Apprentissage Actif de Données Incertaines et Imprécises

Arthur Hoarau

COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, télécommunications, informatique, signal, systèmes, électronique
Spécialité : *INFO*

Par
Arthur HOARAU

Apprentissage actif de données incertaines et imprécises

Présenté et soutenu à Lannion, le 13 juin 2024
Unité de recherche : IRISA - DRUID

Rapporteurs avant soutenance

Eric LEFEVRE	Professeur des universités, Université d'Artois LGI2A
Marie-Jeanne LESOT	Professeur des universités, Sorbonne Université, LIP6

Composition du Jury

Examinateurs :

Eric ANQUETIL	Professeur des universités, INSA Rennes, IRISA (Président)
Sébastien DESTERCKE	Directeur de recherche, Université de Technologie de Compiegne, CNRS
Marie-Jeanne LESOT	Professeur des universités, Sorbonne Université, LIP6
Eric LEFEVRE	Professeur des universités, Université d'Artois, LGI2A
Vincent LEMAIRE	Research Scientist, HDR, Orange Innovation
Zoltan MIKLOS	Maître de conférences HDR, Université de Rennes, IRISA (Direction)
Jean-Christophe DUBOIS	Maître de conférences, Université de Rennes, IRISA (Co-Encadrement)

Invitée :

Yolande LE GALL	Maître de conférences, Université de Rennes, IRISA (Co-Encadrement)

Directeurs de thèse :

Arnaud MARTIN	Professeur des universités, Université de Rennes, IRISA (Posthume)
Zoltan MIKLOS	Maître de conférences HDR, Université de Rennes, IRISA

À mon père,
À Arnaud,
Pour ma mère.

La mort est, comme la naissance, un mystère de la nature : combinaison dans l'une des mêmes éléments qui se séparent dans l'autre. En somme, rien dont on puisse être déshonoré, car mourir n'est pas contraire à la disposition d'un animal raisonnable, ni la logique de sa constitution.

Marc Aurèle, *Pensées pour moi-même*, Livre IV-V.

TABLE DES MATIÈRES

Figure-toi qu'il y a dans la tête, c'est-à-dire dans le cerveau, des nerfs... Ces nerfs ont des fibres, et dès qu'elles vibrent... Tu vois, je regarde quelque chose, comme ça, et elles vibrent, ces fibres... et aussitôt qu'elles vibrent, il se forme une image, pas tout de suite, mais au bout d'un instant, d'une seconde, et il se forme un moment... non pas un moment, je radote... mais un objet ou une action ; voilà comment s'effectue la perception. La pensée vient ensuite... parce que j'ai des fibres, et nullement parce que j'ai une âme et que je suis créé à l'image de Dieu ; Quelle belle chose que la science ! L'homme se transforme, je le comprends... Pourtant, je regrette Dieu !

Fédor Dostoïevski, *Les frères Karamazov*, Livre XI-IV.

Chapitre 1

Introduction

Cette thèse intitulée *Apprentissage actif de données incertaines et imprécises* est articulée autour de deux grandes thématiques qui sont la labellisation [1] en apprentissage automatique et l'imperfection. L'objectif est de les combiner afin de travailler à la fois dans un environnement de labellisation imparfaite, mais également en réduisant les coûts de labellisation, tout en améliorant les performances de classification.

L'introduction suivante vise à présenter l'intérêt d'utiliser les fonctions de croyance et l'apprentissage actif ainsi que les objectifs attendus. L'imperfection dans les données est d'abord présentée au travers de l'incertitude et de l'imprécision, et pour plus de clarté, le terme de qualité des labels est introduit. Ensuite, un rappel sur l'apprentissage supervisé et semi-supervisé introduit l'apprentissage actif afin de travailler sur la quantité de labels. Enfin, ces deux notions de qualité et de quantité sont mises en relation pour décrire les objectifs de cette thèse.

1.1 S'intéresser à la qualité et à la quantité de labels

Apprentissage automatique

L'apprentissage automatique, inscrit dans le champ d'études de l'intelligence artificielle, regroupe de nombreuses techniques visant à apprendre à partir de données. L'apprentissage supervisé, non supervisé, semi-supervisé ou par renforcement, constituent des exemples d'apprentissages automatiques (*cf.* Russell et Norvig 2010). Les deux méthodes d'apprentissage les plus répandues sont l'apprentissage supervisé et l'apprentissage non supervisé. Ces deux méthodes sont présentées ci-après. Une autre méthode, l'apprentissage semi-supervisé, se situe à la frontière entre ces deux notions, elle sera également présentée et permettra d'introduire l'apprentissage actif qui nous intéresse dans ce document.

1. L'anglicisme "label", lui-même issu de l'ancien français, est adopté dans ce document pour faire référence aux étiquettes attribuées aux observations.

Apprentissage supervisé

Il s'agit en apprentissage supervisé d'effectuer une prédiction à partir de données et d'informations relatives à ces données. Elle peut concerner la météo, la race d'un chien présent sur une photo ou encore la détection d'un cancer. Deux scénarios se distinguent alors, la régression, lorsque la prédiction est faite sur une variable quantitative [2] et la classification, lorsque la prédiction est faite sur une variable qualitative. Pour le problème de classification, celui principalement traité dans ce document, un expert (ou oracle) classe (ou labellise) des observations qui seront utilisées par un modèle dans une phase d'apprentissage. Une seconde phase de prédiction intervient ensuite pour prédire la classe d'une nouvelle observation, inconnue jusqu'alors du modèle. Parmi les modèles les plus connus, on retrouve le modèle des K plus proches voisins, introduit par Fix et Hodges 1951, les arbres de décision, de Leo Breiman, Friedman et al. 1984, les forêts aléatoires, de Leo Breiman 2001 ou encore certains réseaux de neurones artificiels (voir Abiodun, Jantan et al. 2018).

Apprentissage non supervisé

Dans l'apprentissage non supervisé, il n'y pas de notion d'expert ou de prédiction, le modèle doit, sans aide extérieure, découvrir les structures existantes au sein des données non labellisées. Le principal objectif est alors de faire des regroupements ou de partitionner les observations, comme par exemple regrouper les photos ressemblantes, partitionner une classe d'étudiants en fonction de leurs résultats, ou encore associer entre eux des individus en fonction de caractéristiques particulières. Les K-moyennes, de Steinhaus 1957, ou encore la classification hiérarchique constituent des méthodes bien connues.

Apprentissage semi-supervisé

Lorsque le problème est divisé en données labellisées et non labellisées, on ne se trouve alors ni en apprentissage non supervisé ni en apprentissage supervisé, on parle ainsi d'apprentissage semi-supervisé. L'objectif est d'améliorer les performances du modèle en utilisant à la fois les données labellisées et les données non labellisées. L'apprentissage semi-supervisé peut être utilisé à la fois pour des problématiques de partitionnement (*cf.* D. Cohn, Caruana et al. 2008) ou pour des problèmes de classification (*cf.* Chapelle et Zien 2005).

2. La notion de variable statistique (quantitative, qualitative) est abordée dans le chapitre 3.1.

Qualité et quantité de labels

L'apprentissage automatique est aujourd'hui démocratisé et de plus en plus répandu dans de nombreux domaines d'applications. Les récents exploits (voir ThinkML-Team 2022) principalement liés à l'apprentissage profond, en vision par ordinateur, en traitement automatique des langues, dans le domaine médical, sur les réseaux sociaux, en robotique ou encore en agriculture, en font un domaine en expansion, lucratif et souvent relayé dans les médias. Parmi les nombreuses thématiques de recherche gravitant autour du *Machine Learning*[3] cette thèse s'intéresse aux labels. Ces étiquettes apportent un contexte significatif permettant à un modèle d'apprendre.

Les récents modèles utilisent une quantité de données labellisées de plus en plus grande, le jeu de données ImageNet, publié par Deng, Dong et al. 2009, compte par exemple 12 197 122 images[4]. L'acquisition d'un tel nombre de labels est coûteuse et peut donner lieu à des erreurs lorsque des humains s'en chargent, ou encore à certaines dérives. Le magazine Time a dévoilé un article, de Perrigo 2023, accusant OpenAI, l'entreprise à l'origine de Chat-GPT, GPT-3 ou encore Dall-E 2, d'utiliser des travailleurs kenyans payés moins de 2$ de l'heure pour labelliser des contenus violents.

C'est donc pour des raisons économiques, écologiques ou sociétales, que les labels sont au cœur de nombreuses thématiques de recherche et cette thèse s'intéresse en particulier à deux problèmes ; *Comment obtenir une meilleure modélisation des labels ?* et *Comment réduire le nombre de données labellisées ?*

1.2 Qualité : Labellisation imparfaite

En classification, où l'on s'intéresse principalement à trouver la classe réelle d'une observation en utilisant des connaissances, des labels durs sont bien souvent utilisés. Autrement dit, s'il existe un label pour une observation, ce label est défini catégoriquement. Le label "Chat" peut être associé à une image de chat, ou encore le label "Feu rouge" à des données issues de capteurs présents sur une voiture autonome.

Cette perfection dans les labels peut s'avérer suffisante pour de nombreux problèmes d'apprentissage automatique et d'apprentissage profond, mais n'est jamais complètement représentative de la réalité. Ce processus de labellisation peut être, et est bien souvent réalisé par des humains, comme évoqué par Fredriksson, Issa Mattos et al. 2020 et par Roh,

3. Apprentissage automatique en anglais.
4. La version la plus utilisée d'ImageNet compte 1 281 167 images d'entrainement labellisées.

Heo et al. 2021. Cette approche ne permet pas de différencier une image labellisée par un individu ignorant la réponse ou hésitant entre plusieurs réponses, de celle labellisée par un individu certain de sa réponse. L'imperfection dans les labels, quant à elle, peut aider à représenter l'information telle qu'elle est, sans altération. Elle peut être représentée à travers plusieurs critères décrits par Philippe Smets 1997 dont l'incertitude et l'imprécision. L'ignorance peut également être modélisée.

Incertitude

L'incertitude se définit par une connaissance partielle de la vraie valeur de la donnée. La phrase "Il va peut-être pleuvoir' représente une incertitude. Dans le cas de la classification, une image labellisée "C'est peut-être un chat" s'inscrit dans cette représentation incertaine.

Imprécision

L'impréciston se traduit par un défaut quantitatif de connaissance. Dans la phrase "Il va pleuvoir cette semaine", l'imprécision est présente, car le jour exact de pluie est inconnu. En classification, une image labellisée "C'est un chien ou un chat" représente une imprécision.

Une autre approche, celle de Klir et Wierman 1998 est de représenter l'imprécision comme un type d'incertitude dans la théorie de l'information (et non de les dissocier). L'incertitude est alors un concept plus grand, regroupant le vague, l'imprécision, et le conflit.

Vague

Le vague, résulte d'un flou dans les frontières d'un ensemble, par exemple dans la phrase "Il est grand", le mot grand est vague, la frontière est floue et dans une représentation en centimètres on ne sait pas si le terme correspond à une taille précise.

Imprécision

L'imprécision, ou non-spécificité, est la même que celle définie précédemment et elle est connectée à la cardinalité de l'événement.

Conflit

Le conflit, ou plutôt la discorde, correspond à l'information conflictuelle au sein de l'événement, par exemple "Il fait beau et il pleut".

Toute cette imperfection qui peut être exprimée par un humain n'est pas représentable dans un label dur pour une observation donnée. Il existe plusieurs cadres, théories et modèles qui permettent de représenter ces imperfections, certains sont présentés ici.

Probabilités

Les probabilités ne permettent de représenter qu'un seul type d'incertitude (*cf.* Klir et Wierman 1998) et la logique bayésienne utilisée pour modéliser, d'après Shafer 1976, des croyances plutôt que des chances, est souvent critiquée pour sa modélisation controversée de l'ignorance.

Exemple : de la vie proche de Sirius ? Des scientifiques se demandent s'il y a de la vie orbitant autour de l'étoile Sirius. Deux événements sont décrits ; ω_1 : "*Il y a de la vie proche de Sirius*" et son complémentaire ω_2 : "*Il n'y a pas de vie proche de Sirius*". Dans le cas d'une ignorance totale et en probabilités bayésiennes, il est parfois défendu de distribuer l'ignorance équitablement entre les événements, on a donc : $P_{\omega_1} = P_{\omega_2} = \frac{1}{2}$, respectivement les probabilités associées aux événements ω_1 et ω_2. Cette influence est attribuée à Pierre Simon Laplace, décrivant les chances uniquement comme une caractéristique de la connaissance. On considère également une autre question, impliquant la présence de planètes, telle que ; θ_1 : "*Il y a de la vie proche de Sirius*", θ_2 : "*Il y a des planètes, mais pas de vie proche de Sirius*" et θ_3 : "*Il n'y a même pas de planètes proches de Sirius*". Suivant le même raisonnement, en présence d'ignorance on a $P_{\theta_1} = P_{\theta_2} = P_{\theta_3} = \frac{1}{3}$. Puisque ω_1 et θ_1 ont le même sens et ω_2 a le même sens que θ_2 et θ_3 on se retrouve dans un cas absurde puisque $\frac{1}{2} \neq \frac{1}{3}$ et $\frac{1}{2} \neq \frac{2}{3}$ (autrement dit $P_{\omega_1} \neq P_{\theta_1}$ et $P_{\omega_2} \neq P_{\theta_2} + P_{\theta_3}$).

Approches floues

Lotfi A. Zadeh définit en 1965 les sous-ensembles flous (voir Zadeh 1965), permettant de représenter le vague. Contrairement aux ensembles classiques, les frontières d'un sous-ensemble flou ne sont pas précises. Le changement d'appartenance à non-appartenance d'un membre est exprimé graduellement par une *fonction d'appartenance*. Un individu mesurant 1m80 peut être plus ou moins *grand* selon sa fonction d'appartenance et un aliment plus ou moins *périmé* selon sa date d'achat, là où la théorie des ensembles définit une appartenance binaire : cet aliment est périmé ou cet aliment n'est pas périmé. La théorie des ensembles est notamment incluse dans celle des sous-ensembles flous. Bien qu'elle soit répandue dans la littérature, la théorie des sous-ensembles flous ne permet de représenter que le vague.

Possibilités

Le même auteur (voir Zadeh 1978) propose une théorie des possibilités pour traiter certains types d'incertitude comme une alternative aux probabilités. D. Dubois et Prade 2001 contribuent par la suite au développement de cette théorie. Les notions de possibilité et de nécessité sont introduites. La possibilité correspond, dans une représentation probabiliste, à la borne supérieure d'une probabilité, par exemple il est possible de *gagner au Loto*, pourtant cela est peu probable. La nécessité est le dual de la possibilité, il correspond à la possibilité non attribuée au complémentaire. Par exemple s'il est possible qu'il neige demain, la nécessité qu'il fasse beau est nulle, ou bien s'il est nécessaire que je sois un humain, il est impossible pour moi de ne pas être humain. Dans l'exemple de la vie proche de Sirius, l'ignorance totale peut être représentée par une possibilité maximale et une nécessité nulle : il est totalement possible qu'il y ait de la vie proche de Sirius et il est totalement possible qu'il n'y en ait pas.

Théorie des fonctions de croyance

Arthur P. Dempster propose en 1967 une nouvelle représentation introduisant des probabilités haute et basse (*cf.* Dempster 1967), ces travaux seront repris par Shafer 1976 dans sa théorie des fonctions de croyance [5] et les probabilités haute et basse sont respectivement renommées plausibilité et crédibilité.

5. *A Mathematical Theory of Evidence* est le nom donné à la théorie, le terme de croyance est utilisé dans ce document pour remplacer l'anglais *evidence*.

