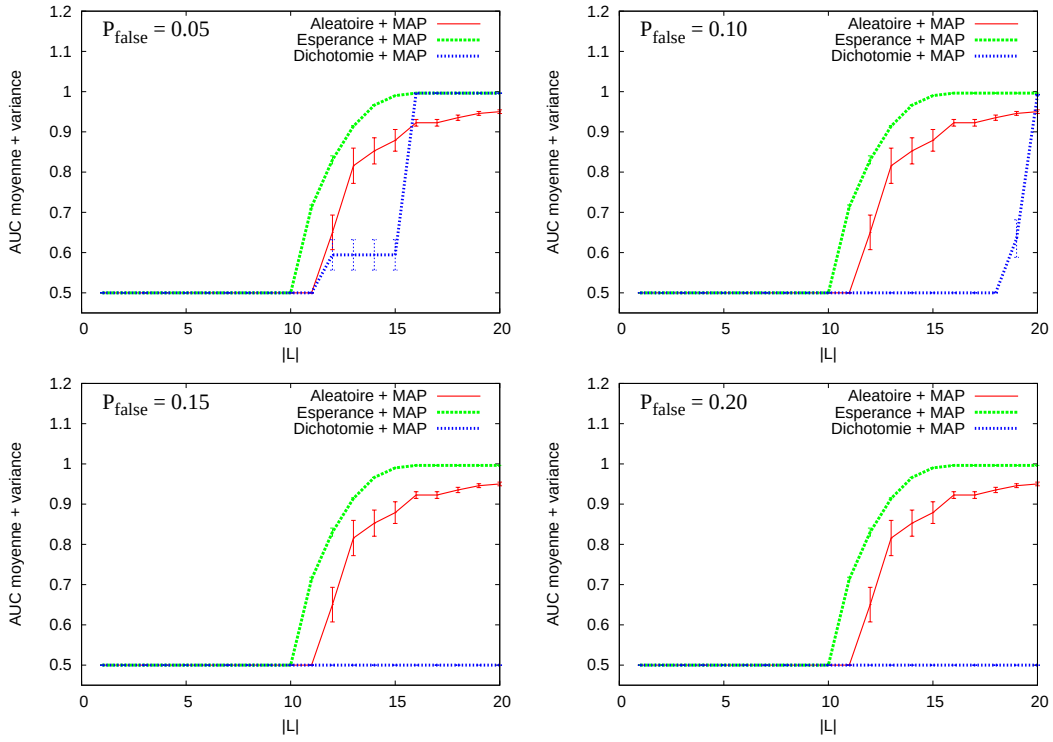

FIG. 5.9 – Performance moyenne des stratégies actives *vs.* le nombre d'exemples étiquetés. Cas où la dichotomie probabiliste est renseignée d'un niveau de bruit erroné.

B Évaluation : utilisation du modèle dichotomique

Dans cette section, les stratégies actives sont évaluées grâce au modèle dichotomique, qui comporte toujours deux intervalles et dont la frontière sépare la distribution $P_\theta(x)$ en deux parties équiprobables. Ici, la famille de modèles $\mathbb{M}$ est restreinte au cas où $I = 2$. Les stratégies actives font donc l'hypothèse que les données sont organisées selon deux intervalles. L'échantillonnage aléatoire n'est pas évalué lors de cette expérience. Les exemples étiquetés étant choisis de manière aléatoire, la mise à jour de $P_\theta(x)$ (étape "C" de l'Algorithme 10) ne permet pas la convergence de la borne prédite vers la position de l'échelon, notée θ.

B.a Cas où la dichotomie probabiliste est correctement renseignée du niveau bruit :

Nous considérons ici le cas où la dichothomie probabiliste est correctement renseignée du niveau bruit p. La Figure 5.10 trace la performance des stratégies actives en fonction du nombre d'exemples étiquetés, pour différentes valeurs de p. Lorsque le niveau de bruit est inférieur à 0.15, la dichotomie probabiliste et la maximisation de l'espérance de $P(M|D, x_{t+1})$ présentent des performances équivalentes. Notre stratégie est légérement moins performante que sa concurrente, pour $|L| \in [3, 6]$. Lorsque $p \geq 0.15$, notre stratégie est moins performante que la dichotomie probabiliste. Cette expérience met en évidence un léger avantage pour la dichotomie probabiliste, lorsque le niveau de bruit est connu.

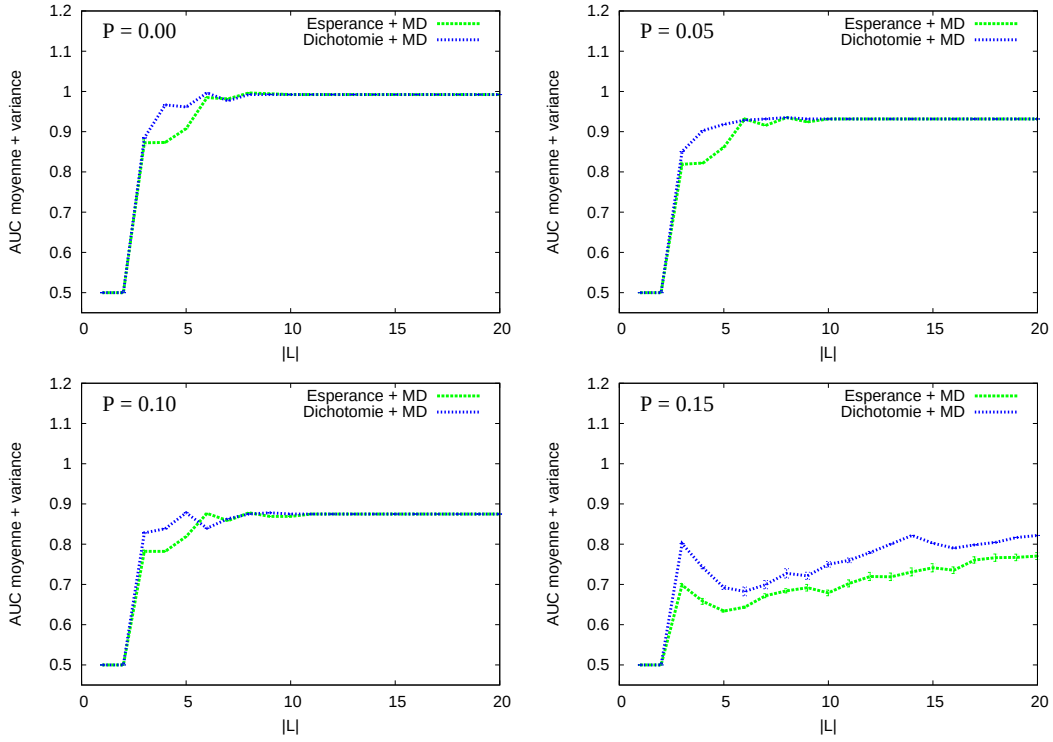

FIG. 5.10 – Performance moyenne des stratégies actives *vs.* le nombre d'exemples étiquetés. Cas où la dichotomie probabiliste est correctement renseignée du niveau de bruit

B.b Cas où la dichotomie probabiliste est renseignée par un niveau de bruit nul :
Lors des expériences présentées ici, la dichothomie probabiliste est renseignée d'un bruit nul, alors que p varie entre 0.05 et 0.35. Comme précédemment, la Figure 5.11 compare les performances des stratégies actives en fonction du nombre d'exemples étiquetés.

Au cours de l'échantillonnage sélectif, les performances de la dichothomie probabiliste augmentent dans un premier temps, puis chutent de manière radicale. Ce phénomène est observé pour toutes les valeurs du niveau de bruit considérées. La dichotomie probabiliste converge vers un modèle de discrétisation qui place tous les exemples étiquetés dans le même intervalle, le deuxième intervalle du modèle étant vide. Cela explique la dégradation des performances de cette stratégie, lorsque $|L|$ augmente.

Notre stratégie basée sur la maximisation de l'espérance de $P(M|D, x_{t+1})$ offre de meilleures performances que la dichotomie probabiliste, lorsque celle-ci est renseignée par un bruit nul. La Figure 5.11 montre que notre approche est relativement robuste au bruit d'étiquetage.

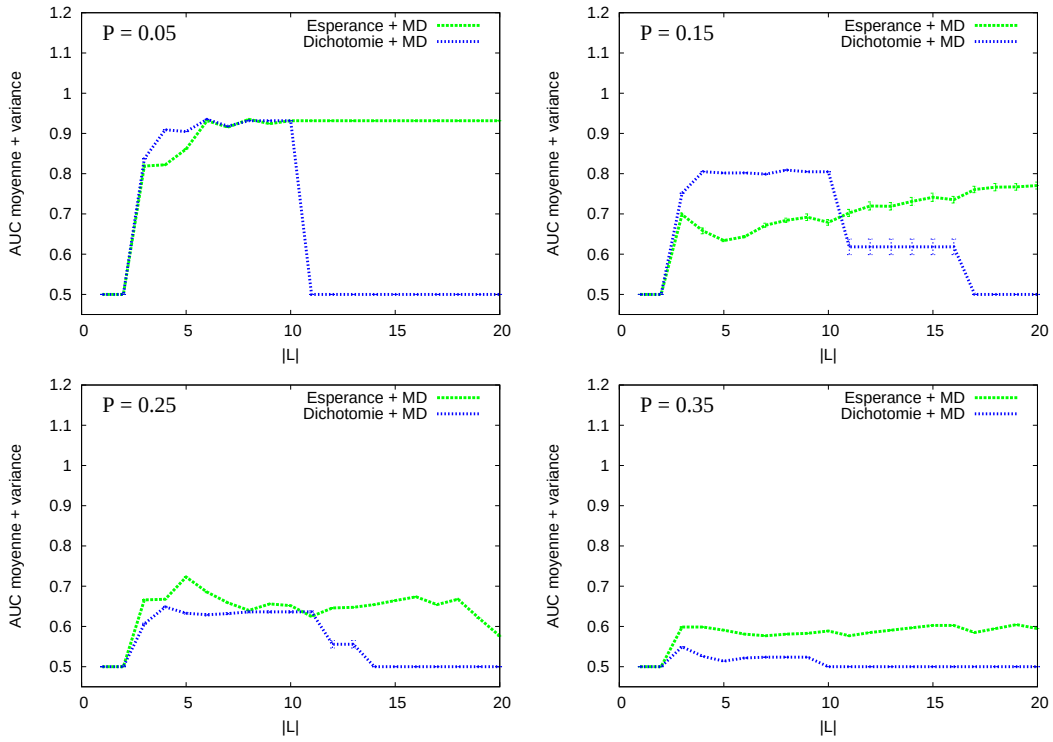

FIG. 5.11 – Performance moyenne des stratégies actives *vs.* le nombre d'exemples étiquetés. Cas où la dichotomie probabiliste est renseignée d'un niveau de bruit nul

B.c Cas où la dichotomie est renseignée par $p_{false} \in [0, 0.20]$, alors que $p = 0$:
Ici, la dichotomie probabiliste est renseignée par un niveau de bruit erroné, noté p_{false}, le véritable niveau de bruit p étant nul. La Figure 5.12 compare les performances des stratégies actives en fonction du nombre d'exemples étiquetés. Les expériences présentées dans ce paragraphe montrent que pour $p_{false} > 0.05$, les performances de la dichotomie probabiliste sont affectées. Cette tendance se confirme lorsque p_{false} augmente. Notre stratégie est performante, car elle ne requiert pas d'être renseignée du niveau de bruit présent dans les données.