Pour Shafer 1976, l'idée de chance et l'idée de croyance ont été réunies sous le même terme de probabilité. La notion de chance décrit une expérience aléatoire, comme le lancer d'un dé, mais la chance associée à cette expérience peut ne pas coïncider avec notre degré de croyance concernant son résultat. "Si nous connaissons les chances, alors nous les adopterons assurément comme degré de croyance, mais sans connaître les chances alors la coïncidence serait extraordinaire pour qu'elles soient égales à notre degré de croyance".

La principale différence entre la théorie des probabilités et la théorie des fonctions de croyance est la non-additivité des mesures pour deux événements disjoints. Pour l'exemple de Sirius, la probabilité *qu'il y ait de la vie ou qu'il n'y ait pas de vie* vaut 1 et se répartit (sous forme d'une somme) entre les deux événements ω_1 : *Il y a de la vie proche de Sirius* et ω_2 : *Il n'y a pas de vie proche de Sirius*. Dans la théorie des fonctions de croyance, si la croyance *qu'il y ait de la vie ou qu'il n'y ait pas de vie* vaut 1, elle n'est pas forcément répartie sur ω_1 et ω_2. Cette différence permet de modéliser mathématiquement plusieurs degrés d'imprécision et d'incertitude, comme dans la phrase "je ne sais pas quel temps il fera demain, mais je pense qu'il ne neigera pas".

Dans le reste de ce document il sera admis, comme interprété dans les écrits de Shafer, que les statistiques fréquentistes décrivent les chances liées à un événement et une incertitude aléatoire, alors que les statistiques bayésiennes décrivent des croyances liées à un événement et une incertitude épistémique. La théorie des fonctions de croyance permet donc de modéliser des croyances de manière plus complète que les probabilités, en représentant une incertitude et une imprécision.

Outre la représentation de plusieurs degrés d'imprécision, ce cadre permet de généraliser la théorie des probabilités et des possibilités [6]. C'est cette dernière théorie qui est retenue, les raisons d'un tel choix sont explicitées plus tard dans cette introduction. Le vague (ou flou) ne sera donc pas étudié dans ce document. Il existe aussi d'autres théories, comme celle des probabilités imprécises [7] ou encore celles de Sugeno 1993 et de Denœux 2021, qui s'attachent à généraliser les approches floues, possibilités et fonctions de croyance, permettant ainsi de représenter toutes les formes d'incertitude présentées.

6. La probabilité est la croyance attribuée par une fonction de masse bayésienne et la possibilité est la plausibilité attribuée par une fonction de masse consonante. Ces notions sont introduites au chapitre 2.

7. La décomposition en probabilités haute et basse des probabilités imprécises est aussi possible avec la théorie des fonctions de croyance, d'ailleurs on pourrait considérer que l'une fait partie de l'autre.

1.3 Quantité : Apprentissage actif

Si plusieurs approches nous permettent de traiter de la qualité des labels, et de représenter des imperfections, notre objectif est aussi de réduire le nombre d'observations à labelliser. Le coût de labellisation, qu'il soit modélisé économiquement, écologiquement, temporellement ou encore par un impact sociétal est donc la variable à minimiser. Voici quelques domaines qui s'inscrivent dans cette thématique.

Apprentissage semi-supervisé

Déjà présenté dans cette introduction, l'apprentissage semi-supervisé combine une faible quantité de données labellisées avec une grande quantité de données non labellisées. Il s'agit donc d'une approche pouvant répondre à la problématique de coûts liés à la labellisation.

Apprentissage par transfert

Cette méthode d'apprentissage, décrite par Zhuang, Qi et al. 2021, s'intéresse à la connaissance stockée lors de la résolution d'un problème et à son application ultérieure sur une tâche différente. Par exemple, l'information emmagasinée par un modèle de détection de voitures, pourrait être utilisée pour détecter des camions ou des vélos. L'intérêt de ce mode d'apprentissage est d'utiliser une connaissance acquise, une des applications est d'utiliser la connaissance présente, avec un nombre réduit de nouvelles données labellisées, pour entraîner un modèle et ainsi réduire les coûts.

Apprentissage actif

Parfois confondu dans la littérature avec l'apprentissage semi-supervisé, l'apprentissage actif (*cf.* Bondu et Lemaire 2008 ; Settles 2009) s'intéresse à un problème où les données ne sont pas labellisées, mais peuvent l'être en interrogeant un oracle. Le modèle a alors le choix d'étiqueter certaines observations afin de gagner rapidement en performance en utilisant le moins de données labellisées possible. La difficulté réside dans la sélection des observations à labelliser, où l'objectif est de choisir à chaque étiquetage, l'observation la plus pertinente, ce processus s'appelle l'échantillonnage.

C'est cet apprentissage actif qui est retenu dans ce document, il offre notamment une réponse au problème de coût lié à la labellisation des données. Cependant, dans notre cas

l'objectif n'est pas seulement de travailler avec peu de données labellisées, mais également avec des données labellisées imparfaitement, et l'échantillonnage par incertitude, utilisé en apprentissage actif, peut répondre à cette problématique.

1.4 Qualité et quantité : Apprentissage actif sur données imparfaites

La théorie des fonctions de croyance et l'apprentissage actif sont donc retenus pour travailler sur des labels plus riches, capables de représenter une incertitude et une imprécision, mais également pour réduire le nombre de données labellisées. L'apprentissage actif et la théorie des fonctions de croyance ont encore été peu couplés dans la littérature scientifique (*cf.* Hemmer, Kühl et al. 2020 ; Ramel, Pichon et al. 2018). Les travaux précédents de Zhu, Martin et al. 2021 seront en partie repris et approfondis avec l'objectif de pouvoir ajouter de l'information durant la phase de labellisation pour permettre à un modèle de gagner en performance lors d'un apprentissage actif. L'intention ici est de pouvoir travailler avec des données réellement labellisées de manière incertaine et imprécise et non plus uniquement des données bruitées, qui ne sont pas représentatives de la réalité. Il est également souhaité d'apporter des nouveautés à la classification crédibiliste [8], dans le but de modifier ensuite les méthodes d'échantillonnage existantes en apprentissage actif pour utiliser le plein potentiel de la labellisation imparfaite.

Des travaux réalisés dans le but de créer des jeux de données labellisées de manière incertaine et imprécise à l'aide de production participative, poursuivis par Thierry, Hoarau et al. 2022, sont également repris pour obtenir des observations labellisées imparfaitement pour les différentes expérimentations.

Contributions de la thèse

Les travaux associés à cette thèse ont été présentés à la communauté scientifique sous la forme de publications dans des revues et conférences internationales à comité de sélection. Au total, neuf contributions ont été soumises [9], dont cinq ont été publiées, deux sont en cours de soumission et deux sont en préparation pour soumission. Ces articles couvrent

8. Classification couplée à la théorie des fonctions de croyance.

9. Seules les publications scientifiques avec comité de sélection sont présentées, les *posters* scientifiques et autres contributions ont été omis.

quatre thématiques principales : “Labellisation imparfaite”, “Classification crédibiliste”, “Apprentissage Actif” et “Représentation et quantification d’incertitude”. Ces concepts sont progressivement explorés tout au long de ce document.

Labellisation imparfaite

La plupart des études antérieures utilisent généralement un bruit synthétique pour simuler des labels plus riches [10], car l’obtention de labels authentiques pose des défis en termes d’interaction et de collecte auprès des utilisateurs. Ainsi, notre première contribution consiste à collecter des données incertaines et imprécises. Nous avons développé une interface de production participative, présentée par Thierry, Hoarau et al. 2022, qui permet à de véritables contributeurs de labelliser des images d’oiseaux tout en représentant leur niveau d’incertitude et d’imprécision. De plus, nous avons proposé une méthode pour produire des labels riches à partir des réponses fournies par les contributeurs.

Ces premières études théoriques ont abouti à des jeux de données qui ne sont pas adaptés aux besoins de l’apprentissage automatique (*e.g.* un nombre insuffisant d’observations, la présence de multiples contributeurs pour une même observation, des contributions sur l’ensemble du jeu de données, etc). Par conséquent, nous avons entrepris de collecter cinq nouveaux jeux de données labellisés de manière incertaine et imprécise, comme décrit dans l’article de Hoarau, Thierry, Martin et al. 2023, spécifiquement conçus pour l’apprentissage automatique. Ces jeux de données ont été rendus accessibles à la communauté scientifique en libre accès [11].

Classification crédibiliste

La contribution de Hoarau, Martin, J.-C. Dubois et Le Gall 2023, majeure pour ce document, consiste en la proposition d’un nouvel arbre de décision crédibiliste ainsi que l’introduction d’une forêt aléatoire crédibiliste. Ces deux modèles utilisent une distance et un degré d’inclusion pour permettre au modèle de regrouper les observations dont les éléments de réponse sont inclus les uns dans les autres en un seul nœud. Les résultats expérimentaux ont démontré une meilleure performance pour les méthodes présentées par rapport à d’autres modèles crédibilistes ainsi qu’aux récentes forêts aléatoires prudentes lorsque les données sont bruitées. De plus, ces modèles offrent une meilleure robustesse face au surapprentissage lors de l’utilisation de jeux de données qui sont effectivement

10. Les notions de “label dur” et de “label riche” sont définies au chapitre 3.
11. Voir le chapitre “Logiciels et Reproductibilité”.

étiquetés avec incertitude et imprécision. En outre, les modèles proposés sont également capables de prédire des étiquettes riches, une information pouvant être exploitée dans d'autres approches telles que l'apprentissage actif.

Apprentissage Actif

La réunion entre la théorie des fonctions de croyance et l'apprentissage actif est présentée par Hoarau, Martin, J.-C. Dubois et Le Gall 2022. Outre la proposition d'une méthode prenant en compte l'incertitude et l'imprécision des utilisateurs dans les labels, cet article introduit un modèle qui s'appuie sur une nouvelle notion de distance moyenne entre éléments de même classe, lorsque la classe des individus n'est pas connue précisément.

Une publication majeure de cette thèse (voir Hoarau, Lemaire, Martin et al. 2024a) propose deux méthodes d'échantillonnage permettant de répondre à trois problématiques en apprentissage actif. Dans un premier temps, les performances de ces méthodes sont comparées avec un échantillonnage par incertitude classique, et une amélioration nette est démontrée statistiquement sur plusieurs jeux de données. Ensuite, le dilemme d'exploration-exploitation en apprentissage actif est traité, et une solution est proposée. Le dernier point, primordial dans cette thèse, est que les deux méthodes permettent de prendre en compte l'incertitude et l'imprécision déjà présentes dans les labels.

En marge des thématiques abordées dans ce document, Hoarau, Shaker et al. 2024 proposent d'abord une critique des outils actuels permettant de capturer les incertitudes épistémique et aléatoire en apprentissage automatique, puis introduisent un nouveau cadre déterministe permettant de saisir une estimation plus fine des différents types d'incertitude. Ces résultats sont notamment appliqués à l'apprentissage actif.

En parallèle de ces travaux, Hoarau et Lemaire 2024 proposent DEMAU, un outil éducatif, exploratoire, analytique et libre d'accès permettant de visualiser et d'explorer plusieurs types d'incertitudes pour les modèles de classification en apprentissage automatique. Cet outil est plus largement lié à la représentation et à la quantification d'incertitude, mais son application à l'apprentissage automatique et plus généralement au domaine de l'intelligence artificielle le place ici dans un cadre plus pratique que la partie suivante.

Représentation et quantification d'incertitude

La thématique de représentation et de quantification d'incertitude est centrale dans cette thèse. Les travaux présentés par Hoarau, Sale et al. 2024 font une distinction entre

la représentation et la quantification de l'incertitude. La première concerne l'espace de représentation des incertitudes et le cadre théorique utilisé pour modéliser un ou plusieurs types d'incertitudes. La seconde implique les mesures au sein de ces espaces de représentation qui attribuent une valeur numérique à une observation. En plus de recenser plusieurs outils de quantification d'incertitude à travers différents cadres théoriques, une étude quantitative est présentée et constitue le cœur de la contribution. Cette étude non biaisée permet notamment d'aboutir à des résultats tangibles, tels que la corrélation entre les mesures d'incertitudes épistémique et aléatoire à travers différents cadres théoriques (par exemple, les Credal Sets [12] et la théorie des fonctions de croyance), ou encore de fournir des éléments de réponse aux questions récentes sur la décomposition de l'incertitude sous forme de somme. Ces travaux résultent d'une mobilité internationale réalisée lors du doctorat.

L'ensemble de ces travaux constitue la majeure partie du travail produit pour aborder la problématique d'apprentissage actif de données incertaines et imprécises défendue dans cette thèse.

Structure du document

Dans ce document, le chapitre 2 présente la théorie des fonctions de croyance ainsi que l'apprentissage actif. Dans un second temps, une introduction à la labellisation imparfaite est proposée dans le chapitre 3, ainsi qu'une présentation de la méthode utilisée pour obtenir de nouveaux jeux de données labellisés imparfaitement et mis à disposition de la communauté scientifique. Une contribution apportée à la version crédibiliste des K plus proches voisins est ensuite exposée au chapitre 4, et deux modèles de classification sont introduits, les arbres de décision crédibilistes et les forêts aléatoires crédibilistes. Les expériences liées à la rencontre entre la théorie des fonctions de croyance et l'apprentissage actif sont présentées dans le chapitre 5. Le chapitre 6 propose deux nouvelles méthodes d'échantillonnage en apprentissage actif, l'échantillonnage par incertitude crédibiliste et l'échantillonnage par incertitude épistémique crédibiliste. Enfin, le chapitre 7 conclut sur les travaux réalisés et expose plusieurs perspectives de recherche liées à cette thèse.

12. Aucune traduction satisfaisante n'a été trouvée, le terme de *Credal Set* sera donc employé dans ce document.

Chapitre 2

Théorie et concepts

La théorie des fonctions de croyance ainsi que l'apprentissage actif sont présentés dans ce chapitre. La première sert par la suite à modéliser l'incertitude et l'imprécision présentes dans les données manipulées et le second est utilisé pour minimiser le nombre d'observations à labelliser. L'objectif est de travailler avec le moins possible de données labellisées et d'enrichir l'information utilisée lors de la phase de labellisation.

2.1 Théorie des fonctions de croyance

La théorie des fonctions de croyance, aussi appelée théorie de Dempster-Shafer, a été introduite par Dempster 1967 sous la forme de probabilités haute et basse. Ces travaux ont été repris et renommés en *theory of evidence* par Shafer 1976. Cette théorie est une extension de la théorie des probabilités, en s'attachant à la notion de croyance plutôt qu'à la notion de chance. Elle permettra ici de modéliser et de manipuler l'incertitude et l'imprécision présentes dans les labels.

2.1.1 Représentation et modélisation

Le cadre de discernement, noté Ω est l'ensemble des C hypothèses exclusives telles que $\Omega = \{\omega_1, \ldots, \omega_C\}$. En classification, Ω représente toutes les classes d'appartenance possibles pour une observation et C le nombre total de classes.

L'ensemble des parties de Ω, noté 2^Ω représente toutes les disjonctions possibles de ses éléments : $2^\Omega = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_1, \omega_2\}, \ldots, \Omega\}$. Une source S peut alors attribuer une croyance à l'un des éléments de 2^Ω. La fonction $m : 2^\Omega \rightarrow [0, 1]$ qui attribue à tout élément de 2^Ω une croyance (ou masse) est appelée fonction de masse et vérifie toujours la condition de normalisation suivante :

$$\sum_{A \in 2^\Omega} m(A) = 1. \tag{2.1}$$

La somme des masses est toujours égale à 1. Le reste de ce document est placé sous l'hypothèse de monde fermé, où $m(\emptyset) = 0$ est également vérifiée. Les travaux de Philippe Smets et Kennes 1994 permettent de passer outre cette restriction grâce à son hypothèse de monde ouvert, mais ce document s'intéresse exclusivement à des problèmes de classification en monde fermé. Les éléments de masse non nulle sont appelés éléments focaux et l'union des éléments focaux forme le noyau.

Certitude et précision

Par opposition à une labellisation dure, la certitude et la précision de la source peuvent être représentées. Plus la masse sur un élément focal est élevée, plus la certitude de la source est forte. Une réponse sur un singleton ω_q pour tout $q \in \{1, \ldots, C\}$ correspond à une précision, au contraire l'imprécision se définit sur une union. Une source peut donc délivrer une croyance plus ou moins certaine et précise. L'ignorance totale est représentée par $m(\Omega) = 1$.

Exemple 1 : $m(\{\omega_2\}) = 0.2$, $m(\Omega) = 0.8$ et $m(A) = 0$ pour tout autre élément A de 2^Ω. Cette fonction de masse est précise sur ω_2 et très peu certaine avec une masse de 0.8 sur l'ignorance. Le chapitre 3 propose de s'appuyer sur des exemples concrets pour illustrer le propos.

Exemple 2 : $m(\{\omega_1, \omega_2\}) = 0.9$, $m(\Omega) = 0.1$ et $m(A) = 0$, $\forall A \in 2^\Omega \backslash \{\Omega, \{\omega_1, \omega_2\}\}$. Cette fonction de masse est imprécise sur ω_1 et ω_2 et quasi certaine.

Fonction de masse catégorique

Une fonction de masse est dite catégorique si et seulement si elle ne possède qu'un seul élément focal :

$$\begin{cases} m(A) = 1\ , & \exists A \in 2^\Omega, \\ m(B) = 0\ , & \forall B \in 2^\Omega, B \neq A. \end{cases} \tag{2.2}$$

Dans ce cas, toute la croyance lui est alors attribuée et la réponse est totalement certaine. Si l'élément focal unique est un singleton, alors la réponse est également précise, sinon, la croyance est donnée sur une union et la réponse est imprécise.

Fonction de masse à support simple

Une fonction de masse à support simple possède deux éléments focaux, dont l'un est un élément $A \in 2^\Omega \backslash \Omega$ et l'autre est toujours le cadre de discernement Ω. Tous les autres éléments de 2^Ω, différents de A et Ω, se voient attribuer une masse nulle. Elle se note :

$$\begin{cases} m(A) = 1 - w_0, \\ m(\Omega) = w_0, \\ m(B) = 0\ , \quad \forall B \in 2^\Omega, B \notin \{\Omega, A\}, \end{cases} \tag{2.3}$$

avec $w_0 \in [0, 1]$. Cette fonction de masse à support simple est également notée A^{w_0}.

Fonction de masse dogmatique

Une fonction de masse est dogmatique lorsque la masse sur le cadre de discernement est nulle : $m(\Omega) = 0$.

Fonction de masse consonante

Une fonction de masse consonante possède tous ses éléments focaux emboités. Par exemple, une fonction de masse ayant pour éléments focaux $\{\omega_1\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_2, \omega_3\}$ et Ω est dite consonante, car elle respecte $\{\omega_1\} \subset \{\omega_1, \omega_2\} \subset \{\omega_1, \omega_2, \omega_3\} \subset \Omega$.

Crédibilité et plausibilité

Les notions de probabilités haute et basse, appelées respectivement plausibilité et crédibilité, permettent d'encadrer la croyance d'une source. La fonction de crédibilité notée Cr permet de représenter la croyance minimale de la source et se calcule de la manière suivante :

$$Cr(A) = \sum_{B \subseteq A} m(B), \quad \forall A \in 2^\Omega. \tag{2.4}$$

Il s'agit de la somme des éléments d'informations qui impliquent A et elle représente la croyance totale en A.
La fonction de plausibilité Pl est la croyance maximale pouvant être accordée à A, elle est définie par :

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B), \quad \forall A \in 2^\Omega. \tag{2.5}$$

Il s'agit de la somme des éléments d'informations non contradictoires avec A. Il est possible de passer de la plausibilité d'un élément A à la crédibilité de son complémentaire $\bar{A}$ avec la formule suivante :

$$Pl(A) = 1 - Cr(\bar{A}). \tag{2.6}$$

Affaiblissement

Une source S peut n'être que partiellement fiable, un coefficient d'affaiblissement est alors introduit et permet de redistribuer une partie de la masse vers l'ignorance. Ce coefficient est noté α et prend ses valeurs dans $[0, 1]$. Lorsque α vaut 1 la source est totalement fiable, et plus ce coefficient s'approche de 0, plus la fiabilité décroît. La fonction de masse affaiblie, notée m^α se calcule avec la formule suivante :

$$\begin{cases} m^\alpha(A) = \alpha m(A) \ , \quad \forall A \in 2^\Omega, A \neq \Omega, \\ m^\alpha(\Omega) = 1 - \alpha(1 - m(\Omega)). \end{cases} \tag{2.7}$$

[1]

Distance entre fonctions de masse

Une notion de distance entre deux fonctions de masse sur le même cadre de discernement peut être calculée par la distance euclidienne entre les deux vecteurs $\mathbf{m_1}$ et $\mathbf{m_2}$ représentant respectivement les fonctions de masse m_1 et m_2 (même ordre d'énumération). La pleine représentation des fonctions de masse n'est alors pas exploitée puisque la distance euclidienne ne prend pas en compte la similarité entre les cadres de réponse.

Exemple : Soit trois fonctions de masse m_1, m_2 et m_3 définies par :

m_1 **:** $m_1(\{chien\}) = 1$, *"C'est un chien"*.

m_2 **:** $m_2(\{chat\}) = 1$, *"C'est un chat"*.

m_3 **:** $m_3(\{chat,\ oiseau\}) = 1$, *"C'est un chat ou un oiseau"*.

La distance euclidienne entre m_1 et m_3 est la même que la distance entre m_2 et m_3 puisque les éléments $\{chien\}$, $\{chat\}$ et $\{chat,\ oiseau\}$ sont différents. Il n'y a pas une proximité plus grande entre $\{chat\}$ et $\{chat,\ oiseau\}$ qu'entre $\{chien\}$ et $\{chat,\ oiseau\}$, malgré l'inclusion de la classe *chat*.

1. Une autre notation présente dans la littérature est la suivante : $m^\alpha(\Omega) = 1 - \alpha + \alpha m(\Omega)$.

Une autre notion de distance entre deux fonctions de masse est introduite par Jousselme, Grenier et al. 2001. Lorsque la distance de Jousselme entre deux fonctions de masse vaut 0, elles sont identiques. À l'inverse, lorsque la distance vaut 1, les fonctions de masse sont les plus éloignées possibles. Cette distance s'appuie sur l'indice de Jaccard, permettant d'étudier la similarité entre des objets d'attributs binaires. Pour deux fonctions de masse m_1 et m_2, représentées par $\mathbf{m_1}$ et $\mathbf{m_2}$, la distance se calcule de la manière suivante :

$$d_J(m_1, m_2) = \sqrt{\frac{1}{2}(\mathbf{m_1} - \mathbf{m_2})^T \underline{D}(\mathbf{m_1} - \mathbf{m_2})}, \tag{2.8}$$

avec $\underline{D}$ la matrice d'éléments D tels que :

$$D(A, B) = \frac{|A \cap B|}{|A \cup B|}, \qquad A, B \in 2^\Omega. \tag{2.9}$$

Dans l'exemple précédent, la distance de Jousselme entre m_1 et m_3 est plus grande que la distance entre m_2 et m_3, avec pour cause l'inclusion de la classe "*chat*" dans "*chat*, *oiseau*".

2.1.2 Combinaison d'information

Lorsque différentes sources s'expriment, il est possible de combiner deux ou plusieurs fonctions de masse définies sur le même cadre de discernement. Une multitude de méthodes existent et certaines sont présentées ici, ayant chacune des avantages et inconvénients en fonction du contexte de fusion.

Règle de combinaison conjonctive

L'opérateur $\oplus$ est appelé règle de combinaison conjonctive et peut être utilisé pour combiner deux fonctions de masse issues de sources fiables et cognitivement indépendantes :

$$m_{1,2}(A) = (m_1 \oplus m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \tag{2.10}$$

Cette opération peut être généralisée à N fonctions de masse avec la formule suivante :

$$\bigoplus_{i=1}^{N} m_i(A) = \sum_{B_1 \cap \ldots \cap B_N = A} \prod_{j=1}^{N} m_j(B_j). \tag{2.11}$$

La nouvelle fonction de masse m peut être normalisée, et doit l'être sous l'hypothèse du monde fermé. La règle de combinaison conjonctive normalisée est donc :

$$\begin{cases} m(A) = \dfrac{1}{1-\kappa} \displaystyle\sum_{B_1 \cap \ldots \cap B_N = A} \prod_{j=1}^{N} m_j(B_j), & \text{si } A \neq \emptyset, \\ m(\emptyset) = 0, \end{cases} \tag{2.12}$$

avec κ l'inconsistance de la fusion :

$$\kappa = \sum_{B_1 \cap \cdots \cap B_N = \emptyset} \prod_{j=1}^{N} m_j(B_j). \tag{2.13}$$

Règle de combinaison prudente

La règle de combinaison prudente $\textcircled{\wedge}$, introduite par Denœux 2006, permet de combiner deux fonctions de masse non dogmatiques dépendantes. Elle se note $m_1 \textcircled{\wedge} m_2 = \bigoplus_{A \in 2^\Omega} A^{w_1(A) \wedge w_2(A)}$ avec $\wedge$ l'opérateur de minimum.

Moyenne des fonctions de masse

Une simple moyenne des fonctions de masse peut aussi être préférée, permettant de lever la plupart des contraintes sur les fonctions de masse initiales et sur les sources :

$$m(A) = \frac{1}{N} \sum_{i=1}^{N} m_i(A), \quad \forall A \in 2^\Omega. \tag{2.14}$$

Autres règles

D'autres règles (*cf.* D. Dubois et Prade 1988 ; Lefevre, Colot et al. 2002 ; Martin 2019 ; Yager 1987) peuvent permettre de combiner plusieurs fonctions de masse, par exemple la règle de combinaison disjonctive, moins restrictive que la règle de combinaison conjonctive, qui élargit les éléments focaux et qui nécessite que toutes les fonctions de masse soient cognitivement indépendantes et qu'au moins une d'entre elles soit fiable.

2.1.3 Prise de décision

Il est possible et parfois nécessaire de prendre une décision sur le cadre de discernement Ω. Choisir le maximum de plausibilité sur les singletons de 2^Ω (les éléments de Ω) permet de choisir l'hypothèse qui maximise la croyance maximale accordée. Le maximum

de crédibilité peut également être choisi, retenant l'hypothèse avec la plus grande croyance minimale. Une prise de décision peut également être faite sur un élément de 2^Ω en fonction, par exemple, d'une mesure de distance (voir Essaid, Martin et al. 2014). Une liste détaillée de mesures appartenant au monde crédal sont présentées par Denœux 2019.

La probabilité pignistique, introduite par Philippe Smets et Kennes 1994 permet d'utiliser les masses sur 2^Ω pour orienter un choix sur Ω. Cette formule approche la crédibilité et la plausibilité et se note $BetP$. Elle se calcule en monde fermé avec $\omega \in \Omega$ comme suit :

$$BetP(\omega) = \sum_{A \in 2^\Omega,\ \omega \in A} \frac{m(A)}{|A|}. \tag{2.15}$$

Cette transformée pignistique sera fréquemment utilisée dans ce document. Il s'agit d'un choix critiquable, notamment due à la comparaison difficile entre des modèles utilisant des labels durs, définis sur Ω, et des modèles utilisant des labels riches, modélisés par la théorie des fonctions de croyance et définis sur 2^Ω.

Plusieurs modèles utilisant la théorie des fonctions de croyance sont présentés dans les chapitres suivants. Ces modèles crédibilistes sont capables de représenter des croyances, mais pour être compatibles avec les standards d'évaluation et de comparaison de modèles plus classiques [2] nous avons décidé de réduire l'information en sortie de ces modèles à un cadre probabiliste. La probabilité pignistique est alors utilisée pour homogénéiser la sortie d'un classifieur crédibiliste avec celle d'un classifieur probabiliste.

Exemple : Soit un cadre de discernement $\Omega = \{\omega_1, \omega_2, \omega_3\}$ et une fonction de masse m définie par :

m **:** $m(\{\omega_1\}) = 0.2$, $m(\{\omega_1, \omega_2\}) = 0.5$, $m(\{\omega_1, \omega_2, \omega_3\}) = 0.3$.

La probabilité pignistique associée à m sur le cadre Ω est donnée par :

— $BetP(\omega_1) = 0.55$,

— $BetP(\omega_2) = 0.35$,

— $BetP(\omega_3) = 0.1$.

La somme des probabilités pignistiques sur Ω vaut 1 et les propriétés probabilistes sont vérifiées (*cf.* P. Smets 1990).

2. Notamment l'*accuracy* des prédictions correctes du modèle (le terme *exactitude* sera utilisé), l'un des outils les plus utilisés.

Aparté sur la transformée pignistique

Une fonction de masse peut être représentée dans un cadre probabiliste sous la contrainte du principe de raison insuffisante, c'est-à-dire qu'en l'absence d'information, la croyance est distribuée uniformément sur chaque sous-élément. C'est un principe couramment utilisé et critiqué dans la représentation de l'ignorance en statistiques bayésiennes. En classification, il y a donc une différence d'interprétation possible entre la sortie d'un modèle probabiliste qui manipule des chances et un modèle crédibiliste qui manipule des croyances.

2.1.4 Incertitude de Klir

Georges Klir fait état des mesures d'incertitude en théorie de l'information (*cf.* Klir et Wierman 1998). Il divise l'incertitude en trois types ; *le flou* (ou vague), *la non-spécificité* (ou l'imprécision) et *la discorde*. La première ne sera pas traitée ici puisqu'elle n'apporte pas d'information supplémentaire exploitable sur les labels. Il peut être noté que cette incertitude n'est pas aussi bien représentée dans la théorie des fonctions de croyance que dans d'autres théories de l'incertain, comme les sous-ensembles flous.

La mesure de non-spécificité N au sein de la théorie des fonctions de croyance permet de mesurer la quantité d'information imprécise d'une fonction de masse m :

$$N(m) = \sum_{A \subseteq \Omega} m(A) \log_2(|A|), \tag{2.16}$$

avec $|A|$ la cardinalité de l'ensemble A.

La notion de désordre est souvent étudiée en probabilité au travers de l'entropie de Shannon 1948, c'est l'un des fondements de la théorie de l'information. Elle mesure l'incertitude moyenne associée à la prédiction d'une expérience aléatoire. La notion de discorde D est une extension de l'entropie de Shannon à la théorie des fonctions de croyance, elle mesure la discorde parmi les éléments d'une fonction de masse et se définit par :

$$D(m) = -\sum_{A \subseteq \Omega} m(A) \log_2(BetP(A)). \tag{2.17}$$

En utilisant la probabilité pignistique $BetP$, le principe de raison insuffisante est implicite-

ment respecté[3]. Cette contrainte peut être levée en remplaçant la probabilité pignistique par la plausibilité Pl, ce qui donne *la dissonance*, ou par la crédibilité Cr, donnant *la confusion*.

L'incertitude globale, est définie par Klir et Wierman 1998 comme devant capturer à la fois la discorde et la non-spécificité, ces deux incertitudes qui coexistent au sein de la théorie des fonctions de croyance. L'incertitude $\mathcal{U}$ de Klir est donc la somme de la discorde et de la non-spécificité :

$$\mathcal{U}(m) = D(m) + N(m). \tag{2.18}$$

Une autre forme de cette incertitude sera présentée dans ce document, avec plus ou moins de non-spécificité ou de discorde en fonction de l'orientation souhaitée (*cf.* Denoeux et Bjanger 2000). Il est aussi à noter que cette formule ne respecte pas la sous-additivité, un point qui remet partiellement en cause son utilisation dans les années 90. Une analyse claire et complète de différentes méthodes de quantification de l'incertitude pour des probabilités imprécises en apprentissage automatique est proposée par Hüllermeier, Destercke et al. 2022. Le chapitre 5 traitera plus en détail de ces notions.