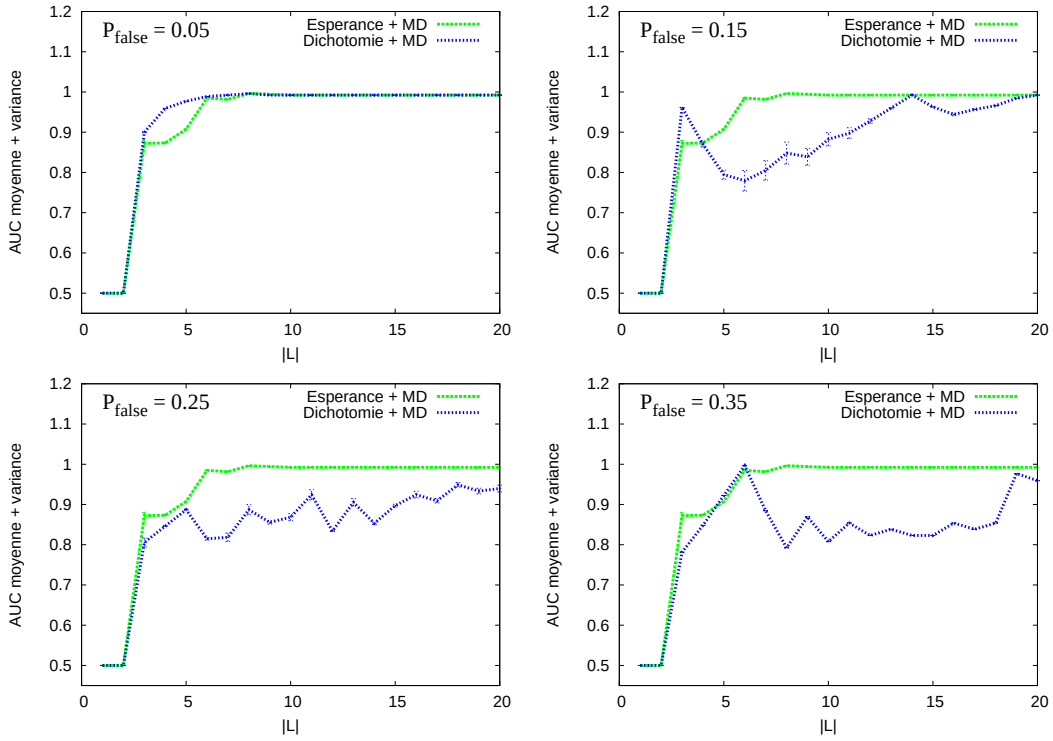

FIG. 5.12 – Performance moyenne des stratégies actives *vs.* le nombre d'exemples étiquetés. Cas où la dichotomie probabiliste est renseignée d'un niveau de bruit erroné.

Conclusion : Les expériences présentées dans cette section comparent notre stratégie basée sur le maximisation de l'espérance de $P(M|D, x_{t+1})$ à une stratégie de référence : la dichotomie probabiliste. Ces expériences montrent que notre approche est plus performante que la dichotomie probabiliste, lorsque celle-ci est mal renseignée du niveau de bruit présent dans les données. Deux modèles prédictifs ont été envisagés : A) le $\mathcal{M}_{map}$, qui ne fait pas l'hypothèse que les données sont organisées selon deux intervalles et qui considère l'éventualité de données complétement mélangées ; B) le modèle dichothomique, qui comporte toujours deux intervalles et qui fait l'hypothèse que le motif à découvrir est constitué de deux intervalles. Dans les deux cas, les résultats obtenus sont similaires.

5.4 Discussion

Ce chapitre présente une stratégie active originale basée sur la méthode de discrétisation semi-supervisée définie au Chapitre 4. Le cadre théorique exploité dans ce chapitre se restreint aux données unidimensionnelles et aux modèles de discrétisation comportant un ou deux intervalles. Notre stratégie active pourrait être exploitée pour améliorer la stratégie d'apprentissage actif par modèles locaux présentée au Chapitre 3. L'objectif serait de sélectionner, localement à la meilleure zone, l'exemple le plus utile à l'apprentissage du modèle local. La mise en œuvre de notre stratégie active dans le cadre d'un apprentissage actif par modèles locaux n'est pas réalisée dans cette thèse, et fait partie des futurs travaux envisageables.

Notre stratégie active consiste à sélectionner l'exemple non-étiqueté qui maximise l'espérance de la probabilité des modèles de discrétisation, connaissant les données et un exemple suplémentaire noté x_{t+1}. Notre démarche aboutit à un critère d'optimisation, noté $\mathcal{C}_{actif}(x_{t+1})$. Nous proposons plusieurs optimisations algorithmiques permettant d'exploiter notre stratégie active en un temps raisonnable. Enfin, notre approche est évaluée et comparée favorablement à une stratégie de référence : la dichotomie probabiliste.

L'étude comparative réalisée à la Section 5.3.4 montre que notre stratégie active est plus performante que la dichotomie probabiliste, lorsque le niveau de bruit présent dans les données n'est pas connu. Les deux approches donnent des résultats comparables lorsque le niveau de bruit est connu. Ce résultat est très encourageant et ouvre nos travaux à d'autres domaines d'application, comme par exemple les arbres de décision [Safavian and Landgrebe, 1991].

Un arbre de décision a pour but d'estimer la probabilité des classes $y \in \mathbb{Y}$ conditionnellement aux exemples $x \in \mathbb{X}$. Pour ce faire, l'espace des variables d'entrée est partitionné récursivement. Chaque feuille de l'arbre correspond à une zone de l'espace $\mathbb{X}$. Lors de la construction d'un arbre, deux décisions sont prises localement à chacun des nœuds : i) le choix de la variable à discrétiser ; ii) la discrétisation de la variable sélectionnée. Le critère utilisé pour prendre ces deux décisions caractérise généralement le gain de pureté réalisé lors du parcours d'un nœud. Il existe un grand nombre de critères fondés sur la théorie de l'information et sur des tests statisques, commme par exemple l'entropie de Shannon [Shannon, 1948] ou encore le critère du Gini. Il existe également des approches qui évaluent les corrélations existant entre les variables explicatives et la variable cible, comme par exemple les approches basées sur le test du χ^2 [Kerber, 1991]. L'approche de discrétisation MODL peut être exploitée pour la construction d'un arbre de décision. Une solution naïve consiste à répéter la discrétisation d'une variable récursivement dans chaque intervalle du modèle optimal. L'arbre ainsi constitué présente le désavantage de surapprendre les données d'apprentissage. Dans ce cas, les feuilles de l'arbre contiennent peu d'exemples et les prédictions réalisées sont mauvaises en généralisation. Pour pallier cette difficulté il est nécessaire d'évaluer les arbres de manière globale, en prenant en compte le coût de leur structure. Un nouveau critère d'évaluation permettant de sélectionner, parmi tous les arbres possibles, l'arbre le plus probable connaissant les données pourrait être défini. Ce critère devrait faire le compromis entre la qualité statistique des arbres et leur qualité prédictive. Classiquement, les arbres de décision sont des approches dédiées à l'ap-

prentissage supervisé. La méthode de discrétisation proposée au Chapitre 4 pourrait être exploitée pour la construction d'un arbre de décision semi-supervisé.

Des techniques incrémentales permettent la construction d'arbres de décision en prenant en compte des données d'apprentissage qui s'enrichissent itérativement [Utgoff, 1989]. Certaines de ces approches, comme par exemple ID5R, sont capables de remettre en cause des nœuds construits lors des itérations précédentes. Dans le cas général, ces méthodes ne permettent pas la sélection des nouveaux exemples à étiqueter. Un arbre binaire actif pourrait être défini grâce au critère $\mathcal{C}_{actif}$ présenté dans ce dernier chapitre. Le critère $\mathcal{C}_{actif}$ désignerait, localement à chacune des feuilles, l'exemple qui maximise l'espérance de $P(M|D, x_{t+1})$. Des travaux futurs sur un arbre actif, exploitant l'apprentissage actif par modèles locaux (Chapitre 3), pourraient être réalisés.

Le critère d'optimisation $\mathcal{C}_{actif}(x_{t+1})$ ne présente aucune limite théorique pour le traitement de données multi-classes ($J > 2$) caractérisées par plusieurs variables ($dim(\mathbb{X}) > 1$). En pratique, notre stratégie active est limitée par sa complexité algorithmique. Le parcours de l'ensemble des modèles de discrétisation lors de l'évaluation de l'espérance de $P(M|D, x_{t+1})$ est une étape coûteuse en temps de calcul. Cette difficulté pourrait être contournée par l'échantillonnage de la famille de modèles $\mathbb{M}$. Des travaux futurs pourraient étendre notre stratégie active à des problèmes d'apprentissage plus complexes, par exemple en exploitant une méthode du type *"monte carlo markov chain"* [Robert, 2006] pour échantillonner la famille de modèles $\mathbb{M}$.

La stratégie active présentée dans ce chapitre est fondée sur notre approche de discrétisation semi-supervisée (Chapitre 4). La méthode de discrétisation supervisée MODL pourrait être exploitée pour évaluer le critère $\mathcal{C}_{actif}(x_{t+1})$. L'influence du choix de la méthode de discrétisation (supervisée ou semi-supervisée) sur les performances de notre stratégie active devrait être étudiée lors de travaux futurs.

Chapitre

Conclusion générale

La particularité des stratégies d'apprentissage actif réside dans leur capacité à interagir avec leur environnement. L'objectif de ces stratégies est de sélectionner les exemples non-étiquetés les plus aptes à améliorer un modèle prédictif. Les exemples sélectionnés sont ensuite étiquetés par un expert, moyennant un coût. Les méthodes d'apprentissage actif ont pour but commun de réduire le coût d'étiquetage, nécessaire à l'apprentissage d'un modèle prédictif. Dans cette thèse, nous adoptons le point de vue de l'échantillonnage sélectif, où les stratégies actives ne sont pas autorisées à générer de nouveaux exemples. Parmi les problèmes d'apprentissage existant, nous choisissons de traiter uniquement les problèmes de classification. La finalité de nos travaux est de proposer une stratégie active innovante par rapport à l'état-de-l'art (Chapitre 2).

Les stratégies actives de la littérature utilisent souvent des modèles prédictifs globaux à tout l'espace d'entrée. L'idée défendue dans cette thèse est de partitionner cet espace et d'entraîner des modèles localement à chacune des zones. La sélection des exemples à étiqueter est réalisée en mettant en compétition les modèles locaux. Le modèle local qui progresse le plus désigne la zone de l'espace la plus intéressante pour l'étiquetage de nouveaux exemples. La curiosité adaptative a été adaptée avec succès à l'échantillonnage sélectif et est utilisée en tant que stratégie active pour résoudre des problèmes de classification (Chapitre 3). Nous avons amélioré cette stratégie issue de la robotique en proposant un nouveau critère de sélection de zones plus performant et plus simple à mettre œuvre que l'original. Notre stratégie d'apprentissage actif par modèles locaux a été comparée à d'autres stratégies de la littérature, et offre de très bonnes performances. Notre critère de sélection de zone ajuste le compromis entre l'exploitation des exemples déjà étiquetés et l'exploration des données, par le biais d'un paramètre. Ce critère constitue un premier pas vers la réduction du nombre de paramètres utilisateur que requière la curiosité adaptative. L'unique paramètre de notre critère se substitue à l'évaluation des progrès des modèles locaux, impliquant une mesure de performance et la dérivation de cette performance sur une fenêtre temporelle.