2.1.5 Conflit

Si l'information conflictuelle entre les éléments d'une même fonction de masse a été vue précédemment sous la forme de *discorde*, il est possible de mesurer le conflit entre plusieurs fonctions de masse. Ce conflit est défini par Martin 2019 et l'une des mesures proposées utilise un degré d'inclusion et une distance.

Une *inclusion stricte* est alors définie comme suit : *On dit qu'une fonction de masse m_i est incluse dans m_j si tous les éléments focaux de m_i sont inclus dans chaque élément focal de m_j.*

Une *inclusion légère* moins restrictive est définie par : *On dit qu'une fonction de masse m_i est incluse dans m_j si tous les éléments focaux de m_i sont inclus dans au moins un élément focal de m_j.*

On note l'inclusion $m_i \subseteq m_j$. Le degré d'inclusion stricte $\delta_s^{i \subseteq j}(m_i, m_j)$ de m_i dans m_j est donné par :

$$\delta_s^{i \subseteq j}(m_i, m_j) = \frac{1}{|\mathcal{F}_i||\mathcal{F}_j|} \sum_{A \in \mathcal{F}_i} \sum_{B \in \mathcal{F}_j} Inc(A, B), \tag{2.19}$$

avec $\mathcal{F}_i$ et $\mathcal{F}_j$ respectivement les sous-ensembles des éléments focaux de m_1 et m_2, l'index

3. Voir l'aparté précédent pour le *principe de raison insuffisante*.

d'inclusion $Inc(A, B)$ vaut 1 quand $A \subseteq B$ et 0 sinon. Pour une formule moins stricte, le degré d'inclusion légère $\delta_l^{i \subseteq j}(m_i, m_j)$ de m_i dans m_j est donné par :

$$\delta_l^{i \subseteq j}(m_i, m_j) = \frac{1}{|\mathcal{F}_i|} \sum_{A \in \mathcal{F}_i} \max_{B \in \mathcal{F}_j}(Inc(A, B)). \tag{2.20}$$

Soit $\delta(m_i, m_j)$ le degré d'inclusion de m_i et m_j défini par :

$$\delta(m_i, m_j) = \max(\delta^{i \subseteq j}(m_i, m_j), \delta^{j \subseteq i}(m_j, m_i)). \tag{2.21}$$

Ce degré donne la proportion maximale d'éléments focaux d'une fonction de masse inclus dans l'autre. Une mesure de conflit $\mathcal{C}$ entre deux fonctions de masse est alors introduite :

$$\mathcal{C}(m_i, m_j) = (1 - \delta(m_i, m_j)) d_J(m_i, m_j), \tag{2.22}$$

avec $d_J(m_j, m_i)$ la distance de Jousselme donnée par l'équation (2.8), entre m_i et m_j. Pour un nombre de sources supérieur à 2, le conflit est la moyenne des conflits deux à deux.

La théorie des fonctions de croyance sera utilisée dans ce document pour représenter l'incertitude et l'imprécision relatives au label d'une observation. Elle sera également utilisée pour la construction de modèles d'apprentissage dits crédibilistes, fournissant une fonction de masse et non plus un label dur pour une observation à étiqueter. L'objectif ici est de comprendre si en ajoutant de l'information, même imparfaite, durant la phase de labellisation, de meilleurs résultats peuvent être obtenus lors de la classification.

2.2 Apprentissage actif

2.2.1 Apprentissage supervisé

En apprentissage automatique, l'apprentissage supervisé consiste à apprendre à partir d'exemples annotés et à restituer une prédiction. Les exemples annotés forment la base d'apprentissage. Soit $\mathcal{L} = \{(x_n, y_n) | 1 \leq n \leq N\}$ la base de données d'apprentissage de N observations avec $x_n \in X$ et $y_n \in Y$ les réalisations respectives des variables aléatoires de X et Y. L'espace X est appelé espace d'entrées, il s'agit des variables de prédiction, et Y est l'espace de sortie, ce sont les labels (ou étiquettes) des observations.

L'objectif d'un modèle d'apprentissage supervisé est de généraliser l'apprentissage effectué sur des données labellisées par des experts, autrement dit d'apprendre à partir d'un jeu d'entraînement labellisé.

2.2.2 Apprentissage actif

L'apprentissage actif, synthétisé plus exhaustivement par Bondu et Lemaire 2008, par Settles 2009 et par Aggarwal, Kong et al. 2014, est une branche de l'apprentissage automatique où l'apprenant peut choisir les observations qui doivent être labellisées (voir figure 2.1). Il permet notamment de ne travailler qu'avec une fraction labellisée du jeu de données, s'inscrivant ainsi dans une réponse à des problématiques de coûts liées à la labellisation. C'est une approche utilisée pour la réduction des coûts (*cf.* Hacohen, Dekel et al. 2022), mais aussi dans d'autres domaines tels que la détection d'anomalies, comme Abe, Zadrozny et al. 2006 et Martens, Perini et al. 2023.

En apprentissage actif, les observations sont appelées des *instances*. La requête permettant d'obtenir la classe d'une instance est faite à l'*oracle* qui délivre les labels. L'ensemble $\mathcal{U} = \{x_\nu | 1 \leq \nu \leq N - |\mathcal{L}|\}$ désigne les instances non labellisées et $\mathcal{L} = \{(x_\mu, y_\mu) | 1 \leq \mu \leq N - |\mathcal{U}|\}$ les instances labellisées. La difficulté réside donc dans le choix des données à labelliser, afin de sélectionner les instances qui apporteront le meilleur gain de performance au modèle. Cette étape est appelée *échantillonnage*.

2.2.3 Stratégie active d'apprentissage

Deux scénarios sont distingués par Willett, Nowak et al. 2005, l'apprentissage adaptatif et l'apprentissage sélectif.

Apprentissage adaptatif

En échantillonnage adaptatif (*cf.* Singh, Nowak et al. 2006) des vecteurs descripteurs sont utilisés pour être étiquetés à la place des instances, ils sont ainsi labellisés par l'oracle. Ce scénario peut induire la labellisation de vecteurs qui n'ont pas de sens réel, ou qui ne sont pas représentatifs du problème, ce qui rend l'interprétation de ce scénario plus complexe.

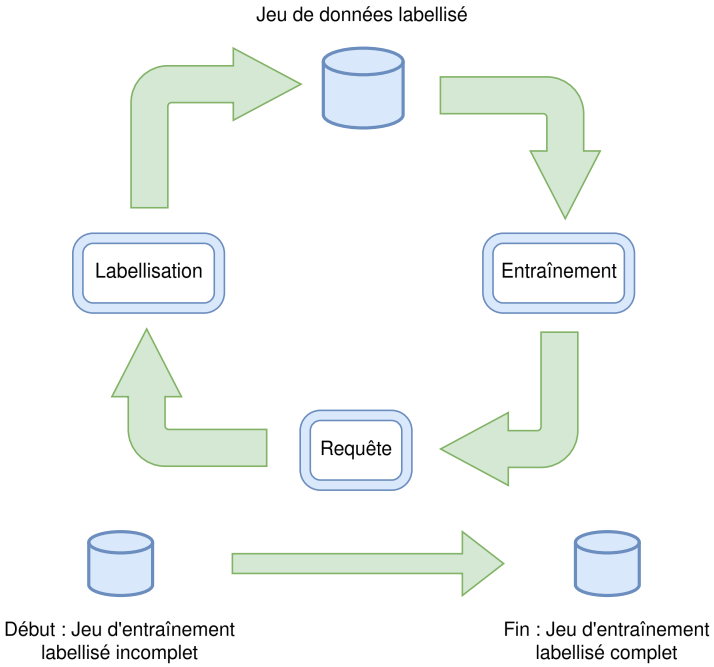


FIGURE 2.1 – Schéma de l'Apprentissage Actif.

Apprentissage sélectif

En échantillonnage sélectif, les données labellisées par l'oracle sont les instances elles-mêmes. L'information labellisée par l'oracle sera donc une des données d'entraînement, ce qui permet de ne sélectionner que des labels sur des observations qui existent réellement. C'est l'échantillonnage qui sera retenu dans la suite de cette étude. Cependant, il existe plusieurs méthodes permettant de réaliser cet échantillonnage, elles sont présentées dans la partie suivante.

2.2.4 Méthodes d'échantillonnage

Lors de l'apprentissage actif, il convient de sélectionner la meilleure instance à labelliser. Ce choix est au cœur du processus et il existe plusieurs procédures possibles, les plus répandues sont présentées ci-dessous.

Échantillonnage aléatoire

L'échantillonnage aléatoire est le plus simple à aborder ; le modèle choisit aléatoirement les instances qui vont être labellisées pour la phase d'entraînement. Une instance est tirée

aléatoirement dans l'ensemble U des instances non labellisées, son label est requêté à l'oracle et l'instance ainsi labellisée est ajoutée à l'ensemble $\mathcal{L}$ des instances labellisées.

Échantillonnage par incertitude

L'échantillonnage par incertitude (*cf.* Lewis et Gale 1994) s'appuie sur la capacité du modèle d'apprentissage à délivrer avec sa réponse une fiabilité. Une prédiction sous forme de probabilité associée à une classe pour une instance peut par exemple être utilisée. Cet échantillonnage vise donc à choisir l'instance pour laquelle le modèle délivre, lors de la classification, la certitude la plus faible. Cette instance de U se voit alors attribuer un label par une requête à l'oracle, puis est ajoutée à l'ensemble $\mathcal{L}$ des instances labellisées.

L'instance x^* à labelliser est alors, parmi les meilleures prédictions, celle pour laquelle le modèle est le moins confiant :

$$x^* = \underset{x \in U}{\operatorname{argmin}}\ P(y^*|x), \tag{2.23}$$

avec $y^* = \underset{y \in \Omega}{\operatorname{argmax}}\ P(y|x)$ la classe de Ω la plus probable pour une instance x.

Une autre approche d'échantillonnage par incertitude vise à utiliser l'entropie pour sélectionner la meilleure instance à labelliser. En maximisant la formule de l'entropie proposée par Shannon 1948 on a :

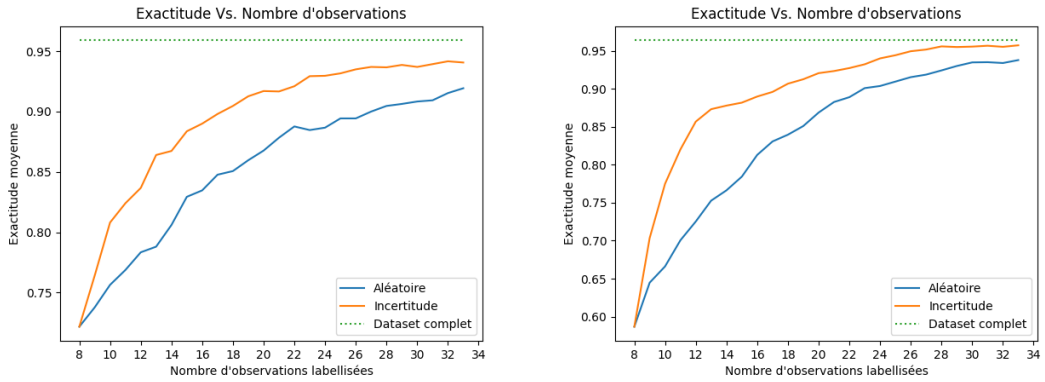
$$x^* = \underset{x \in U}{\operatorname{argmax}}\ -\sum_{i=1}^{C} P(y_i|x) log_2 P(y_i|x) \tag{2.24}$$

avec y_i couvrant les possibles labels.

Échantillonnage par comité de modèles

L'échantillonnage par comité de modèles (*cf.* Seung, Opper et al. 1992) consiste à utiliser plusieurs modèles entraînés en parallèle sur les mêmes données. Le désaccord entre les modèles est ensuite mesuré et l'instance maximisant le désaccord est choisie pour être labellisée. L'entropie peut encore une fois être utilisée pour mesurer le désaccord, et ainsi sélectionner la meilleure instance. L'objectif et la difficulté ici, sont d'obtenir un comité de modèles qui représente au mieux l'espace des versions.

Il existe d'autres méthodes d'échantillonnage comme les arbres sur l'espace des versions, décrits par S. Tong et Koller 2002, ou encore les échantillonnages par changement



(a) K-NN - Iris (b) K-NN - Wine

FIGURE 2.2 – Évolution des performances du modèle des K plus proches voisins en fonction du nombre d'observations labellisées, sur le jeu Iris (2.2a) et sur le jeu Wine (2.2b)

attendu (voir D. A. Cohn, Ghahramani et al. 1996).

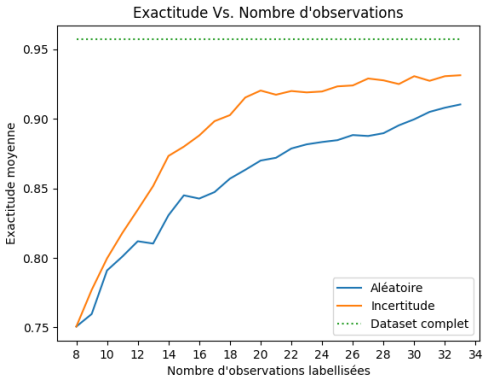
2.2.5 Illustration sur données réelles

Le jeu de données Iris de Fisher est composé de trois différentes espèces d'Iris, qui sont des plantes vivaces de la famille des Iridacées. Ce jeu de données est composé de 150 observations équitablement divisées en trois classes *Setosa*, *Versicolor* et *Virginica*, selon 4 variables caractéristiques. Entraîner un modèle de classification sur ce jeu de données est un exercice courant en science des données. Cependant, si l'on s'intéresse à sa construction, il a fallu labelliser les 150 observations [4] en leur attribuant une classe parmi $\Omega = \{Setosa,\ Versicolor,\ Virginica\}$. L'intérêt de l'apprentissage actif est de réduire ce nombre d'observations à labelliser en essayant d'avoir des performances peu impactées. Le même exemple est aussi illustré sur le jeu de données Wine, composé de 178 observations réparties sur 3 classes, selon 13 variables caractéristiques.

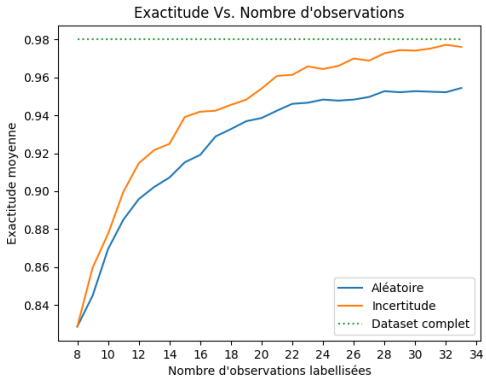
Les figures 2.2, 2.3 et 2.4 représentent trois illustrations d'apprentissage actif sur les jeux de données Iris et Wine pour différents modèles de classification [5]. L'échantillonnage aléatoire est souvent utilisé comme repère, si l'on ne prend pas en compte le modèle et que l'on se contente de tirer aléatoirement les observations, on obtient les performances de cette méthode, représentées par la courbe bleue. L'échantillonnage par incertitude

4. Il s'agit en réalité de 120 observations à labelliser, puisque 80% du jeu de données est ici utilisé pour entraîner le modèle.

5. Ces modèles sont utilisés ici à titre indicatif et sont introduits au chapitre 4.

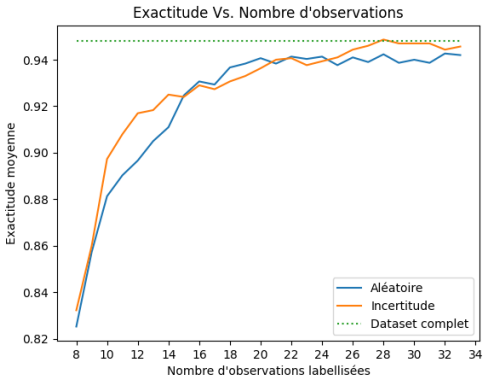


(a) Régression logistique - Iris

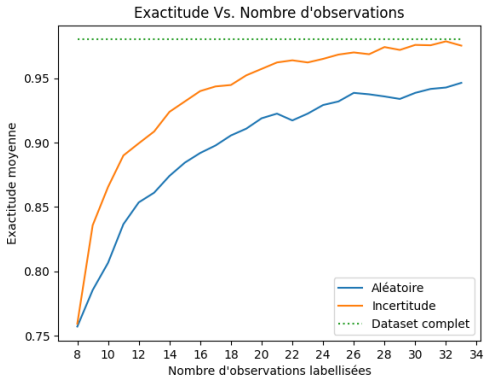


(b) Régression logistique - Wine

FIGURE 2.3 – Évolution des performances d'une régression logistique en fonction du nombre d'observations labellisées, sur le jeu Iris (2.3a) et sur le jeu Wine (2.3b)



(a) Forêt aléatoire - Iris



(b) Forêt aléatoire - Wine

FIGURE 2.4 – Évolution des performances d'une forêt aléatoire en fonction du nombre d'observations labellisées, sur le jeu Iris (2.4a) et sur le jeu Wine (2.4b)

est représenté en orange et ses performances sont presque systématiquement meilleures que celles de l'échantillonnage aléatoire. Le modèle peut alors choisir les observations à labelliser pour lesquelles il est le plus incertain. Les performances asymptotiques du modèle sur un jeu de données complet sont représentées par des pointillés verts. [6]

On voit qu'avec moins de 35 observations (sur 150 pour Iris et sur 178 pour Wine) les performances des modèles sont souvent proches des performances sur le jeu complet tout en permettant de diviser au minimum par 4 le nombre d'observations labellisées et donc de réduire les coûts liés à la labellisation.

2.2.6 Dilemme d'exploration-exploitation

Au cours du processus d'apprentissage actif, le choix des exemples à labelliser peut être considéré comme un dilemme entre exploration et exploitation dans l'espace des données d'entrée X. L'exploration consiste à labelliser des instances dans une zone non échantillonnée, X tend alors vers un échantillonnage uniforme. La sélection d'une instance dans une zone échantillonnée de l'espace sur X correspond à l'exploitation des données, dans ce cas l'apprentissage actif se concentre sur une zone déjà peuplée avec des instances labellisées, et affine localement les prédictions du modèle.

Le dilemme d'exploration-exploitation en apprentissage actif correspond alors à la possibilité de choisir l'une ou l'autre, ou même un compromis comme le suggère Bondu, Lemaire et Boullé 2010. Ce dilemme est mis en évidence sur la figure 2.5. Pour cette tâche de classification en deux dimensions, la figure de gauche illustre le cas où toutes les observations sont labellisées, tandis que la figure de droite montre la situation après seulement quelques itérations du processus d'échantillonnage, où seules certaines instances sont labellisées. Si la stratégie d'échantillonnage n'effectue que de l'exploitation [7], il est évident qu'aucune information supplémentaire ne sera requêtée sur les instances situées dans le coin en haut à gauche. Il en résultera une baisse en performances puisque le modèle de classification devra modifier un grand nombre de paramètres lorsqu'il découvrira les exemples rouges dans la zone inexploitée. Inversement, si la stratégie d'échantillonnage n'effectue que de l'exploration pure et simple, le temps nécessaire pour affiner la prédiction des deux motifs sera très long. Par conséquent, une bonne stratégie doit tenir compte du compromis exploration-exploitation.

6. Les performances sont évaluées selon l'exactitude moyenne des modèles sur 100 expériences (*i.e.* la moyenne des taux de bonnes classifications sur 100 tirages d'un jeu de test).

7. Le modèle de classification est représenté par la ligne bleue.

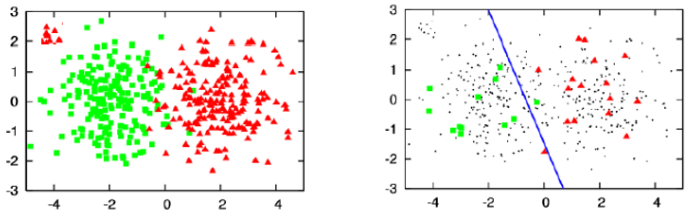


FIGURE 2.5 – Dilemme d'exploration-exploitation.

2.3 Conclusion du chapitre

Deux outils utilisés dans ce document ont été présentés. D'abord la théorie des fonctions de croyance, dont les probabilités sont un cas particulier, qui permet par la suite de modéliser l'incertitude et l'imprécision dans les données manipulées. La représentation et modélisation de croyances a été explicitée, plusieurs méthodes pour combiner l'information provenant de plusieurs sources ont été présentées ainsi que des outils liés à la prise de décision. L'apprentissage actif, branche de l'apprentissage automatique, a également été exposé. Il permet de travailler dans un contexte où le nombre de labellisations peut être minimisé, en utilisant uniquement les observations les plus pertinentes. Plusieurs méthodes d'échantillonnage ont été abordées et illustrées, dont l'échantillonnage par incertitude, pour lequel plusieurs travaux de recherche ont été entrepris durant cette thèse.

Chapitre 3

Labellisation imparfaite

Labelliser imparfaitement n'est pas synonyme de perte d'information ou de fiabilité. Au contraire, cette imperfection se traduit par la capacité à retranscrire une incertitude ou une imprécision lors de la labellisation. Laisser la possibilité à une source, ou un utilisateur, d'exprimer une croyance et non une affirmation peut permettre de mieux représenter l'information. L'objectif est de travailler non plus avec des labels durs décrivant une observation, mais avec des labels plus riches, pouvant représenter plusieurs degrés d'ignorance. Ainsi, un utilisateur qui se trompe lors d'une labellisation et qui est capable de représenter son ignorance sur le sujet peut induire une erreur plus faible. L'approche dans un premier temps orientée pour de la classification, prend en compte des labels représentés de manière qualitative. Certaines approches (*cf.* Fiche, Martin et al. 2010 ; Philippe Smets 2005 ; Strat 1984) permettent une modélisation d'informations quantitatives, d'autres travaux ont été réalisés en proposant une formalisation pour des bases de données crédibilistes (voir Lee 1992 ; Samet, Lefèvre et al. 2014). Ce chapitre introduit donc la différence entre données parfaitement et imparfaitement labellisées afin de présenter les jeux de données utilisés dans le reste de cette thèse ainsi que leur processus de génération. Toutes ces ressources ont été formalisées et mises à disposition pour la communauté scientifique.

Publications

— Constance Thierry, Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2022), « Real bird dataset with imprecise and uncertain values », in : *Belief Functions : Theory and Applications*, p. 275-285

— Arthur Hoarau, Constance Thierry, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2023), « Datasets with Rich Labels for Machine Learning », in : *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, p. 1-6

3.1 Labels riches

Soit un ensemble de descripteurs $\mathcal{X} = \{x^n = (x_1^n, \ldots, x_P^n) | n = 1, \ldots, N\}$ d'éléments à P dimensions pour N observations. La variable statistique x_i^n est le i-ème élément décrivant l'observation n. Le processus de labellisation vise à attribuer un label à chacune des observations parmi $\Omega = \{\omega_1, \ldots, \omega_C\}$ l'ensemble des C classes possibles pour un problème de classification. On distingue deux cas : la labellisation dure où chaque observation est représentée par une classe [1] et la labellisation imparfaite où chaque observation est représentée par label riche, ici modélisé par la théorie des fonctions de croyance.

Variable statistique

Une variable statistique est une caractéristique attribuée à une observation, elle peut être :
Qualitative, si l'observation est décrite par un trait ou une catégorie ; elle est dite ordinale s'il existe un ordre entre toutes les valeurs possibles (petit, moyen, grand) et nominale si ces valeurs ne peuvent pas être ordonnées (ensoleillé, pluvieux, nuageux)
Quantitative, si la description est numérique, mesurable. La variable peut-être quantitative discrète (à valeur entière) ou quantitative continue (à valeurs dans $\mathbb{R}$).

Labellisation dure

Lors d'une labellisation dure, une observation n se voit attribuer une variable qualitative y^n à valeurs dans Ω. L'information contenue dans cette variable peut être fausse, mais la labellisation peut être dite parfaite, car elle ne contient ni incertitude ni imprécision. Lorsque la variable à expliquer est quantitative, on parle de régression. Dans ce document, seul le problème de classification, avec des variables qualitatives, sera traité. Un exemple de jeu de données parfaitement labellisé est présenté dans le tableau 3.1.

Labellisation imparfaite

La théorie des fonctions de croyance, présentée dans la partie 2.1, sera utilisée pour modéliser des labels plus riches. Ici, la variable attribuée à l'observation est remplacée par

1. Label avec une probabilité de 1 sur la classe de l'observation pour un modèle probabiliste.

n	x_1	x_2	x_3	y
1	11 $°C$	53 %	10 km/h	Ensoleillé
2	8 $°C$	86 %	26 km/h	Pluvieux
3	16 $°C$	55 %	23 km/h	Nuageux

TABLEAU 3.1 – Jeu de données contenant 3 observations parfaitement labellisées, x_1 représente la température en degrés Celsius, x_2 l'humidité en pourcentage, x_3 la vitesse du vent en kilomètres par heure et y est le label dur représentant la condition météorologique parmi $\Omega = \{Ensoleillé,\ Pluvieux,\ Nuageux\}$.

une fonction de masse m_n à valeurs dans 2^Ω. La labellisation est dite imparfaite, mais l'information contenue dans cette fonction de masse peut être une meilleure représentation de la classe réelle de l'observation n, car elle tient compte de l'incertitude et de l'imprécision de la source. Un exemple de jeu de données imparfaitement labellisé est représenté dans le tableau 3.2. Les observations sont décrites avec les mêmes variables x_1, x_2 et x_3, mais contrairement au tableau 3.1, les labels y sont exprimés par des masses et non plus par des variables qualitatives.

Un autre cas, non traité ici, est celui de jeux de données caractérisés entièrement de manière incertaine et imprécise avec la théorie des fonctions de croyance. L'élément x_i^n n'est plus une variable statistique, mais une fonction de masse décrivant l'observation n. Lorsque l'information est qualitative, elle est représentée par une fonction de masse. Cependant, quand l'information est quantitative, il convient de travailler différemment, par exemple avec des fonctions de masse continues. Certains travaux de Strat 1984, de Philippe Smets 2005 et de Fiche, Martin et al. 2010 ont été réalisés en ce sens.

3.2 Campagnes de création de jeux de données

Dans l'objectif de travailler avec des données imparfaitement labellisées, plusieurs campagnes de production participative ont été réalisées. L'interface permet aux utilisateurs de labelliser des images avec incertitude et imprécision. Les observations ainsi récoltées se voient attribuer un label riche. Il n'existe pas à notre connaissance de jeu de données pour l'apprentissage automatique labellisé de manière incertaine et imprécise par des contributeurs qui ont eu la possibilité de représenter une imperfection. La plupart du temps, des jeux de données avec des labels durs sont bruités, ou comme présenté par Schmarje, Brünger et al. 2021, des labels flous sont extraits du jeu de données original. D'autres approches font directement référence à des labels flous (voir Schmarje, Zelenka et al. 2019),

n	x_1	x_2	x_3	y
1	11 $°C$	53 %	10 km/h	$m_1(\textit{Ensoleillé}) = 0.7$, $m_1(\textit{Ensoleillé, Nuageux}) = 0.2$, $m_1(\Omega) = 0.1$
2	8 $°C$	86 %	26 km/h	$m_2(\textit{Pluvieux}) = 0.3$, $m_2(\textit{Pluvieux, Nuageux}) = 0.1$, $m_2(\Omega) = 0.6$
3	16 $°C$	55 %	23 km/h	$m_3(\textit{Nuageux}) = 0.5$, $m_3(\textit{Ensoleillé, Nuageux}) = 0.2$, $m_3(\Omega) = 0.3$

TABLEAU 3.2 – Jeu de données contenant 3 observations imparfaitement labellisées, x_1 représente la température en degrés Celsius, x_2 l'humidité en pourcentage, x_3 la vitesse du vent en kilomètres par heure et y représente la condition météorologique exprimée par une fonction de masse avec $\Omega = \{\textit{Ensoleillé, Pluvieux, Nuageux}\}$.

mais l'imprécision est liée aux observations. Aucune d'entre elles ne représente plusieurs degrés d'imprécision de l'utilisateur.

Interface

L'interface utilisée est issue du modèle et des outils développés par Thierry, Martin et al. 2021. Les utilisateurs y ont accès et doivent labelliser l'image qui leur est présentée. Deux étapes se succèdent, une première permet de sélectionner plusieurs réponses en attribuant une certitude globale. La deuxième étape offre la possibilité, soit de préciser leur sélection en réduisant le nombre de réponses lorsque le premier choix est imprécis, soit de l'élargir en augmentant le nombre de réponses lorsque le choix précédent est précis. Dans les deux cas, une nouvelle certitude peut y être associée. À chacune des deux étapes, une fonction de masse est générée de la manière suivante. Une masse est attribuée à l'élément focal sélectionné par l'utilisateur, la certitude associée (de 1 à 7, pour suivre une échelle de Likert [2]) réduite à valeurs dans $[0, 1]$ lui est affectée. Les deux fonctions de masse à support simple obtenues sont combinées à l'aide de la règle de combinaison prudente pour obtenir le label imparfait associé à l'image.

De nombreuses campagnes ont été réalisées avec pour objectif de récolter suffisamment de labels riches pour combler l'absence de jeux de données similaires dans la communauté scientifique. Il en résulte cinq jeux de données, sur un spectre diversifié de caractéristiques,

2. L'échelle de Likert est un outil psychométrique utilisé pour mesurer le degré d'accord ou de désaccord de la personne interrogée.

allant de 2 à 10 classes et de 40 à 700 observations. Malgré les efforts de variation, il reste des contraintes non résolues par les données proposées. Par exemple, les jeux de données présentés possèdent tous des variables explicatives quantitatives, et la classe est exclusivement représentée avec une variable qualitative. Il reste donc une grande porte ouverte pour la recherche sur ce sujet. Utiliser le cadre des fonctions de croyance permet notamment de généraliser les probabilités, mais aussi d'être compatible avec d'autres théories du raisonnement avec l'incertain, comme les possibilités ou les approches floues. L'interface complète est présentée sur la figure 3.2 et de manière rapprochée sur la figure 3.1.

Plus le nombre de classes est élevé, plus la théorie des fonctions de croyance permet de représenter un grand degré d'ignorance [3]. Nous avons donc opté pour certains jeux de données avec un nombre de classes important. Cependant, pour des domaines de recherche où il est primordial de comprendre en détail le processus de classification par certains modèles, il est préférable de travailler avec des jeux de données plus simples, à deux classes, comme cela est courant dans la littérature en apprentissage actif. C'est pourquoi nous présentons aussi deux jeux de données à deux classes, ainsi qu'un jeu de données intermédiaire à 4 classes.

Credal Dog-7

Le premier jeu de données présenté, *Credal Dog-7*, est composé d'images de 7 races de chiens parmi les races connues en Europe [4]. Des utilisateurs non spécifiques et rémunérés, au nombre de 50, ont labellisé au total 700 observations réparties uniformément sur toutes les classes. Chaque image est composée de 400 pixels de longueur et 400 pixels de largeur suivant le système de codage informatique des couleurs *rouge, vert, bleu*. Les observations sont donc les images représentées par un vecteur descripteur de dimension 480000. Pour que les jeux de données soient compatibles avec tous types de modèles, nous avons décidé de fournir les caractéristiques des images sous trois formes. Les images en elles-mêmes, interprétables par des humains ou par des modèles d'apprentissage profonds, un vecteur réduit de 512 variables décrivant l'image avec des caractéristiques extraites et un vecteur de composantes principales très réduit à destination des modèles classiques d'apprentissage automatique ou de l'analyse exploratoire de données.