Les travaux réalisés ensuite ont pour objectif d'améliorer notre stratégie d'apprentissage actif par modèles locaux en rendant automatiques les décisions relatives au partitionnement d'une zone : i) *quand* couper une zone ; ii) *ou* couper la zone. Nous choisissons

d'exploiter une méthode de discrétisation supervisée, basée sur un formalisme Bayésien. La méthode MODL est particulièrement intéressante pour l'apprentissage actif par modèles locaux, puisqu'elle n'est pas assujettie au sur-apprentissage, qu'elle est très robuste et non-paramétrique. Le principal inconvénient de cette méthode est qu'elle n'exploite pas les exemples non-étiquetés, qui sont pourtant abondants lors d'un échantillonnage sélectif. L'extension de l'approche MODL au cas de l'apprentissage semi-supervisé constitue un des apports majeurs de cette thèse (Chapitre 4). Plusieurs résultats théoriques importants sont établis lors d'une étude approfondie de notre méthode de discrétisation semi-supervisée. Nous montrons que la prise en compte des exemples non-étiquetés dans la définition du prior n'est pas informative. Nous montrons également que notre approche semi-supervisée est asymptotiquement équivalente à l'approche supervisée, munie d'un post-traitement sur la position des bornes du modèle optimal. Dans notre cas, le meilleur moyen d'exploiter les exemples non-étiquetés est d'utiliser l'approche supervisée, puis de placer les bornes du modèle optimal au milieu des zones non-étiquetées. Ce résultat montre que, malgré les hypothèses faiblement informatives adoptées sur les données, les exemples non-étiquetés apportent de l'information utile à la discrétisation.

Notre stratégie d'apprentissage actif par modèles locaux définit les zones de l'espace des variables d'entrées où l'étiquetage de nouveaux exemples est le plus utile à l'apprentissage des modèles locaux. Les exemples étiquetés sont choisis de manière aléatoire dans les zones sélectionnées. Cette stratégie peut être améliorée en sélectionnant, localement à la meilleure zone, l'exemple qui est le plus utile à l'apprentissage du modèle local. Pour ce faire, nous proposons une stratégie originale d'apprentissage actif fondée sur notre méthode de discrétisation semi-supervisée (Chapitre 5). Cette stratégie sélectionne l'exemple non-étiqueté qui maximise l'espérance de la probabilité des modèles de discrétisation, connaissant les données. Nous proposons plusieurs optimisations algorithmiques permettant d'exploiter notre stratégie active en un temps raisonnable. Enfin, notre approche est évaluée et comparée favorablement à une autre stratégie active de la littérature : la dichotomie probabiliste [Horstein, 1963]. Ce résultat est très encourageant et ouvre nos travaux à d'autres heuristiques, qui exploitent elles aussi une méthode dichotomique et des données bruitées.

Les travaux réalisés dans les Chapitres 4 et 5 n'ont pas encore été exploités dans la cadre d'un apprentissage actif par modèles locaux. Notre méthode de discrétisation semi-supervisée pourrait être utilisée pour décider *"ou"* et *"quand"* couper une zone. Notre stratégie active maximisant l'espérance de la probabilité des modèles de discrétisation connaissant les données pourrait être utilisée pour sélectionner, localement à la meilleure zone, l'exemple le plus utile à l'apprentissage du modèle local. Ces deux améliorations possibles de l'apprentissage actif par modèles locaux font partie des futurs travaux envisageables. La méthode de discrétisation semi-supervisée présentée au Chapitre 4 a été élaborée dans un cadre théorique restreint : i) les modèles de discrétisation incluent un ou deux intervalles ; ii) les données sont caractérisées par une seule variable explicative. Des travaux futurs pourraient étendre les démonstrations réalisées dans cette thèse au cas des modèles multi-intervalles, et au cas de données caractérisées par plusieurs variables. La stratégie active présentée au Chapitre 5 peut également être améliorée. Le critère d'optimisation $\mathcal{C}_{actif}(x_{t+1})$ ne présente aucune limite théorique pour les problèmes multi-classes et

pour les données caractérisées par plusieurs variables. Pour réduire la complexité de notre approche, le parcours de l'ensemble des modèles de discrétisation pourrait être effectué de manière non-exhaustive, en échantillonnant la famille de modèles.

Chapitre

Annexes

A Critères d'évaluation

Cette Annexe présente les critères utilisés dans cette thèse pour évaluer les stratégies d'apprentissage actif. La Section A.a introduit l'aire sous la courbe de ROC (AUC). Ce critère est capable d'évaluer la qualité d'un modèle prédictif en exploitant un ensemble de test. La Section A.b présente un algorithme efficace pour l'évaluation de l'AUC. La Section A.c définit une mesure de déficit basée sur l'AUC. Ce critère compare une stratégie active à l'échantillonnage aléatoire, sur l'intégralité d'un jeu de données (quel que soit le nombre d'exemples étiquetés). Dans la Section A.d, nous présentons le calcul théorique de l'AUC dans le cas d'un modèle de discrétisation comportant exactement deux intervalles. Ce calcul est basé sur la position de la meilleure borne qui est supposée être connue, et ne nécessite pas d'ensemble de test.

A.a Présentation de l'AUC

Cette section présente l'aire sous la courbe de ROC, aussi appelée AUC. Ce critère évalue la qualité d'un modèle prédictif, en se basant sur un ensemble de test.

Courbe de ROC :
Pour introduire les courbes de ROC, nous considérons un problème de classification binaire. Chaque instance de ce problème de classification appartient à la classe y_1 ou à la classe y_2. Un classifieur est entraîné sur des données d'apprentissage T, l'objectif est d'évaluer les performances de ce modèle prédictif sur un ensemble de test. La matrice de confusion (voir figure A.a) compte pour chacune des classes le nombre d'instances bien classées et mal classées, sur l'ensemble de test. Les valeurs de la diagonale de cette matrice représentent les décisions correctes prises par le modèle prédictif, les autres valeurs représentent les erreurs de prédiction.

Les courbes de ROC sont tracées dans un espace à deux dimensions, l'axe des ordonnées représente le taux de prédictions correctes et l'axe des abscisses représente le taux de prédictions erronées. Une courbe de ROC peut être interprétée comme le "profit" (les y_1 bien classés) en fonction du "coût" (les y_1 mal classés) propre au modèle prédictif.

	y_1	y_2
prédit y_1	y_1 classés y_1	y_2 classés y_1
prédit y_2	y_2 classés y_1	y_2 classés y_1

FIG. A.a – Matrice de confusion

Nous supposons ici que la valeur de sortie du modèle prédictif estime la probabilité qu'un exemple soit de la classe y_1. La prédiction de la classe d'un exemple est réalisée en comparant la sortie du modèle à un seuil de décision. Pour chaque valeur du seuil, une matrice de confusion peut être construite, et un point peut être tracé dans l'espace de représentation des courbes de ROC. Une courbe de ROC est tracée en faisant varier le seuil de décision.

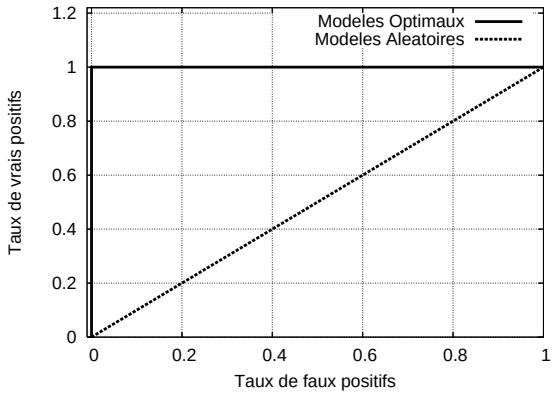

FIG. A.b – Espace de représentation des courbes de ROC

La Figure A.b illustre l'espace de représentation des courbes de ROC. Les courbes tracées sur cette figure représentent l'ensemble des modèles optimaux et l'ensemble des modèles aléatoires. Pour illustrer ceci, quelques points de la figure peuvent être commentés. Le point $(0, 0)$ représente le classifieur qui prédit toujours la mauvaise classe (y_2), puisque aucune "détection" de y_1, ni de fausses alertes ne sont détectées. Le point $(1, 1)$ représente quant à lui un modèle qui prédit toujours la bonne classe. La stratégie aléatoire prédit la classe y_1 une fois sur deux et est représentée par le point $(0.5, 0.5)$. Dans le cas où les classes sont équilibrées, ce modèle commet 50% de fausses alertes et fait 50% de bonnes prédictions. Enfin le point $(0, 1)$ représente un modèle qui ne se trompe jamais.

Les courbes de ROC sont des courbes en escalier, car elles sont construites à partir d'un ensemble fini d'exemples de test. En effet, les nouvelles instances prises en compte lorsqu'on augmente le seuil de décision sont soit des fausses alertes, ce qui dessine le "plat" d'une marche, soit des alertes avérées, ce qui dessine le "front montant" d'une nouvelle marche.

Calcul de l'AUC :
Les courbes de ROC représentent les performances d'un modèle prédictif dans un espace

à deux dimensions. Pour comparer plusieurs modèles, il est nécessaire de "résumer" les courbes de ROC en une valeur scalaire. Une méthode classique pour réaliser cela est d'intégrer l'aire sous la courbe de ROC, cette mesure est appelée l'AUC (*Area Under ROC Curves*) [Bradley, 1997]. L'AUC peut être vue comme une proportion de l'espace de description des courbes de ROC. En se reportant à la Figure A.b il apparaît que cette aire est comprise entre 1 (si le modèle est parfait) et $\frac{1}{2}$ (si le modèle est aléatoire). L'AUC a des propriétés statistiques intéressantes [Fawcett, 2003], cette mesure correspond à la probabilité que le modèle attribue une probabilité plus importante à un exemple de classe y_1 qu'à un exemple de classe y_2, ces exemples étant tirés de manière aléatoire dans l'ensemble de test [Hanley and McNeil, 1982]. L. Breiman montre [Breiman *et al.*, 1984] que l'AUC est liée au critère de Gini par la relation suivante : $Gini + 1 = 2.AUC$. L'annexe A.b présente un algorithme [Fawcett, 2003] efficace qui calcule l'AUC en une complexité de l'ordre de $O(n \log n)$, avec n le nombre d'exemples de l'ensemble de test.

Cas des problèmes multi-classes :
Lorsque le nombre de classes augmente, la matrice de confusion devient une matrice carrée de dimension $|\mathbb{Y}| \times |\mathbb{Y}|$, avec $|\mathbb{Y}|$ le nombre de classes. Les $\mathbb{Y}$ valeurs de la diagonale principale correspondent aux classifications correctes et les $\mathbb{Y}^2 - \mathbb{Y}$ autres valeurs représentent les classifications erronées sur l'ensemble de test. Il est difficile de représenter sur un même graphique le taux de bonne classification d'une classe en fonction des erreurs commises pour chacune des classes. Si la définition des courbes de ROC reste inchangée, l'espace de représentation contient $\mathbb{Y}^2 - \mathbb{Y}$ dimensions. Pour des problèmes multi-classes, nous nous ramènerons au cas de la classification binaire.

Dans la pratique, $\mathbb{Y}$ courbes de ROC sont manipulées. Pour construire une courbe de ROC, une meta-classe cible $Y_1 = y_i$ est considérée, toute les autres classes sont regroupées et constituent la deuxième meta-classe $Y_2 = \bigcup_{j=1\,,\,j\neq i}^{m} y_j$. Nous nous ramenons au cas d'une matrice de confusion de dimension 2×2. L'AUC peut être calculée pour chacune des m courbes de ROC obtenues. Pour estimer la performance globale du classifieur, l'espérance de l'AUC sur toutes les classes est calculée. Cela s'écrit de la manière suivante :

$$AUC_{global} = \sum_{i=1}^{m} P(y_i).AUC(y_i)$$

A.b Un algorithme efficace pour le Calcul de l'AUC

Cette section présente un algorithme efficace, permettant de calculer l'AUC en une complexité temporelle de l'ordre de $O(N \log N)$. L'idée principale est de calculer l'AUC au cours de la construction des courbes de ROC. Pour chaque valeur du seuil de décision testée, l'aire du trapèze formé par deux points consécutifs de la courbe de ROC est calculée. La Figure A.c illustre cela. La partie "A" de cette figure montre deux points consécutifs qui forment le "front montant" d'une marche, dans ce cas l'aire calculée est nulle. La partie "B" représente le cas où les deux points forment le "plat" d'une marche. Enfin, la partie "C" de la Figure A.c correspond à plusieurs exemples pour lesquels le modèle prédictif a attribué le même score, dans ce cas l'aire calculée forme un trapèze. L'Algorithme 11 décrit de manière détaillée la méthode utilisée dans [Fawcett, 2003] pour calculer l'AUC.

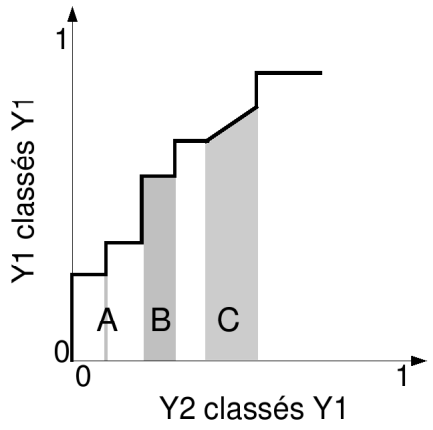

FIG. A.c – Illustration : Intégration sous la courbe de ROC

Notations :

- Des instances de test $u \in Test$ associées à leur classe $f(u)$
- Un classifieur $\mathcal{M}$ associant une sortie $OutPut(u)$ à l'instance u

*/*initialisation des variables*/*
$AUC \leftarrow 0$ */*L'Aire sous la courbe de ROC qu'on cherche à calculé*/*
$FA \leftarrow 0$ */*le nombre de fausses alertes*/*
$D \leftarrow 0$ */*le nombre de détection*/*
$FA_{prec} \leftarrow 0$ */*le nombre précédent de fausses alertes*/*
$D_{prec} \leftarrow 0$ */*le nombre précédent de détections*/*
$OutPut_{prec} \leftarrow -\infty$ */*la sortie précédente du modèle*/*

Trier $Test$ dans l'ordre décroissant des scores de sortie $OutPut(u)$
Pour tout $u \in Test$, **faire**
  */*Pas de nouveaux points tant qu'il y a des instances qui ont la même sortie*/*
  Si $OutPut(u) \neq OutPutprec$ **Alors**
    */*On calcule l'aire du nouveau trapèze*/*
    $AUC \leftarrow AUC + (FA - FA_{prec}).(\frac{D+D_{prec}}{2})$
  Fin Si
  $OutPut_{prec} \leftarrow OutPut(u)$
  $FA_{prec} \leftarrow FA$
  $D_{prec} \leftarrow D$
  */*On incrémente les compteurs de détection et de fausse alerte*/*
  Si $f(u) = y_1$ **Alors**
    $D \leftarrow D + 1$
  Sinon
    $FA \leftarrow FA + 1$
  Fin Si
Fin Pour
*/*On ajoute le dernier point et le point (1,1)*/*
$AUC \leftarrow AUC + (FA - FA_{prec}).(\frac{D+D_{prec}}{2})$
$AUC \leftarrow AUC + (1 - FA).(\frac{1+D}{2})$
*/*On normalise l'AUC par rapport au nombre d'instances*/*
$AUC \leftarrow \frac{AUC}{|Test|}$

Algorithme 11: Calcul de l'aire sous la courbe de ROC en $O(n \log n)$

A.c Mesure de déficit basée sur l'AUC

Cette section présente une mesure de déficit basée sur l'AUC, permettant d'obtenir une valeur scalaire représentative de la performance d'une stratégie active sur l'intégralité d'un jeu de données.

L'évaluation d'une stratégie d'apprentissage actif prend la forme d'une courbe qui trace la performance moyenne du modèle prédictif, en fonction du nombre d'exemples étiquetés. Cette représentation est problématique lorsque plusieurs stratégies actives sont comparées entre elles. Les courbes de performance des différentes stratégies peuvent se croiser, si tel est le cas, il est difficile de déterminer si une stratégie est meilleure qu'une autre. Y. Baram propose une mesure qui évalue le "déficit" d'une stratégie d'apprentissage actif sur l'intégralité d'un jeu de données [Baram *et al.*, 2004].

Cette mesure de déficit compare les performances d'une stratégie active (notée "*Active*") aux performances de la stratégie aléatoire (notée "*Rand*"). La stratégie aléatoire sélectionne les exemples à étiqueter de manière uniforme, parmi les exemples non-étiquetés. Le déficit est donc une mesure comparative et globale, qui intègre une mesure de performance sur l'ensemble des valeurs de $|L|$. Dans le cas de l'AUC, le déficit d'une stratégie *Active* s'exprime comme suit :

$$Deficit_n(Active) = \frac{\sum_{t=0}^{n} [AUC_n(Rand) - AUC_t(Active)]}{\sum_{t=0}^{n} [AUC_n(Rand) - AUC_t(Rand)]} \quad (.1)$$

Avec n le nombre maximum d'exemples étiquetés pendant l'apprentissage actif ($n = \max |L|$).

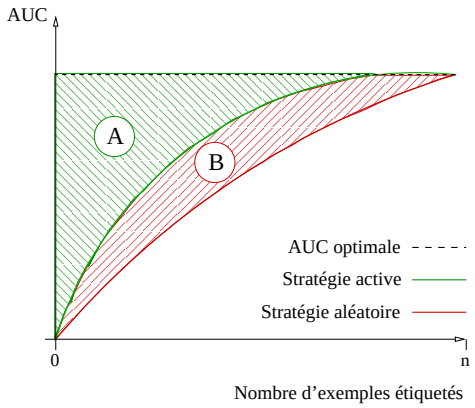

FIG. A.d – Définition du déficit

La Figure A.d illustre le calcul du déficit, deux courbes de performance correspondant aux stratégies *Active* et *Rand* sont représentées. L'Équation .1 peut être simplement définie par le ratio des aires $\frac{A}{A+B}$. Lorsque la stratégie *Active* donne exactement les mêmes performances que la stratégie *Rand*, le déficit est égal à 1. Lorsque le nombre d'exemples nécessaires à la stratégie *Active* pour atteindre la performance optimale tend vers 0, le déficit tend vers 0. La valeur du déficit peut être négative lorsque la stratégie *Active* atteint une meilleure performance que la stratégie *Rand*, pour $n = max|L|$.

A.d AUC théorique pour un modèle à deux intervalles

Cette section présente le calcul de l'AUC théorique, pour un modèle de discrétisation à deux intervalles et des données qui forment un échelon bruité (Section 5.3.4). Nous nous plaçons dans le cas où les étiquettes forment deux intervalles mélangés par un bruit uniforme, les deux intervalles présentant le même niveau de bruit. D'une part, la vraie borne correspondant à la position de l'échelon est notée "TB". D'autre part, le modèle M possède deux intervalles dont la frontière est notée "PB". La Figure A.e représente le jeu de données et la borne du modèle M.

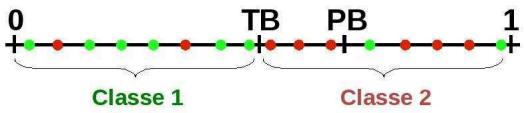

FIG. A.e – Echelon bruité

L'objectif est de calculer l'AUC théorique du modèle M, en tenant compte du niveau de bruit. Il s'agit d'un calcul exact qui ne requiert pas d'ensemble de test, habituellement utilisé pour le calcul de l'AUC. Soit "p" la probabilité qu'une instance soit étiquetée par la mauvaise étiquette. Comme le montre la Figure A.f, les courbes de ROC théoriques comportent deux régimes. Les deux segments des courbes de ROC correspondent aux deux intervalles du modèle de prédiction. La transition entre ces deux régimes est défini par le point "a" dont les coordonnées sont notées (x_a, y_a).

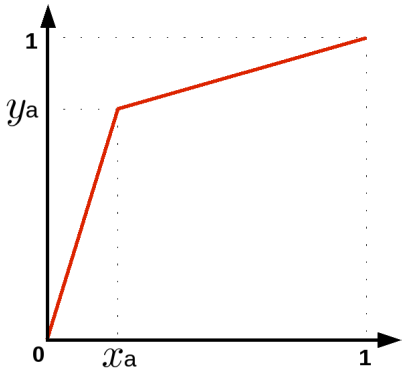

FIG. A.f – Courbe de ROC théorique pour l'échelon bruité

Les coordonnés du point a peuvent être calculés facilement. Sur la Figure A.e, la borne prédite est à droite de la vraie borne ($PB > TB$). Dans ce cas, et pour la courbe de ROC dédiée à la classe "1", on a :

$x_a = \frac{TB.p+|TB-PB|.(1-p)}{1-TB}$
$y_a = \frac{TB.(1-p)+|TB-PB|.p}{TB}$

x_a correspond à la proportion d'exemples prédits comme étant de la classe "1" dans l'intervalle $[0, PB]$, alors qu'ils sont en réalité de la classe "2". À gauche de la borne théorique ($[0, TB]$), $TB.p$ exemples sont mal étiquetés à cause du bruit. Entre les deux bornes ($[TB, PB]$), il y a $|TB - PB|.(1 - p)$ exemples dont la prédiction est erronée. Le dénominateur de l'expression de x_a représente le nombre total d'exemples de la classe "2" qui auraient pu être mal classées par le modèle prédictif. y_a représente la proportion d'exemples prédits à juste titre comme étant de la classe "1". Dans le cas où $PB > TB$, la courbe de ROC dédiée à la classe "2" est définie tel que :

$x_a = \frac{(1-PB).p}{TB}$
$y_a = \frac{(1-PB).(1-p)}{1-TB}$

Nous démontrons également que pour $PB < TB$, la courbe de ROC dédiée à la classe "1" est définie par :

$x_a = \frac{PB.p}{1-TB}$
$y_a = \frac{PB.(1-p)}{TB}$

Et la courbe de ROC dédiée à la classe "2" est définie par :

$x_a = \frac{(1-TB).p+|TB-PB|.(1-p)}{TB}$
$y_a = \frac{(1-TB).(1-p)+|TB-PB|.p}{1-TB}$

Dans tous les cas, l'aire sous la courbe de ROC se calcule selon l'expression suivante :

$AUC = \frac{1}{2}.x_a.y_a + y_a.(1 - x_a) + \frac{1}{2}.(1 - x_a).(1 - y_a)$

B Problème d'arrondi pour les N_{ij} optimaux

Dans la Section 4.3.2.1 (page 79), nous caractérisons les paramètres optimaux $N_{ij}^{\diamond}$ par une formule analytique. Cependant, des problèmes d'arrondis peuvent être observés tel que $\sum_{j=1}^{J} N_{ij}^{\diamond} = N_i - 1$. Pour résoudre de problème, il suffit de choisir un des $N_{ij}^{\diamond}$ et de lui ajouter 1. Tous les choix possibles sont équivalents et optimaux au sens du critère $\mathcal{C}_{semi\,super}$.