Pour extraire des caractéristiques de l'image, un réseau de neurones ResNet, publié

3. Pour M classes, la théorie des fonctions de croyance permet de représenter une certitude sur 2^M éléments.

4. Les races de chiens présentes sont : *Berger des Shetland, Corgi, Colley, Epagneul Breton, Basset, Foxhound* et *Beagle*.

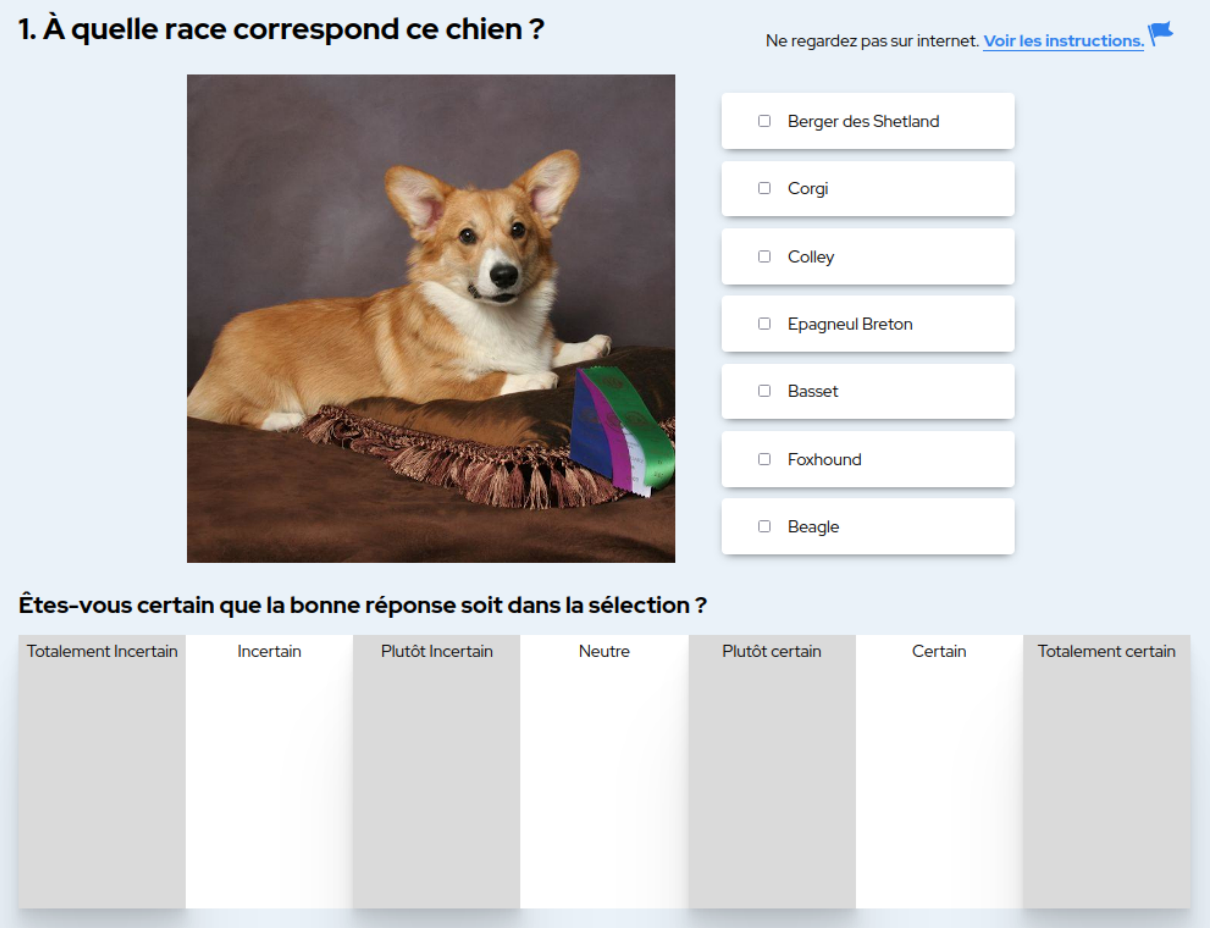


FIGURE 3.1 – Interface utilisateur rapprochée. Campagne Credal Dog-7.

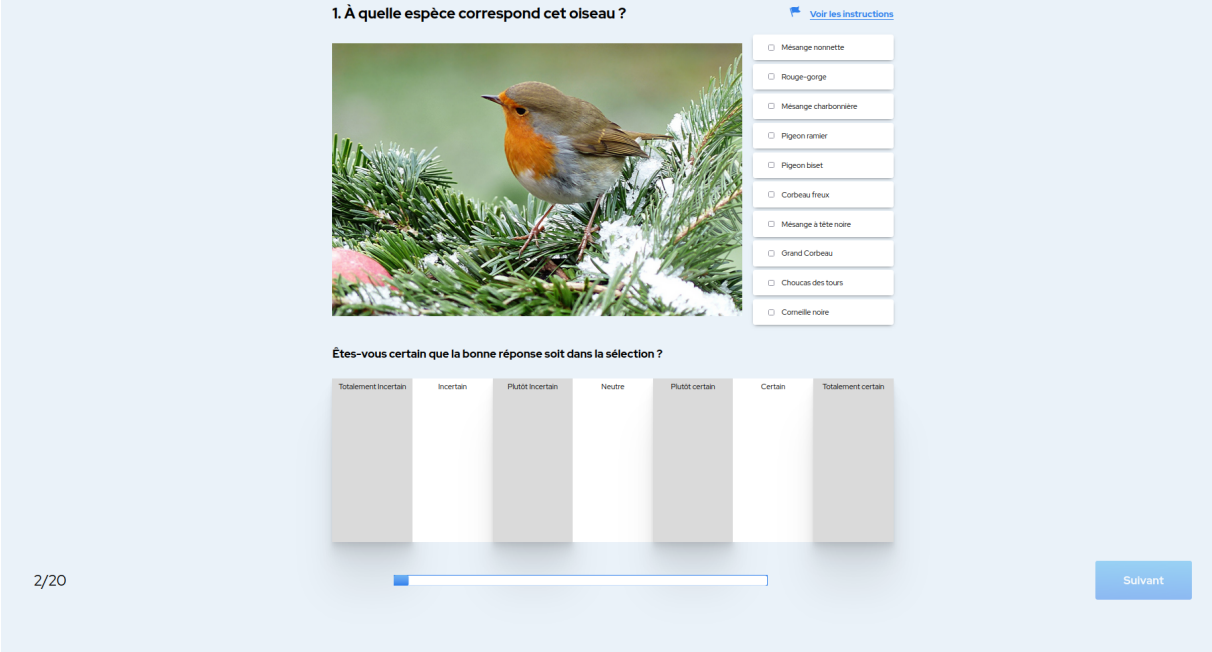


FIGURE 3.2 – Interface utilisateur complète permettant la récolte de labels incertains et imprécis. Campagne Credal Bird-10.

par He, X. Zhang et al. 2016, est entraîné afin de récupérer les 512 variables de l'avant-dernière couche du réseau [5]. Pour réduire à nouveau l'espace de représentation, une analyse en composantes principales est réalisée et l'ensemble des premières composantes retenant un minimum de 70% de l'information forment le vecteur de représentation des caractéristiques de l'image. Pour le jeu de données *Credal Dog-7*, 43 variables sont retenues pour caractériser une observation.

Un exemple du processus qui vise à faire labelliser en deux étapes une observation par un utilisateur est présenté sur la figure 3.3. Lors de la première étape il est demandé à l'utilisateur de choisir un nombre de races de chiens suffisant pour s'assurer avec une certitude élevée, que la bonne réponse est incluse dans la sélection. Sur l'exemple, l'utilisateur choisit *Epagneul Breton*, *Berger des Shetland* et *Beagle* avec une certitude de 6 sur 7. La seconde étape permet à l'utilisateur de réduire son choix, la certitude de sa réponse est donc vouée à baisser, comme le montre Thierry, Martin et al. 2021 [6]. Dans l'exemple, l'utilisateur a choisi *Epagneul Breton* avec une certitude de 3 sur 7. Dans le cas où l'uti-

5. Le ResNet32 utilisé est fortement inspiré de : `https://github.com/ecm200/caltech_birds/blob/master/example_notebooks/002_Train_pytorch_resnet_Caltech_birds.ipynb`

6. L'hypothèse de Smets est la suivante : Plus l'homme est imprécis dans sa réponse, plus sa certitude augmente. On en déduit que plus il est précis plus sa certitude baisse.

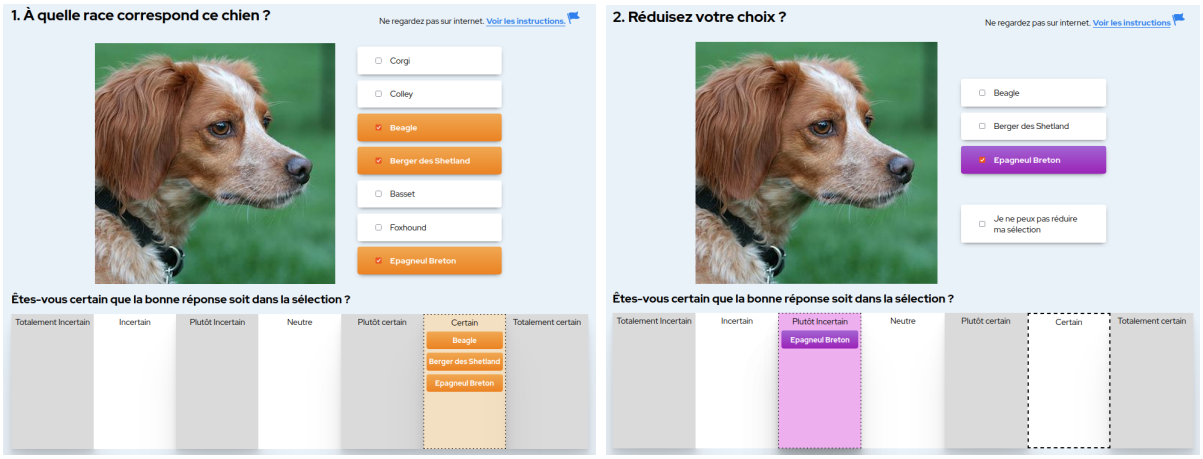


FIGURE 3.3 – Exemple de la réponse d'un utilisateur d'abord imprécise, avec une certitude élevée lors de la première itération, puis plus précise avec une certitude plus faible lors de la seconde itération.

lisateur choisit dans un premier temps une réponse précise mais non totalement certaine, la seconde étape ne lui propose pas de réduire son choix, mais de l'élargir si possible dans l'objectif de gagner en certitude. Si par contre la première réponse est précise (une seule réponse cochée) et certaine (certitude maximale), il n'est pas demandé de seconde étape.

Les deux fonctions de masses m_1 et m_2 respectivement issues du premier et du second choix de l'utilisateur sont donc :

— $m_1 : m_1(\{\omega_1, \omega_2, \omega_3\}) = 0,86,\ m_1(\Omega) = 0,14,$

— $m_2 : m_2(\{\omega_1\}) = 0,43,\ m_2(\Omega) = 0,57,$

avec $\omega_1 =$ *Epagneul Breton*, $\omega_2 =$ *Berger des Shetland* et $\omega_3 =$ *Beagle*. Le label riche m de l'observation est obtenu par la combinaison prudente, introduite à la section 2.1.2, des deux fonctions de masse :

— $m : m(\{\omega_1\}) = 0,43,\ m(\{\omega_1, \omega_2, \omega_3\}) = 0,49,\ m_1(\Omega) = 0,08.$

Les labels du jeu de données sont donc des fonctions de masse qui correspondent à chaque observation labellisée par un utilisateur.

Pour un modèle qui ne prendrait pas en compte des données labellisées imparfaitement, c'est le singleton maximisant la probabilité pignistique qui est choisi comme label dur. Par exemple, sur la figure 3.3, le label associé à l'image est *Epagneul Breton* pour un modèle non crédibiliste. Lorsque plusieurs probabilités sont équivalentes, le label dur est tiré au sort parmi les probabilités maximales.

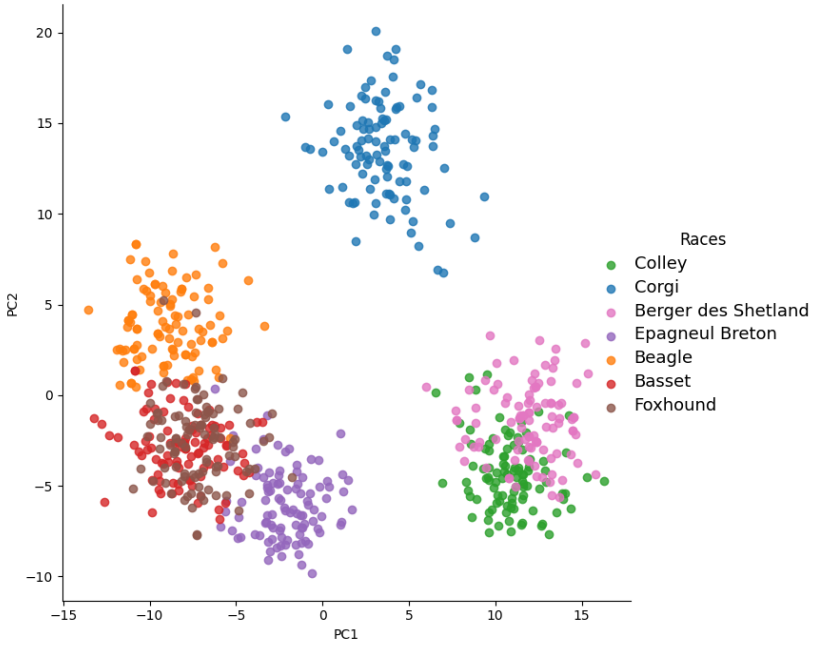


FIGURE 3.4 – Représentation du jeu de données Credal Dog-7 sur le premier plan factoriel d'une ACP retenant 22% de la variance totale.

La figure 3.4 représente le jeu de données sur le premier plan factoriel d'une analyse en composantes principales. On voit clairement apparaître des métaclasses [7], les classes *Berger des Shetland* et *Colley* sont très proches et effectivement, ce sont deux races de chiens très similaires et difficilement discernables pour un utilisateur non expérimenté. Le *Corgi* est laissé à part, il se distingue des autres races assez facilement, même pour un néophyte. Les classes restantes *Epagneul Breton*, *Beagle*, *Basset* et *Foxhound* forment le dernier regroupement.

Pour ce jeu de données, 50% des labels sont donnés avec une itération (partie droite de la figure 3.3), ce qui signifie que la moitié des réponses collectées ont nécessité deux étapes [8]. Parmi les 700 réponses collectées, 237 sont certaines (case *totalement certain* cochée) et 189 d'entre elles contiennent la bonne réponse. Ainsi, lorsqu'un utilisateur pense détenir la bonne réponse, dans 80% des cas celle-ci est effectivement bien présente dans la sélection. Au total, 192 réponses sont à la fois certaines et précises (l'ensemble sélectionné ne contient qu'une race de chien) et 79% de ces réponses sont correctes. Cette

7. Une métaclasse est un regroupement de plusieurs classes, par exemple la classe *café* et la classe *thé* pourraient être regroupées en une classe *boissons chaudes*.

8. L'autre moitié correspond aux utilisateurs qui annoncent savoir parfaitement quelle race est représentée, ou ceux qui ne peuvent pas répondre.

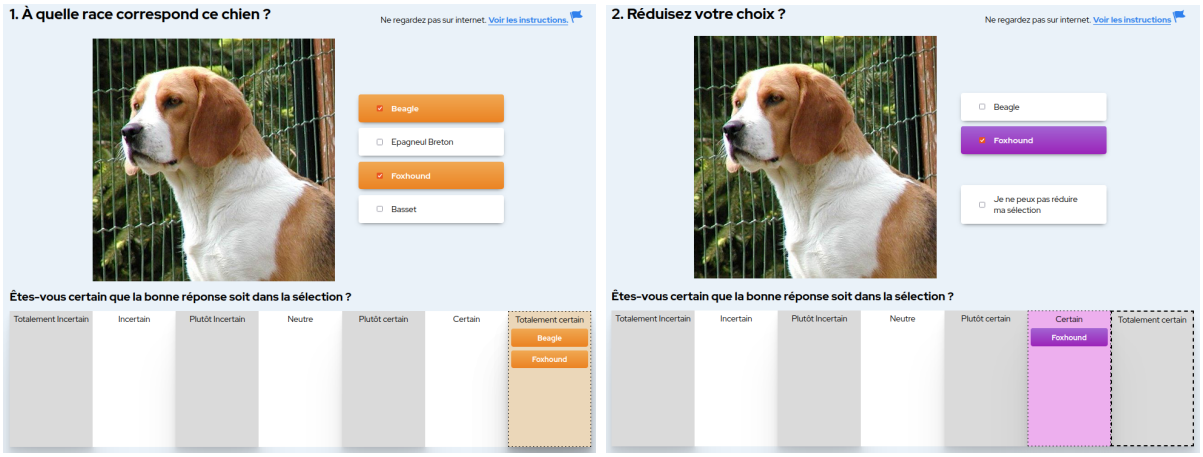


FIGURE 3.5 – Exemple de la réponse d'un utilisateur d'abord imprécise, avec une certitude élevée lors de la première itération, puis plus précise avec une certitude plus faible lors de la seconde itération.

information est importante puisqu'elle montre que les utilisateurs qui affirment connaître la véritable classe ont souvent raison, ce sont les utilisateurs expérimentés.

Credal Dog-4

Ce jeu de données possède de nombreuses similarités avec le précédent, mais il s'agit bien d'un jeu de données totalement distinct, avec de nouvelles variables descriptives et des labels différents, donnés par de nouveaux contributeurs. Un total de 400 observations réparties uniformément sur 4 races de chiens[9] sont labellisées par 50 utilisateurs. Les caractéristiques des images de *Credal Dog-4* sont identiques au premier jeu de données et de manière analogue, 512 variables sont extraites pour caractériser une observation. Une analyse en composantes principales permet ensuite de ne retenir que 47 variables (les composantes principales portant 70% de l'information). L'interface de cette campagne, avec un exemple de labellisation, est présentée sur la figure 3.5. Les deux mêmes étapes sont présentes, ici l'utilisateur sélectionne d'abord les classes *Foxhound* et *Beagle* avec une certitude de 7 sur 7, ce qui veut dire qu'il sait que la bonne réponse se trouve dans sa sélection. Lors de la seconde étape, il choisit *Foxhound* avec une forte certitude de 6 sur 7, ce qui veut dire que parmi sa sélection précédente il a une forte croyance que la vraie classe est *Foxhound*.

Les deux fonctions de masses m_1 et m_2 issues de l'exemple précédent sont données

9. Les races de chiens présentes sont : *Epagneul Breton, Basset, Foxhound* et *Beagle*.

par :

— $m_1 : m_1(\{\omega_1, \omega_2\}) = 1$,

— $m_2 : m_2(\{\omega_1\}) = 0,86$, $m_2(\Omega) = 0,14$,

avec $\omega_1 =$ *Foxhound* et $\omega_2 =$ *Beagle*. Le label riche m de l'observation est obtenu par la combinaison prudente des deux masses :

— $m : m(\{\omega_1\}) = 0,86$, $m(\{\omega_1, \omega_2\}) = 0,14$.

La figure 3.6 représente le jeu de données sur le premier plan factoriel d'une analyse en composantes principales. Les deux races de chiens *Foxhound* et *Beagle* sont très proches, et sont aussi plus difficiles à différencier pour la plupart des utilisateurs, alors que les classes *Epagneul Breton* et *Basset* ont leur propre regroupement.

Pour ce jeu de données, 45% des réponses sont formulées en deux étapes. Ces résultats semblent constants par rapport à *Credal-7* et aux autres expériences présentées. Exactement 124 des 400 réponses sont certaines et parmi elles, 100 contiennent la vraie classe de l'observation, soit 81% de ces réponses. Un total de 107 réponses sont à la fois certaines et précises et 83 sont correctes, soit 78%. Ici aussi, les utilisateurs qui pensent connaître la bonne réponse ont souvent raison, cette information est vérifiée pour toutes les expériences avec des taux de bonnes réponses similaires.

Credal Dog-2

Le jeu de données Credal Dog-2 compte 200 observations réparties uniformément sur deux classes. Des images de *Foxhound* et de *Beagle* sont labellisées par 50 contributeurs. L'extraction de caractéristiques reprend le principe utilisé pour *Credal Dog-7* et *Credal Dog-4* et une analyse en composantes principales permet de retenir 42 variables décrivant chaque observation. La représentation de l'ignorance sur deux classes ne permet pas de représenter une grande quantité d'imprécision, mais souvent en apprentissage automatique, des jeux de données simples, avec peu de classes, sont utilisés pour introduire de nouveaux modèles ou de nouvelles définitions. L'interface utilisée pour cette campagne est présentée sur la figure 3.7. Pour ce jeu de données, lorsqu'un utilisateur sélectionne les deux races de chiens à la première étape, cela n'a pas d'impact sur le label final, puisqu'il s'agit de l'ignorance totale. Seules deux races peuvent donc être sélectionnées et pour la première étape, l'utilisateur choisit *Epagneul Breton* et *Beagle* avec une certitude de 5 sur 7. Lors de la seconde étape, l'utilisateur choisit *Beagle* avec une certitude de 2 sur 7, ce qui veut

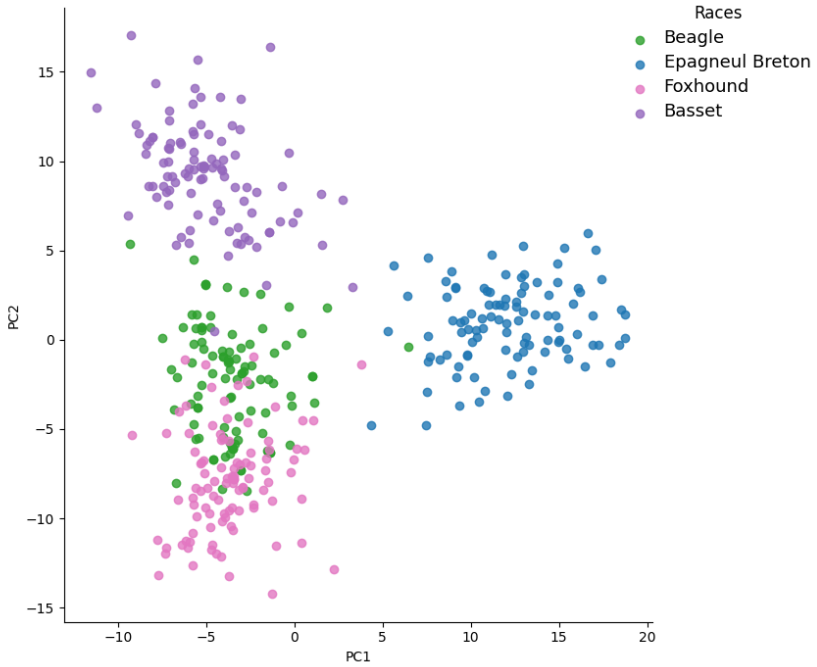


FIGURE 3.6 – Représentation du jeu de données Credal Dog-4 sur le premier plan factoriel d'une ACP retenant 20% de la variance totale.

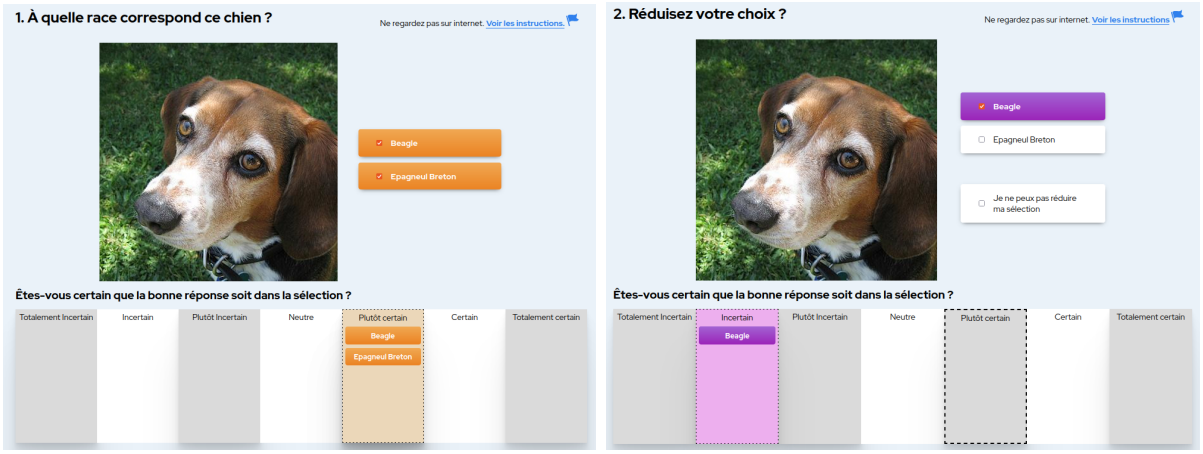


FIGURE 3.7 – Exemple de la réponse d'un utilisateur d'abord imprécise, avec une certitude élevée lors de la première itération, puis plus précise avec une certitude plus faible lors de la seconde itération.

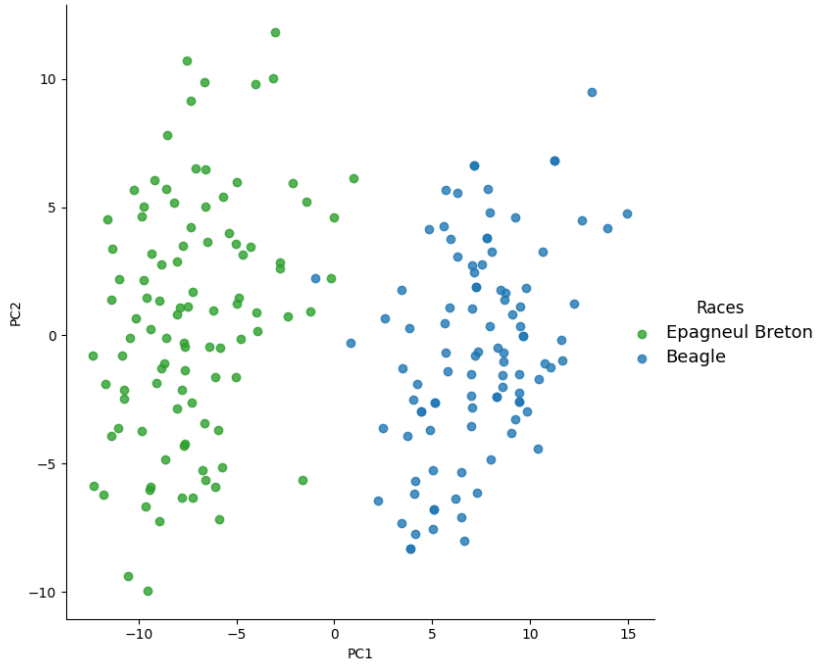


FIGURE 3.8 – Représentation du jeu de données Credal Dog-2 sur le premier plan factoriel d'une ACP retenant 16% de la variance totale.

dire que parmi la sélection précédente, il n'a pas une croyance forte que la vraie classe soir *Beagle* mais il est tout de même plus confiant que pour la race *Epagneul Breton*.

Les deux fonctions de masses m_1 et m_2 résultant de l'exemple précédent sont données par :

— $m_1 : m_1(\Omega) = 1$,

— $m_2 : m_2(\{\omega_1\}) = 0,29,\ m_2(\Omega) = 0,71$,

avec $\omega_1 =$ *Beagle* et $\omega_2 =$ *Epagneul Breton*. Le label riche m de l'observation est obtenu par la combinaison prudente des deux masses :

— $m : m(\{\omega_1\}) = 0,29,\ m(\Omega) = 0,71$.

La figure 3.8 représente *Credal Dog-2* sur le premier plan factoriel d'une analyse en composantes principales. Les deux classes sont distinctes et bien représentées.

Credal Bird-10

Credal Bird-10 est un jeu de données contenant 200 images d'oiseaux imparfaitement labellisées. Chacune de ces images appartient à une classe correspondant à l'une des 10

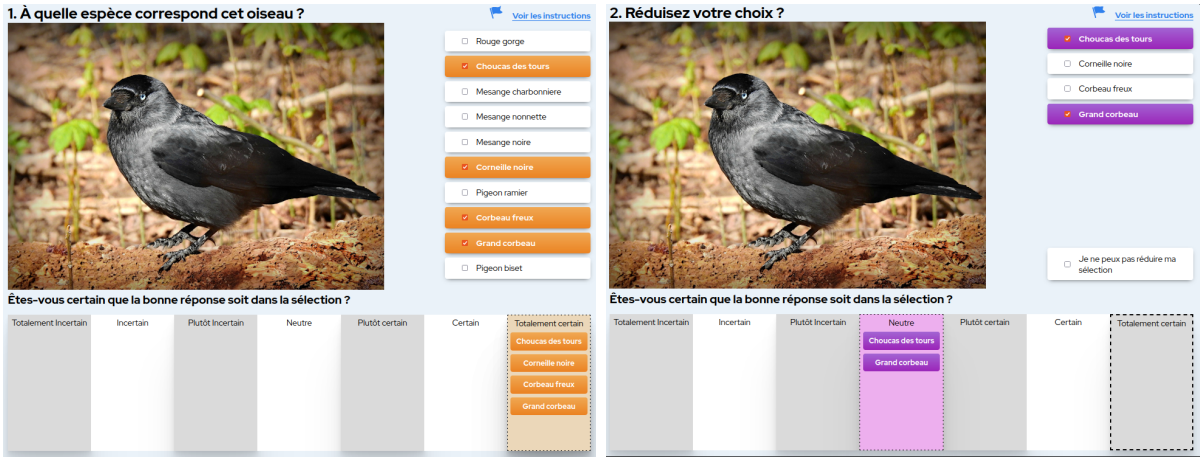


FIGURE 3.9 – Exemple de la réponse d'un utilisateur d'abord imprécise, avec une certitude élevée lors de la première itération, puis plus précise avec une certitude plus faible lors de la seconde itération.

espèces d'oiseaux [10] distribuées uniformément sur le jeu de données. Ce jeu de données est obtenu à partir de la même interface, mais cette fois-ci, le nombre d'observations n'est pas égal au produit entre le nombre de contributeurs et le nombre de contributions par contributeur. Précédemment, chaque observation se voyait attribuer un label unique par un seul utilisateur. Ici, plusieurs contributeurs peuvent labelliser la même image, un seul label est ensuite tiré aléatoirement pour caractériser l'observation. À l'exception du format des images, les autres modalités de l'expérience sont similaires, 512 caractéristiques sont extraites pour chaque observation et une analyse en composantes principales permet d'extraire 30 variables retenant 70% de l'information.