Démonstration.
Cette démonstration se restreint au cas d'un modèle de discrétisation incluant un seul intervalle ($I = 1$) et au cas d'une classification binaire ($J = 2$). Soit la fonction $f(N_{i1}, N_{i2})$ correspondant au critère $\mathcal{C}_{semi\,sup}$, pour lequel tous les paramètres sont fixés exceptés N_{i1} et N_{i2}.

$$f(N_{i1}^{\diamond}+1, N_{i2}^{\diamond}) - f(N_{i1}^{\diamond}, N_{i2}^{\diamond}+1) = \log\left(\frac{\left(\frac{N_i \times N_{i1}^l - N_{i2}^l}{N_{i1}^l + N_{i2}^l} + 1 - N_{i1}^l\right)!}{\left(\frac{N_i \times N_{i1}^l - N_{i2}^l}{N_{i1}^l + N_{i2}^l} + 1\right)!}\right) + \log\left(\frac{\left(\frac{N_i \times N_{i2}^l - N_{i1}^l}{N_{i1}^l + N_{i2}^l} - N_{i2}^l\right)!}{\left(\frac{N_i \times N_{i2}^l - N_{i1}^l}{N_{i1}^l + N_{i2}^l}\right)!}\right)$$
$$- \log\left(\frac{\left(\frac{N_i \times N_{i1}^l - N_{i2}^l}{N_{i1}^l + N_{i2}^l} - N_{i1}^l\right)!}{\left(\frac{N_i \times N_{i1}^l - N_{i2}^l}{N_{i1}^l + N_{i2}^l}\right)!}\right) - \log\left(\frac{\left(\frac{N_i \times N_{i2}^l - N_{i1}^l}{N_{i1}^l + N_{i2}^l} + 1 - N_{i2}^l\right)!}{\left(\frac{N_i \times N_{i2}^l - N_{i1}^l}{N_{i1}^l + N_{i2}^l} + 1\right)!}\right)$$

$$= \log\left(\frac{N_i \times N_{i1}^l - N_{i2}^l}{N_{i1}^l + N_{i2}^l} + 1 - N_{i1}^l\right) - \log\left(\frac{N_i \times N_{i1}^l - N_{i2}^l}{N_{i1}^l + N_{i2}^l} + 1\right)$$
$$- \log\left(\frac{N_i \times N_{i2}^l - N_{i1}^l}{N_{i1}^l + N_{i2}^l} + 1 - N_{i2}^l\right) + \log\left(\frac{N_i \times N_{i2}^l - N_{i1}^l}{N_{i1}^l + N_{i2}^l} + 1\right)$$

$$= \log\left(\frac{\left(\frac{N_i \times N_{i1}^l - N_{i2}^l}{N_{i1}^l + N_{i2}^l} + 1 - N_{i1}^l\right)\left(\frac{N_i \times N_{i2}^l - N_{i1}^l}{N_{i1}^l + N_{i2}^l} + 1\right)}{\left(\frac{N_i \times N_{i1}^l - N_{i2}^l}{N_{i1}^l + N_{i2}^l} + 1\right)\left(\frac{N_i \times N_{i2}^l - N_{i1}^l}{N_{i1}^l + N_{i2}^l} + 1 - N_{i2}^l\right)}\right)$$

$$= \log\left(\frac{\left(N_i \times N_{i1}^l + N_{i1}^l - N_{i1}^l \times N_{i2}^l - (N_{i1}^l)^2\right)\left(N_i \times N_{i2}^l + N_{i2}^l\right)}{\left(N_i \times N_{i1}^l + N_{i1}^l\right)\left(N_i \times N_{i2}^l + N_{i2}^l - N_{i1}^l \times N_{i2}^l - (N_{i2}^l)^2\right)}\right)$$

$$= \log(1) = 0$$

$$f(N_{i1}^{\diamond}+1, N_{i2}^{\diamond}) = f(N_{i1}^{\diamond}, N_{i2}^{\diamond}+1)$$

□

C Implémentation de l'espérance de $P(M|D, x_{t+1})$

La stratégie de sélection d'exemples basée sur l'espérance de $P(M|D, x_{t+1})$ s'écrit de la manière suivante :

$$\underset{x_{t+1}\in U}{ArgMax} \sum_{M\in\mathbb{M}} \left[P(M)P(D|M) \times \sum_{y\in\mathbb{Y}} \left[\frac{P(y|M, D, x_{t+1}) \times P(M) \times P(D, x_{t+1}, y|M)}{\sum_{M'\in\mathbb{M}} P(M') \times P(D|M') \times P(y|D, M', x_{t+1})} \right] \right]$$

Pour implémenter cette formule il faut prendre quelques précautions. Les produits de probabilités, qui sont des termes très petits, peuvent induire des problèmes d'arrondis machine. Nous présentons ici une forme de cette expression qui limite ce risque.

Démonstration.
Le critère $\mathcal{C}_{semi\ super}$ défini à la section 4.2.4 correspond à $I(M|D)$ l'information d'un modèle, conditionnellement aux données observées. L'expression précédente peut s'écrire de la manière suivante, car $P(M) \times P(D|M) = e^{-I(M|D)}$:

$$\underset{x_{t+1}\in U}{ArgMax} \sum_{M\in\mathbb{M}} \left[e^{-I(M|D)} \times \sum_{y\in\mathbb{Y}} \left[\frac{P(y|M, D, x_{t+1}) \times e^{-I(M|D, x_{t+1}, y)}}{\sum_{M'\in\mathbb{M}} e^{-I(M'|D)} \times P(y|D, M', x_{t+1})} \right] \right]$$

L'expression est normalisée en utilisant le maximum a posteriori, $\mathcal{M}_{map} = ArgMax_{M\in\mathbb{M}}\, e^{-I(M|D)}$. Les termes du type "$e^{-I(M|D)}$" ont des valeurs proches de zéro, la multiplication de ces termes peut causer des problèmes d'arrondis. Ces termes sont divisés par $e^{-I(\mathcal{M}_{map}|D)}$ de telle sorte qu'ils aient une valeur proche de 1. Finalement :

$$\underset{x_{t+1}\in U}{ArgMax} \sum_{M\in\mathbb{M}} \left[e^{[-I(M|D)+I(\mathcal{M}_{map}|D)]} \times \sum_{y\in\mathbb{Y}} \left[\frac{P\left(y|M, D, x_{t+1}\right) \times e^{[-I(M|D, x_{t+1}, y)+I(\mathcal{M}_{map}|D)]}}{\sum_{M'\in\mathbb{M}} \left[P(y|D, M', x_{t+1}) \times e^{-I(M'|D)+I(\mathcal{M}_{map}|D)}\right]} \right] \right]$$

□

D Nombre d'exemples étiquetés à chaque itération

Cette Annexe présente une étude expérimentale caractérisant l'influence du nombre d'exemples étiquetés à chaque itération d'un échantillonnage sélectif , noté e, sur les performances de deux stratégies d'apprentissage actif [Lemaire *et al.*, 2007]. Les enjeux de ce problème peuvent être illustrés grâce à deux situations extrêmes. Lorsque $e = 1$, le temps de calcul nécessaire à l'apprentissage du modèle prédictif et à la sélection des exemples est élevé. L'application de stratégies actives à de grosses bases d'apprentissage peut s'avérer problématique. À l'inverse, l'étiquetage d'un grand nombre d'exemples à chaque itération amoindrit l'apport des stratégies actives. L'ajustement du paramètre e peut être vu comme la recherche d'un compromis entre le temps de calcul et l'efficacité d'une stratégie active.

Protocole expérimental :

Les stratégies évaluées lors de nos expériences comparatives sont les suivantes :

- l'echantillonnage par incertitude (Section 2.2.1, page 9) ;
- l'échantillonnage par réduction de l'erreur de généralisation (Section 2.2.4, page 16) ;
- l'échantillonnage aléatoire (Section 3.2.2, page 39).

Les données sont préalablement centrées et réduites. Au début de chaque expérience, l'ensemble d'apprentissage T ne possède que deux exemples choisis de manière aléatoire. À chaque itération de l'échantillonnage sélectif, e exemples sont étiquetés et ajoutés à l'ensemble d'apprentissage. Ces exemples sont sélectionnés par une stratégie active. Les expériences sont réalisées pour plusieurs valeur de e (1, 2, 4, 8, 16). Chaque expérience est répétée dix fois de manière à obtenir une performance moyenne munie de sa variance, pour chaque point des courbes de résultat.

L'AUC est utilisée pour évaluer la performance du modèle prédictif, en fonction du nombre d'exemples étiquetés (Annexe A.a). La performance du modèle prédictif reflète la qualité des exemples étiquetés. Le modèle prédictif utilisé est une fenêtre de Parzen à noyau gaussien et de norme L2 [Parzen, 1962]. Ce modèle est présenté à la Section 3.4 (page 51) :

$$\hat{p}(y|x) = \frac{\sum_{n=1}^{|L|} \mathbb{1}_{\{f(l_n)=y\}} \, K(x, l_n)}{\sum_{n=1}^{|L|} K(x, l_n)} \qquad l_n \in L \ , \ x \in \Phi \qquad (.2)$$

avec

$$K(x, l_n) = e^{\frac{||x-l_n||^2}{2\sigma^2}}$$

La fenêtre de Parzen utilise la même valeur du paramètre σ pour chaque dimension du problème traité. La valeur optimale de σ est déterminée au début de chaque expérience en entraînant le modèle grâce à la totalité des données d'apprentissage, en effectuant une *cross-validation* de l'AUC sur l'ensemble de test [Chappelle, 2005]. Cette valeur est utilisée ensuite pour fixer le paramètre de la fenêtre de Parzen durant l'expérience. L'apprentissage se réduit au "comptage" des exemples au sens du noyau gaussien, l'unique paramètre du

modèle prédictif étant fixé. La performance du modèle dépend uniquement de la qualité des exemples étiquetés.

Jeux de données :
Les jeux de données utilisés dans cette section sont issus de *"l'UCI Repository"* [D.J. Newman and Merz, 1998]. Les problèmes de classification traités sont les suivants :

- **Glass - Classification de différents types de verre :** Ce jeu de données comporte 214 instances (Apprentissage : 146, Test : 68) caractérisées par 9 attributs continus. Ces attributs correspondent respectivement à l'indice de réflexion du verre et aux taux de différents éléments chimiques (Sodium, Magnesium, Aluminum, Silicium, Potassium, Calcium, Barium et fer). Les 6 classes considérées sont : "fenêtre fabriquée par flottage", "fenêtre", "vitre de voiture", "récipient", "vaisselle" et "phares".

- **Iris - Classification d'espèces de fleurs :** Ce jeu de données décrit 150 fleurs (Apprentissage : 90 Test : 60) grâce à 4 attributs numériques (longueur et largeur des pétales et des sépales). Ces fleurs appartiennent à trois espèces d'iris (Setosa, Versicolour, Virginica).

- **Segment - Recherche de textures dans des images :** Ce jeu de données comporte 2310 images de 9 pixels (Apprentissage : 310 Test : 2000) décrites par 19 attributs numériques. Ces petites images sont en réalité extraites de photographies plus grandes et correspondent à des textures. Les 7 classes prises en compte dans ce problème de classification sont : "façade en briques", "ciel", "feuillage", "ciment", "fenêtre", "chemin" et "pelouse".

Résultats :
Les figures D.g, D.h et D.i montrent les résulats obtenus pour chacun des jeux de données. Les 5 graphiques de chaque figure représentent, pour différentes valeurs de e, les performances des stratégies (axe vertical) en fonction du nombre d'exemples étiquetés (axe horizontal). Les moustaches sur les courbes représentent la variance des résultats ($\pm 2\sigma$).

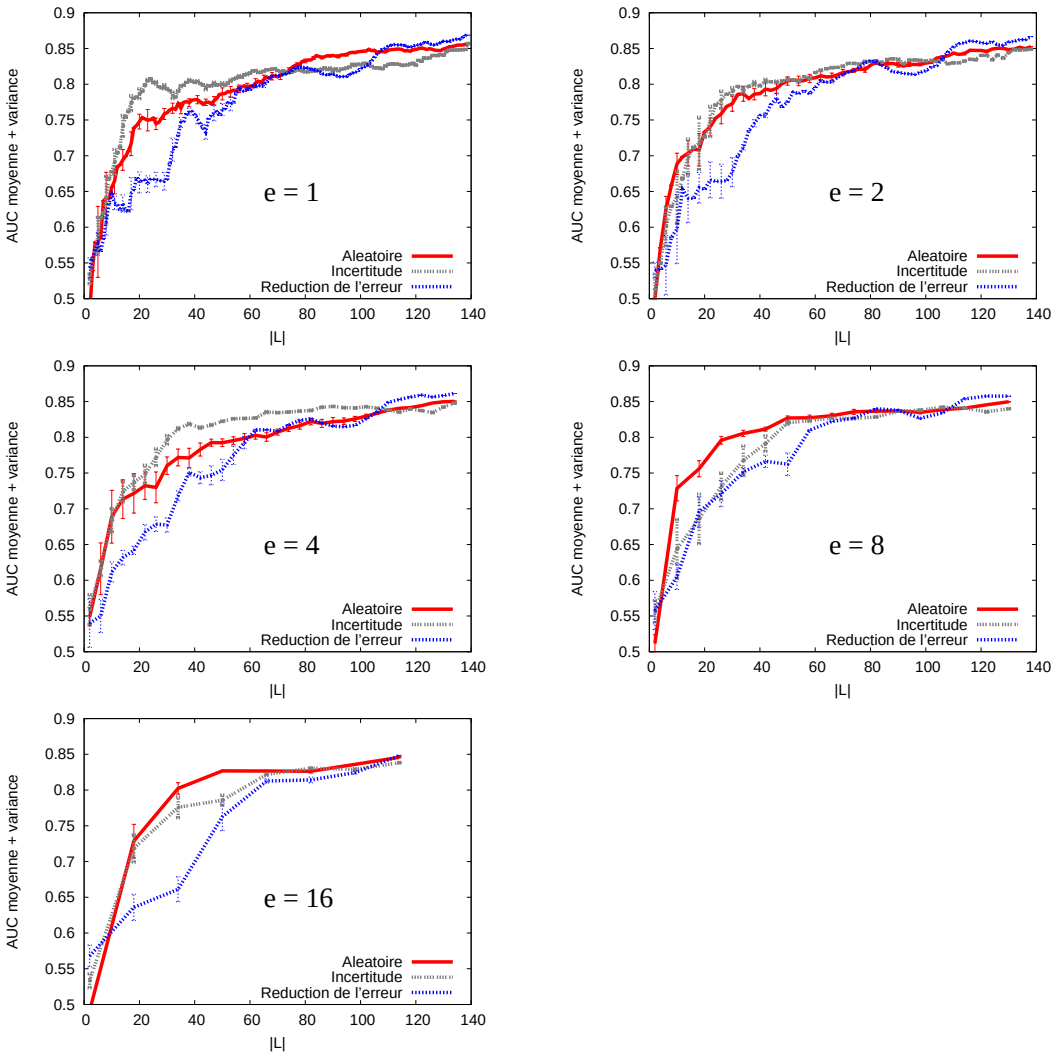

FIG. D.g – Résultats complets pour le jeu de données "Glass"

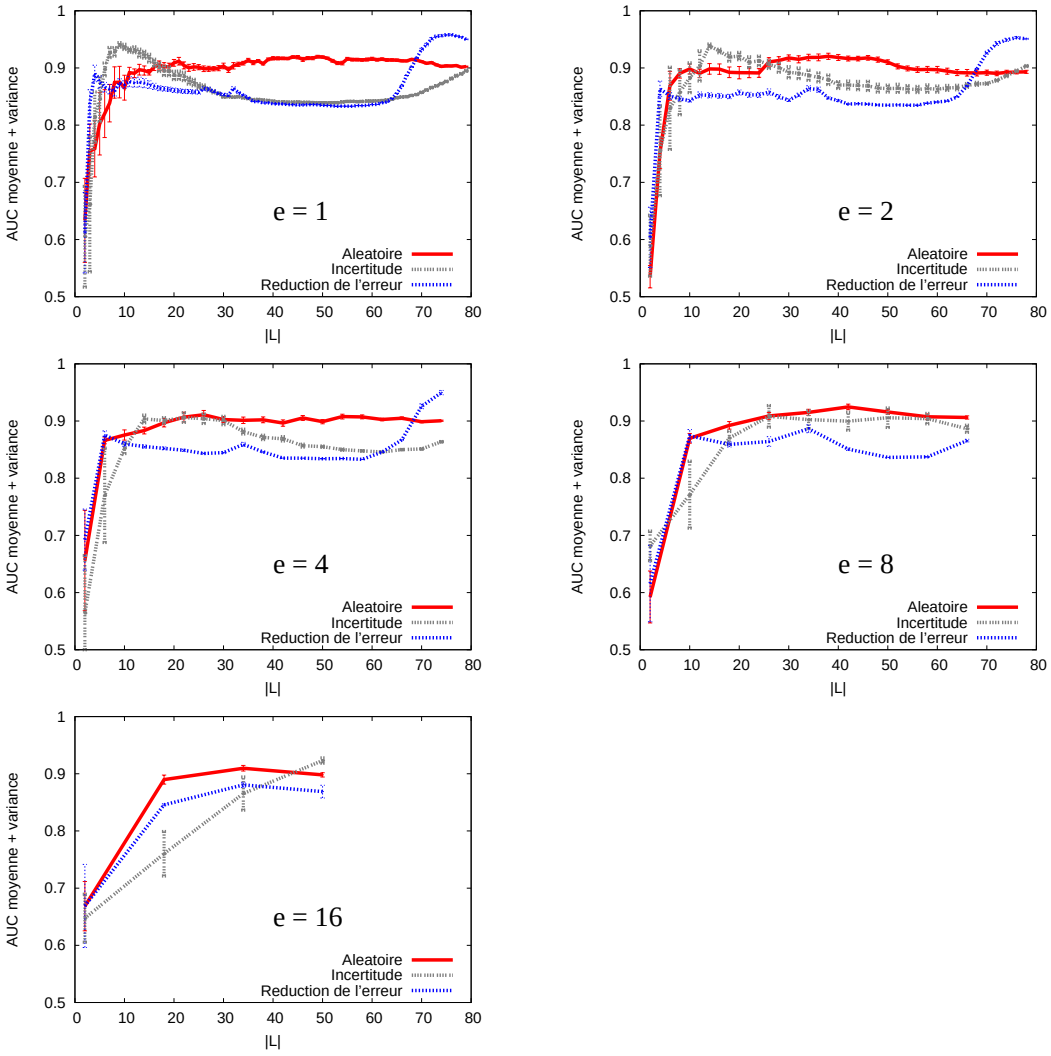

FIG. D.h – Résultats complets pour le jeu de données “Iris”

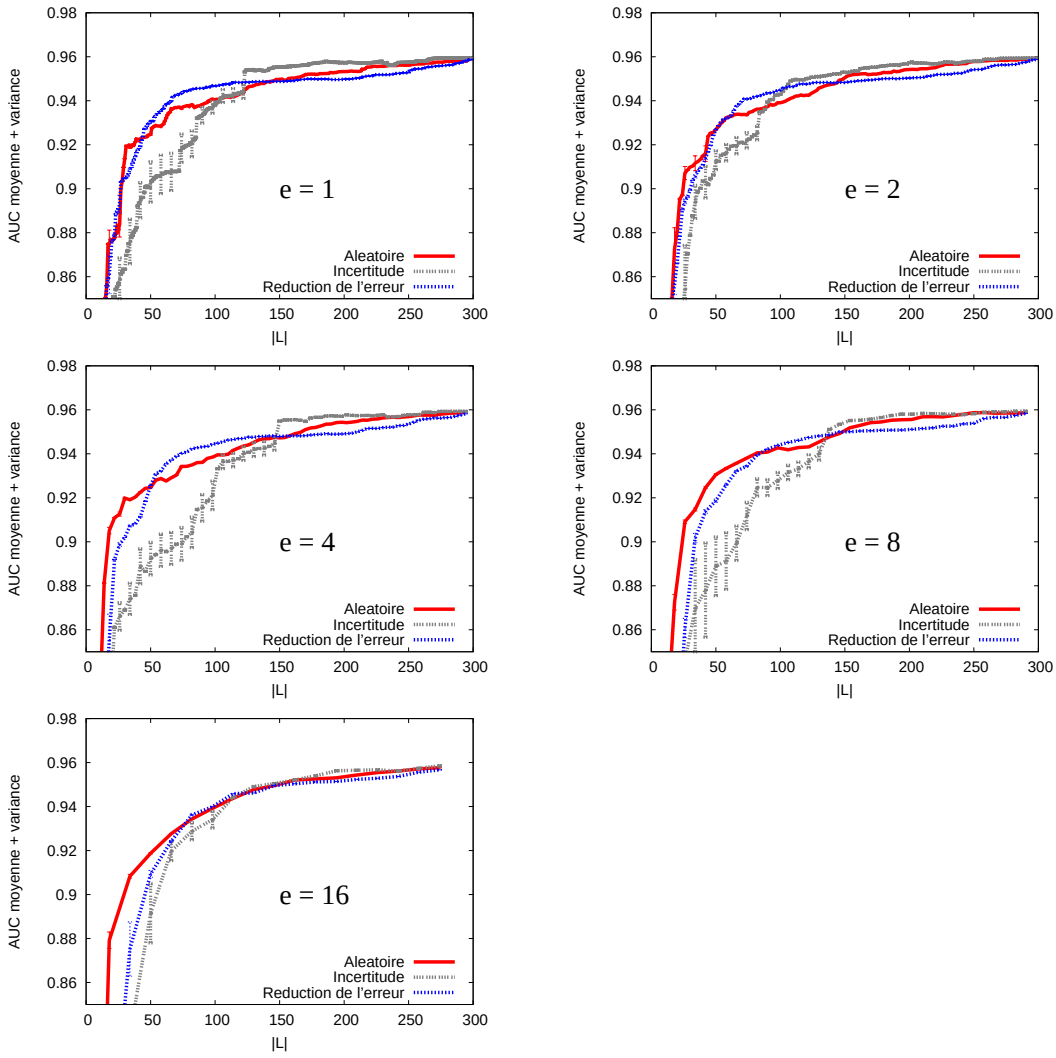

FIG. D.i – Résultats complets pour le jeu de données "Segment"

En augmentant le nombre d'exemples étiquetés à chaque itération, les stratégies actives sont de moins en moins compétitives par rapport à la stratégie aléatoire. Les Figures D.g, D.h et D.i montrent les performances des différentes stratégies pour e=1,2,4,8,16 sur les trois jeux de données traités. Dans chacun des cas, la stratégie aléatoire devient plus performante que les deux stratégies actives quand e augmente, particulièrement pour $e \geq 8$. Les résultats obtenus à l'issue de cette étude montrent que le nombre d'exemples étiquetés à chaque itération d'un échantillonnage sélectif influence la qualité du modèle prédictif. Les expériences réalisées ici confirment notre intuition : l'apport d'une stratégie active (relativement à la stratégie aléatoire) diminue lorsque le nombre d'exemples étiquetés à chaque itération augmente. Cette tendance reste à confirmer par des expériences plus étendues, impliquant des jeux de données de plus grande taille. L'utilisation de jeux de données synthétiques dont on maîtrise les caractéristiques (zones de mélange, bruit ...etc) pourrait aider à la compréhension des comportements des différentes stratégies d'apprentissage actif lorsque e varie.

Liste des figures

Références bibliographiques

[Andrew *et al.*, 1998] cité page 11, 12
K. Andrew, A. McCallum, and K. Nigam. Employing EM in pool-based active learning for text classification. In Jude W. Shavlik, editor, *Proceedings of ICML-98, 15th International Conference on Machine Learning*, pages 350–358, Madison, US, 1998. Morgan Kaufmann Publishers, San Francisco, US.

[Baram *et al.*, 2004] cité page 20, 25, 27, 28, 136
Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *J. Mach. Learn. Res.*, 5 :255–291, 2004.

[Berger, 2006] cité page 24, 71, 86
J. Berger. The case of objective bayesian analysis. *Bayesian Analysis*, 1(3) :385–402, 2006.

[Blum and Mitchell, 1998] cité page 62
A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT : Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, pages 92–100, 1998.

[Bondu and Lemaire, 2007a] cité page 31
A. Bondu and V. Lemaire. Active Learning using Adaptive Curiosity. *ICER*, 2007.

[Bondu and Lemaire, 2007b] cité page 5
A. Bondu and V. Lemaire. Etat de l'art sur les méthodes statistiques d'apprentissage actif. *Revue des Nouvelles Technologie de l'Information (RNTI), Numéro spécial sur l'apprentissage et la fouille de données*, 2007.

[Bondu and Lemaire, 2008a] cité page 31
A. Bondu and Lemaire. A Multi-criterion Active Learning strategy : Application to Emotion Detection in Speech. In *CAP (Conférence francophone sur l'apprentissage automatique)*, pages 21–36, Ile de Porquerolles, May 2008.

[Bondu and Lemaire, 2008b] cité page 31
A. Bondu and V. Lemaire. Adaptive Curiosity for Emotion Detection in Speech. In *IJCNN (International Joint Conference on Neural Networks)*, Hong Kong, june 2008.

[Bondu *et al.*, 2007a] cité page 31
A. Bondu, V. Lemaire, and B. Poulain. Active Learning Strategies : a case study for detection of emotions in speech. In *ICDM' (Industrial Conference of Data Mining)*, Leipzig, july 2007.

[Bondu *et al.*, 2007b] cité page 31
A. Bondu, V. Lemaire, and B. Poulain. Apprentissage actif d'émotions dans les dialogues Homme-Machine. In *EGC (Extraction et Gestion de Connaissances)*, volume 2, pages 427–433, Namur, 2007.

[Bondu *et al.*, 2008] cité page 59
A. Bondu, M. Boullé, and V. Lemaire. A Non-parametric Semi-supervised Discretization Method. In *Soumis à ICDM (International Conference on DataMining)*, Pise, december 2008.

[Boullé, 2004] cité page 69
M. Boullé. Khiops : a statistical discretization method of continuous attributes. *Machine Learning*, 55(1) :53–69, 2004.

[Boullé, 2006a] cité page 52
M. Boullé. An enhanced selective naive bayes method with optimal discretization. In I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors, *Feature extraction, foundations and Application*, pages 499–507. Springer, August 2006.

[Boullé, 2006b] cité page 3, 3, 29, 57, 60, 70, 73, 76, 79, 84, 84, 87, 92, 95
M. Boullé. MODL : A bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1) :131–165, 2006.

[Boullé, 2007a] cité page 29, 57
M. Boullé. Performance prediction challenge. In *IJCNN : International Joint Conference on Neural Networks*, pages 2958–2965, 2007.

[Boullé, 2007b] cité page 3, 67, 71, 71, 79, 83
M. Boullé. *Recherche d'une représentation des données efficace pour la fouille des grandes bases de données*. Phd thesis, ENST (Ecole Nationale Supérieur des Télécommunications), 2007.

[Bradley, 1997] cité page 132
A.P. Bradley. The use of area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7) :1145–1159, 1997.

[Breiman *et al.*, 1984] cité page 132
L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[Brinker, 2003] cité page 21, 28
K. Brinker. Incorporating Diversity in Active Learning with Support Vector Machines. In *ICML : International Conference on Machine Learning*, pages 59–66, 2003.

[Castro and Nowak, 2005] cité page 2, 24
R. Castro, R. andWillett and R. Nowak. Faster rate in regression via active learning. In *NIPS (Neural Information Processing Systems)*, Vancouver, 2005.

[Castro and Nowak, 2008] cité page 57, 108
R. Castro and R. Nowak. *Foundations and Application of Sensor Management*, chapter Active Learning and Sampling. Springer-Verlag, 2008.

[Catlett, 1991] cité page 69
J. Catlett. On changing continuous attributes into ordered discrete attributes. In *EWSL-91 : Proceedings of the European working session on learning on Machine learning*, pages 164–178, New York, NY, USA, 1991. Springer-Verlag New York, Inc.

[Chapelle *et al.*, 2007] cité page 1
O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 2007.

[Chappelle, 2005] cité page 22, 24, 28, 54, 60, 141
O Chappelle. Active learning for parzen windows classifier. In *AI & Statistics*, pages 49–56, Barbados, 2005.

[Christopher and Schütze, 1999] cité page 51
D. Christopher and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[Cohn *et al.*, 1994] cité page 12
David A. Cohn, Les Atlas, and Richard E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2) :201–221, 1994.

[Cohn *et al.*, 1995] cité page 16
David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press, 1995.

[Cohn *et al.*, 2003] cité page 61
D. Cohn, R. Caruana, and A. McCallum. Semi-supervised clustering with user feedback. Technical Report TR2003-1892, Cornell University, 2003.

[Dasgupta, 2005] cité page 13, 16
S. Dasgupta. Analysis of greedy active learning strategy. In *NIPS (Neural Information Processing Systems)*, San Diego, 2005.

[Dima and Hebert, 2005] cité page 10
Cristian Dima and Martial Hebert. Active learning for outdoor obstacle detection. In *Proceedings of Robotics : Science and Systems*, Cambridge, June 2005.

[D.J. Newman and Merz, 1998] cité page 67, 142
C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases. http ://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[Dougherty *et al.*, 1995] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and Unsupervised Discretization of Continuous Features. In *International Conference on Machine Learning*, pages 194–202, 1995.

[Fawcett, 2003] cité page 27, 132, 132, 134
T. Fawcett. Roc graphs : Notes and practical considerations for data mining researchers. T. Fawcett. ROC Graphs : Notes and Practical Considerations for Data Mining Researchers. Technical Report HPL-2003-4, HP Labs, 2003., 2003.

[Fayyad and Irani, 1992] cité page 69
U. Fayyad and K. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8 :87–102, 1992.

[Ferrière, 1922] cité page 1
A. Ferrière. *L'école active*. Editions Forums, 1922.

[Fralick, 1967] cité page 61
S.C. Fralick. Learning to recognize patterns without a teacher. *IEEE Transaction on Information Theory*, 13 :57–64, 1967.

[Freinet, 1964] cité page 1
C. Freinet. *Les invariants pédagogiques*. Bibliothèque de l'école moderne, 1964.

[Freund *et al.*, 1997] cité page 11
Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3) :133–168, 1997.

[Guide *et al.*, 2003] cité page 53
V.A. Guide, Rakotomamonjy, and S. Canu. Méthode à noyaux pour l'identification d'émotion. In *RFIA (Reconnaissance des Formes et Intelligence Artificielle)*, 2003.

[Hanley and McNeil, 1982] cité page 132
A.P. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143 :29–36, 1982.

[Harmon, 1996] cité page 26
M. Harmon. Reinforcement learning : a tutorial. http ://eureka1.aa.wpafb.af.mil/ rltutorial/, 1996.

[Holte, 1993] cité page 68
R.C. Holte. Very simple classification rules perform well on most commonly used dataset. *Machine Learning*, 11 :63–90, 1993.

[Horstein, 1963] cité page 95, 98, 108, 111, 128
M. Horstein. Sequential decoding using noiseless feedback. In *IEEE Transmition Information Theory*, volume 9, pages 136–143, 1963.

[Huang and Ling, 2005] cité page 27
J. Huang and C.X. Ling. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3) :299–310, 2005.

[Jain *et al.*, 1999] cité page 1
A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering : a review. *ACM Computing Surveys*, 31(3) :264–323, 1999.

[Jamy *et al.*, 2005] cité page 1
Irène Jamy, Tao-Yuan Jen, Dominique Laurent, Georges Loizou, and Oumar Sy. Extraction de règles d'association pour la prédiction de valeurs manquantes. *Revue Africaine de la Recherche en Informatique et Mathématique Appliquée ARIMA*, Spécial CARI04 :103–124, 2005.

[Junqua and Halton, 1995] cité page 51
J.C. Junqua and J.P. Halton. *Robustness in Automatic Speech Recognition : Fundamentals and Applications*. Springer, 1995.

[Kerber, 1991] cité page 69, 124
R. Kerber. Chimerge discretization of numeric attributes. In *Procceedings of the 10th International Conference on Artificial Intelligence*, pages 123–128, 1991.

[Kohavi and Sahami, 1996] cité page 68
R. Kohavi and M. Sahami. Error-Based and Entropy-Based Discretization of Continuous Features. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 114–119, 1996.

[Kononenko *et al.*, 1984] cité page 69
I. Kononenko, I. Bratko, and E. Roskar. Experiments in automatic learning of medical diagnostic rules. Technical report, Ljubljana : Joseph Stefan Institute, Faculty of Electrical Engineering and Computer Science, 1984.

[Kothari and Jain, 2003] cité page 25
R. Kothari and V. Jain. Learning From Labeled and Unlabeled Data Using a Minimal Number of Queries. *IEEE Transactions on Neural Network*, 14(6) :1496–1505, 2003.

[Lechevalier, 1990] cité page 69
Y. Lechevalier. Recherche d'une partition optimale sous contrainte d'ordre total. Rapport technique 1247, INRIA, 1990.

[Lemaire *et al.*, 2007] cité page 5, 141
V. Lemaire, A. Bondu, and F. Clérot. Purchase of data labels by batches : study of the impact on the planning of two active learning strategies. *ICONIP*, 2007.

[Lewis and Gale, 1994] cité page 9
D. Lewis and A. Gale. A sequential algorithm for training text classifiers. In W. Bruce Croft and Cornelis J. van Rijsbergen, editors, *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin, 1994. Springer Verlag, Heidelberg.

[Liscombe *et al.*, 2005] cité page 51
J. Liscombe, G. Riccardi, and D. Hakkani-Tür. Using context to improve emotion detection in spoken dialog systems. In *InterSpeech*, Lisbon, 2005.

[Maass, 1994] cité page 68
W. Maass. Efficient agnostic pac-learning with simple hypothesis. In *COLT'94 : Proccedings of the seventh annual conference on Computational learning theory*, pages 67–75, 1994.

[Maeireizo *et al.*, 2004] cité page 62
B. Maeireizo, D. Litman, and R. Hwa. Co-training for predicting emotions with spoken dialogue data. In *ACL '04 : The Companion Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.

[Muslea *et al.*, 2002] cité page 22, 28, 60
I. Muslea, S. Minton, and C.A. Knoblock. Active + Semi-supervised Learning = Robust Multi-View Learning. In *ICML '02 : Proceedings of the Nineteenth International Conference on Machine Learning*, pages 435–442. Morgan Kaufmann Publishers, 2002.

[Muslea, 2002] cité page 7
I. Muslea. *Active Learning With Multiple View.* Phd thesis, University of southern california, 2002.

[Nagai *et al.*, 2002] cité page 33
Y. Nagai, M. Asada, and K. Hosoda. Developmental learning model for joint attention. In *Proceedings of the 15th International Conference on Intelligent Robots and Systems (IROS)*, pages 932–937, 2002.

[Naoki and Hiroshi, 1998] cité page 11, 11
A. Naoki and M. Hiroshi. Query learning strategies using boosting and bagging. In *ICML '98 : Proceedings of the Fifteenth International Conference on Machine Learning*, pages 1–9, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[Nguyen and Smeulders, 2004] cité page 20, 28
H.T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *ICML '04 : Proceedings of the twenty-first international conference on Machine learning*, page 79, New York, NY, USA, 2004. ACM.

[Nigam and Rayid, 2000] cité page 63
K. Nigam and G. Rayid. Analyzing the effectiveness and applicability of co-training. In *CIKM*, pages 86–93, 2000.

[Osugi *et al.*, 2005a] cité page 20
T. Osugi, D. Kun, and S. Scott. Balancing exploration and exploitation : A new algorithm for active machine learning. In *ICDM '05 : Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 330–337, Washington, DC, USA, 2005. IEEE Computer Society.

[Osugi *et al.*, 2005b] cité page 50
T. Osugi, D. Kun, and S. Scott. Balancing exploration and exploitation : A new algorithm for active machine learning. In *Proceedings of the Fith IEEE International Conference on Data Mining (ICDM'05)*, 2005.

[Oudeyer and Kaplan, 2004] cité page 4, 32, 33, 33, 33, 36, 36, 36, 37
P-Y. Oudeyer and F. Kaplan. Intelligent adaptive curiosity : a source of self-development. In Luc Berthouze, Hideki Kozima, Christopher G. Prince, Giulio Sandini, Georgi Stojanov, G. Metta, and C. Balkenius, editors, *Proceedings of the 4th International Workshop on Epigenetic Robotics*, volume 117, pages 127–130. Lund University Cognitive Studies, 2004.

[Parzen, 1962] cité page 17, 22, 53, 141
E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[Poulain, 2006] cité page 52, 52
B Poulain. Sélection de variables et modélisation d'expressions d'émotions dans des dialogues hommes-machine. In *EGC (Extraction et Gestion de Connaissance)*, Lille. + Note technique http ://perso.rd.francetelecom.fr/lemaire, 2006.

[Quinlan, 1986] cité page 69
J.R. Quinlan. Introduction of decision trees. *Machine Learning*, 1 :81–106, 1986.

[Quinlan, 1993] cité page 69
J.R. Quinlan. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann, 1993.

[Richard and Cappé, 2004] cité page 51
G. Richard and O. Cappé. Synthèse de la parole à partir du texte. Rapport technique ENST, Télécom Paris, 2004.

[Robert, 2006] cité page 70, 125
C.P. Robert. *Le choix bayésien Principes et pratique*. Springer, 2006.

[Roy and McCallum, 2001] cité page 2, 6, 17
Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. 18th International Conf. on Machine Learning*, pages 441–448. Morgan Kaufmann, San Francisco, CA, 2001.

[Safavian and Landgrebe, 1991] cité page 124
S.C. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics*, 21(3) :660–674, 1991.

[Sarle, 1994] cité page 37
Warren S. Sarle. Neural networks and statistical models. In *Proceedings of the Nineteenth Annual SAS Users Group International Conference, April, 1994*, pages 1538–1550, Cary, NC, 1994. SAS Institute.

[Scott, 1979] cité page 68
D.W. Scott. Averaged shifted histograms : Effective nonparametric density estimator in several dimensions. *Annals of statistic*, 13(3) :1024–1040, 1979.

[Seung *et al.*, 1992] cité page 10
H. S. Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Computational Learning Theory*, pages 287–294, 1992.

[Shannon, 1948] cité page 43, 69, 70, 88, 124
C.E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3) :379–423, 1948.

[Singh *et al.*, 2006] cité page 2, 6
Aarti Singh, Robert Nowak, and Parmesh Ramanathan. Active learning for adaptive mobile sensing networks. In *IPSN '06 : Proceedings of the fifth international conference on Information processing in sensor networks*, pages 60–68, New York, NY, USA, 2006. ACM Press.

[Sturges, 1926] cité page 68
H.A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21 :65–66, 1926.

[Sugiyama and Müller, 2005] cité page 64
M. Sugiyama and K.R. Müller. Model Selection Under Covariate Shift. In *ICANN, International Conference on Computational on Artificial Neural Networks : Formal Models and Their Applications*, 2005.

[Sugiyama *et al.*, 2007] cité page 22, 25, 29, 64, 64
M. Sugiyama, M. Krauledat, and K.R. Müller. Covariate Shift Adaptation by Importance Weighted Cross Validation. 8 :985–1005, 2007.

[Thrun and Möller, 1992] cité page 9
Sebastian B. Thrun and Knut Möller. Active exploration in dynamic environments. In John E. Moody, Steve J. Hanson, and Richard P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 531–538. Morgan Kaufmann Publishers, Inc., 1992.

[Tong and Koller, 2000] cité page 12, 23
Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. In Pat Langley, editor, *Proceedings of ICML-00, 17th International Conference on Machine Learning*, pages 999–1006, Stanford, US, 2000. Morgan Kaufmann Publishers, San Francisco, US.

[Utgoff, 1989] cité page 125
P.E. Utgoff. Incremental Induction of Decision Trees. *Machine Learning*, 4 :161–186, 1989.

[White, 1959] cité page 33
R. White. Motivation reconsidered : The concept of competence. *Psychological Review*, 66 :297–333, 1959.

[Xu *et al.*, 2007] cité page 21, 28
Z. Xu, R. Akella, and Y. Zhang. Incorporating diversity and density in active learning for relevance feedback. In *ECIR (European Conference on Information Retrieval)*, volume 4425, pages 246–257, 2007.

[Zhu *et al.*, 2003] cité page 17, 60
X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML (International Conference on Machine Learning)*, Washington, 2003.

[Zhu, 2005] cité page 62, 63, 65
X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[Zighed and Rakotomalala, 2000] cité page 81
D.A. Zighed and R. Rakotomalala. *Graphes d'induction*. Hermes, France, 2000.

[Zighed *et al.*, 1998] cité page 69
D.A Zighed, S. Rabaseda, and R. Rakotomalala. Fusinter : a method for discretization of continuous attributes for supervised learning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(33) :307–326, 1998.

[Zoller and Buhmann, 2000] cité page 20, 21, 28
T. Zoller and M. Buhmann. Active learning for hierarchical pairwise data clustering. In *ICPR '00 : Proceedings of the 15th International Conference on Pattern Recognition*, pages 186–189, 2000.

Apprentissage Actif par Modèles Locaux

Résumé

Les méthodes d'apprentissage statistiques exploitent des exemples, pour enseigner un comportement à un modèle prédictif. La classification supervisée requiert des exemples étiquetés. En pratique, l'étiquetage des exemples peut se révélé coûteux. Dans certain cas, l'étiquetage implique un expert humain, un instrument de mesure, un temps de calcul élevé...etc. Les méthodes *d'apprentissage actif* réduisent le coût de préparation des données d'apprentissage. Ces méthodes cherchent à étiqueter uniquement les exemples les plus utiles à l'apprentissage d'un modèle. Les travaux présentés dans ce manuscrit sont réalisés dans le cadre de *l'échantillonnage sélectif*, qui n'autorise pas les stratégies actives à générer de nouveaux exemples d'apprentissage. Les stratégies actives de la littérature utilisent généralement des modèles globaux à l'espace des variables d'entrées. Nous proposons dans ce manuscrit une stratégie originale qui effectue un partitionnement dichotomique récursif de l'espace d'entrée. Cette stratégie met en compétition les modèles locaux à chacune des zones, pour choisir les exemples à étiqueter. Notre stratégie décide "*quand*" couper une zone et "*où*" la couper. Une amélioration possible consiste à exploiter une méthode de discrétisation pour prendre ces deux décisions. L'extension de l'approche de discrétisation MODL au cas de l'apprentissage semi-supervisé constitue un des apports majeurs de cette thèse. Nous proposons une deuxième amélioration qui consite à sélectionner, localement à la meilleure zone, l'exemple le plus utile à l'apprentissage du modèle local. Nous proposons une stratégie active originale, qui maximise la probabilité des modèles de discrétisation connaissant les données et l'exemple candidat à l'étiquetage.

Mots-clés : Apprentissage Actif, Modèles locaux, Discrétisation Bayésienne.

Active Learning using Local Models

Abstract

Supervised classification problems requires labelled examples, and labelling step can be costly in practice. *Active learning* strategies reduce the cost of preparing learning data. These strategies aim to label only the most useful examples for the learning of the predictive model. This thesis proposes a new active learning strategy which carries out a recursive binary partitionning of the input space. This strategy handles severals local predictive models in each zones, and chooses examples to be labelled. Our strategy decides "*when*" and "*where*" a zone must be cut. A possible improvement consists in exploiting a discretization methode to make both decisions. The extension of the MODL discretisation approach to the semi-supervised learning constitutes an importante contribution of this thesis. We propose a second improvement which aims to select, locally in the best zone, the most useful example for the training of the local model. We propose an active learning strategy based on the semi-supervised MODL approach, which maximizes the probability of discretization models given the data.

Keywords : Active Learning, Local Models, Bayesian Discretization.