La figure 3.9 représente un exemple de labellisation du jeu de données. L'utilisateur choisit d'abord *Choucas des tours*, *Corneille noire*, *Corbeau freux* et *Grand corbeau* avec une certitude de 7 sur 7. Pour la seconde étape, il choisit *Choucas des tours* et *Grand corbeau* avec une certitude de 4 sur 7. Les deux fonctions de masses m_1 et m_2 issues des réponses sont donc :

— $m_1 : m_1(\{\omega_1, \omega_2, \omega_3, \omega_4\}) = 1$,

— $m_2 : m_2(\{\omega_1, \omega_2\}) = 0,57,\ m_2(\Omega) = 0,43$,

avec $\omega_1 =$ *Choucas des tours*, $\omega_2 =$ *Grand corbeau*, $\omega_3 =$ *Corneille noire* et $\omega_1 =$ *Corbeau freux*. Le label riche m de l'observation est obtenu par la combinaison prudente des deux

10. Les espèces d'oiseaux présentes sont : *Rouge gorge, Choucas des tours, Corneille noire, Corbeaux freux, Grand corbeau, Mésange charbonnière, Mésange nonnette, Mésange noire, Pigeon ramier* et *Pigeon biset.*

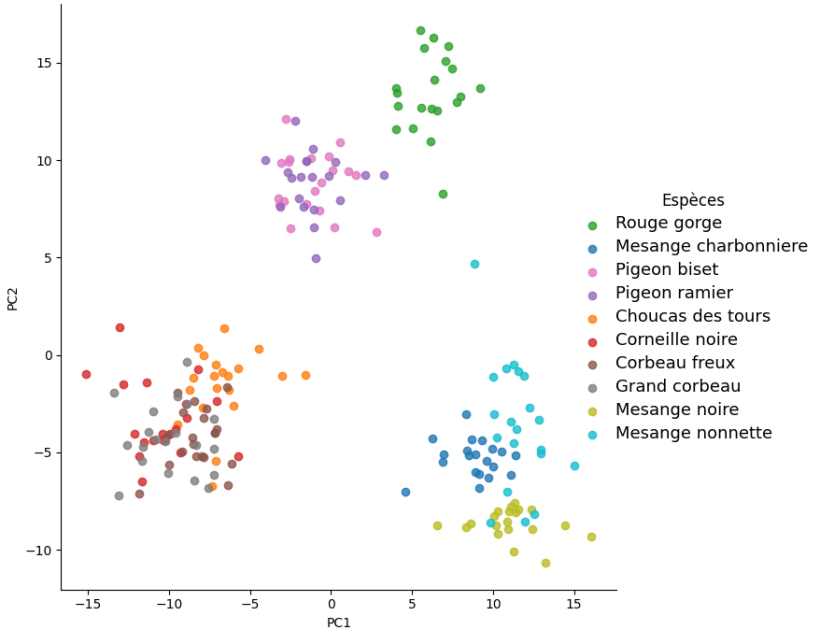


FIGURE 3.10 – Représentation du jeu de données Credal Bird-10 sur le premier plan factoriel d'une ACP retenant 25% de la variance totale.

masses :

— m : $m(\{\omega_1, \omega_2\}) = 0,57$, $m(\{\omega_1, \omega_2, \omega_3, \omega_4\}) = 0,43$.

Ce label contient deux degrés d'ignorance, un sur $\{\omega_1, \omega_2\}$ et un sur $\{\omega_1, \omega_2, \omega_3, \omega_4\}$, mais il n'y a pas d'incertitude sur l'ignorance totale puisque $m(\Omega) = 0$.

La représentation du jeu de données sur le premier plan factoriel de l'analyse en composantes principales est donnée sur la figure 3.10. Ici aussi la formation de métaclasses est visible, toutes les mésanges (*Mésange nonnette*, *Mésange charbonnière* et *Mésange noire*) sont regroupées dans l'espace de représentation ainsi que les corvidés (*Choucas des tours*, *Corneille noire*, *Corbeau freux* et *Grand corbeau*) et les columbidés (*Pigeon ramier* et *Pigeon biset*). Seul le *Rouge gorge* est mis à part, avec une ressemblance plus importante au groupe des columbidés qu'aux autres groupes.

Pour ce jeu de données, 46% des réponses sont à deux étapes, avec 82% de bonnes réponses parmi les réponses certaines et 91% des bonnes réponses parmi les réponses certaines et précises.

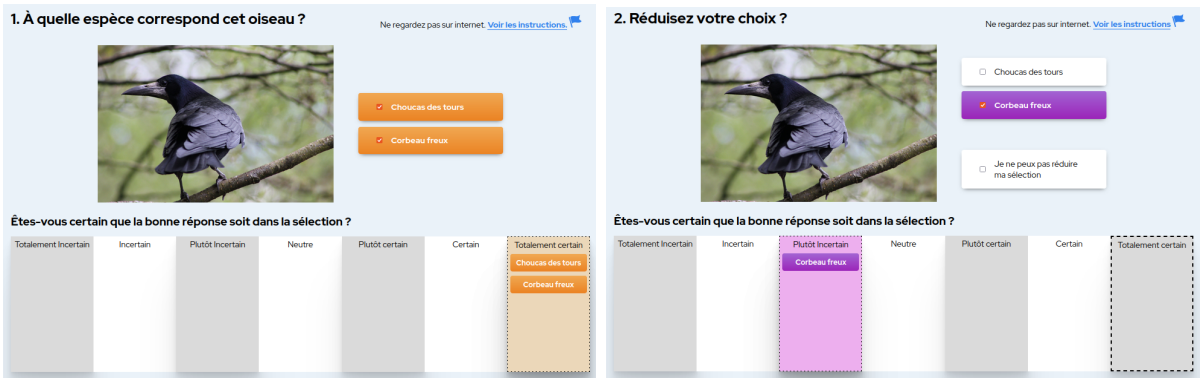


FIGURE 3.11 – Exemple de la réponse d'un utilisateur d'abord imprécise, avec une certitude élevée lors de la première itération, puis plus précise avec une certitude plus faible lors de la seconde itération.

Credal Bird-2

Le dernier jeu de données comporte deux classes, *Choucas des tours* et *Corbeau freux*. Il s'agit du plus petit jeu de données avec 40 observations. Les caractéristiques du jeu sont réduites à 17 variables labellisées par 50 contributeurs de la même manière que *Credal Bird-10*. La combinaison des réponses est la même que pour *Credal Dog-2* et un exemple de l'interface utilisateur est présenté sur la figure 3.11. L'utilisateur sélectionne d'abord les deux classes avec une certitude de 7 sur 7, puis lors de la seconde étape, il sélectionne *Corbeau freux* avec une certitude de 3 sur 7. Sa certitude la plus importante est donc sur la classe *Corbeau freux*, mais celle-ci reste faible. Les deux fonctions de masses m_1 et m_2 sont donc :

— $m_1 : m_1(\Omega) = 1$,

— $m_2 : m_2(\{\omega_1\}) = 0,43,\ m_2(\Omega) = 0,57$,

avec $\omega_1 =$ *Corbeau freux* et $\omega_2 =$ *Choucas des tours*. Le label riche m de l'observation est obtenu par la combinaison prudente des deux masses :

— $m : m(\{\omega_1\}) = 0,43,\ m(\Omega) = 0,57$.

La représentation du jeu de données sur le premier plan factoriel d'une analyse en composantes principales est donné sur la figure 3.12.

Expérience de classification

Plusieurs applications sont possibles pour de tels jeux de données, allant de l'apprentissage supervisé à l'apprentissage non supervisé, en passant par l'apprentissage actif.

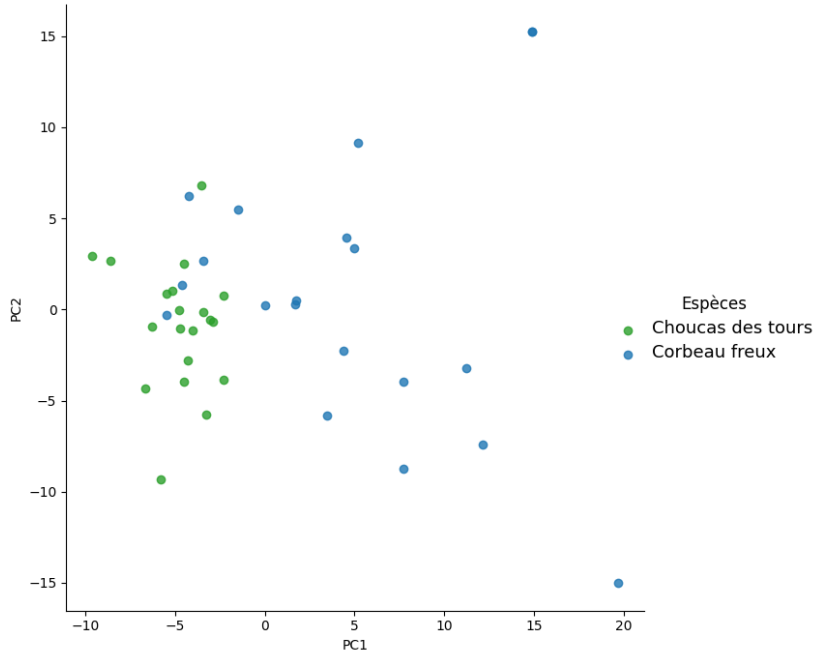


FIGURE 3.12 – Représentation du jeu de données Credal Bird-2 sur le premier plan factoriel d'une ACP retenant 16% de la variance totale.

Ils peuvent être utilisés sur un large spectre d'applications et grâce à l'utilisation des fonctions de croyance, de Shafer 1976 ; des probabilités, des possibilités ou encore des sous-ensembles flous peuvent en être extraits. Ces jeux de données sont compatibles avec des modèles crédibilistes, comme ceux de Elouedi, Mellouli et al. 2001, de Denœux 1995 et de Denoeux et Bjanger 2000 mais aussi avec de nombreux autres, comme par exemple des modèles flous ou des modèles issus des probabilités imprécises. L'information ajoutée dans les labels peut aussi servir la représentation de l'incertitude du modèle, par exemple en apprentissage actif, comme le montre Hoarau, Martin, J.-C. Dubois et Le Gall 2022.

L'idée de se rapprocher de ce que pense réellement l'utilisateur lors de la labellisation peut aussi améliorer les performances du modèle utilisé. Par exemple, il serait préférable qu'un contributeur soit capable de dire qu'il est incertain de sa réponse fausse plutôt que d'avoir simplement un faux label dur. Ce type de jeux de données permet de montrer qu'en donnant la possibilité à un utilisateur de répondre de manière incertaine et imprécise, il peut donner une information plus fiable qu'avec un label dur. Pour l'expérience présentée dans le tableau 3.3, toutes les campagnes précédentes ont été réitérées, mais cette fois-ci, un seul label dur est demandé aux contributeurs. Le modèle des K plus proches voisins crédibiliste, introduit par Denœux 1995, est ensuite utilisé en classification, avec

Datasets	Labels durs	Labels riches
Credal Dog-7	68.7 ± 0.8	75.8 ± 0.7
Credal Dog-4	70.8 ± 1.0	69.3 ± 1.0
Credal Dog-2	98.4 ± 0.5	98.0 ± 0.4
Credal Bird-10	52.8 ± 1.5	60.7 ± 1.5
Credal Bird-2	51.6 ± 3.1	63.5 ± 3.7

TABLEAU 3.3 – Performance moyenne (± un intervalle de confiance à 95% pour l'estimation de la moyenne), quand l'utilisateur a la possiblité de répondre de manière incertaine et imprécise (labels riches) et quand il n'a pas cette possibilité (labels durs).

Nom	Classes	Observations	Caractéristiques	Contributeurs
Credal Dog-7	7	700	43	50
Credal Dog-4	4	400	47	50
Credal Dog-2	2	200	42	50
Credal Bird-10	10	200	30	50
Credal Bird-2	2	40	17	50

TABLEAU 3.4 – Jeux de données *Credal*.

une validation croisée à 5 plis pour estimer le meilleur paramètre K. La moyenne d'exactitude du modèle est présentée sur 100 expériences, et montre que donner la possibilité à un utilisateur de répondre imparfaitement peut en effet augmenter les performances du modèle.

Accès aux données

Ces jeux de données sont disponibles à l'adresse suivante : `https://github.com/ArthurHoa/credal-datasets`. La composition de chaque répertoire est la suivante ; le fichier *README.md* est un résumé du jeu de données, *classes.csv* contient les classes ordonnées, *X_512.csv* contient les vecteurs larges de caractéristiques, *X.csv* contient les caractéristiques des observations, *X_pictures.csv* contient le nom de l'image pour chaque observation (les images elles-mêmes sont aussi disponibles, mais le répertoire n'est pas le même, suivre les instructions de *README.md*), *y_true.csv* contient la vraie classe de chaque observation et *y.csv* contient les labels riches. Les autres outils sont disponibles dans le répertoire *experiment*. Tous les jeux de données présentés sont résumés dans le tableau 3.4.

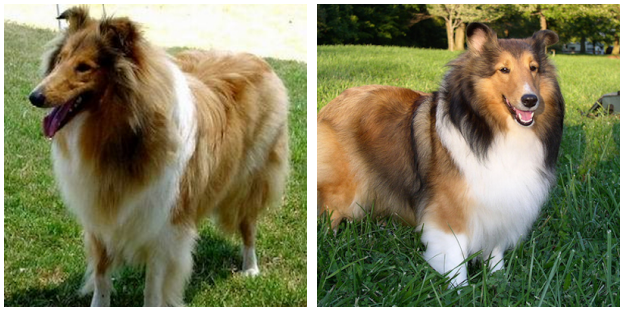
3.3 Inconsistance liée à la difficulté du problème

Lors des campagnes de production participatives, réalisées pour obtenir les labels riches des observations, une inconsistance a été observée quant à la qualité de la représentation de son incertitude et de son imprécision par un utilisateur. Lorsque le problème était simple, les utilisateurs sont parvenus avec plus de facilité à représenter leur ignorance. Intuitivement, il serait possible de se dire que plus le problème est difficile, plus l'ignorance des utilisateurs va augmenter, mais que la capacité de l'utilisateur à dire qu'il sait ou ne sait pas va rester constante. Cependant, pour un problème plus complexe, les réponses incohérentes se multiplient. Ces réponses prennent la forme de "Je suis certain qu'il s'agit d'un *Beagle*" alors qu'il s'agit d'un *Foxhound*, ou plus généralement de réponses où les utilisateurs sélectionnent une certitude qui ne s'avère pas représenter leur connaissance.

Deux versions de la campagne *Credal Dog-2* sont présentées ici pour illustrer cette inconsistance. La figure 3.13 représente les races de chiens utilisées pour les deux versions de la campagne. Lors de la première campagne les deux races *Colley* (Figure 3.13a) et *Berger des shetland* (Figure 3.13b) sont utilisées. Ces deux races de chiens sont plus difficiles à différencier, et seul un utilisateur très expérimenté serait capable de reconnaître l'une des deux races avec certitude. Pour la seconde campagne, les races *Beagle* (Figure 3.13c) et *Epagneul breton* (Figure 3.13d) sont utilisées et sont beaucoup plus simples à différencier.

La figure 3.14 représente les labels riches, issus de la récolte des réponses des utilisateurs pour ces deux campagnes. Plus la couleur est affirmée, plus la réponse de l'utilisateur est certaine, plus elle est claire, plus il est ignorant. Pour les deux expériences, les classes sont fortement séparables, ce qui veut dire que sur les deux représentations, les deux couleurs doivent être séparées (à droite et à gauche) lorsque l'on utilise les vrais labels. Une hypothèse serait de s'attendre à avoir une représentation plus claire pour la campagne présente sur la figure 3.14a et une représentation plus foncée pour la campagne de la figure 3.14b, l'une plus difficile, donc avec plus d'ignorance représentée et l'autre plus simple, avec des utilisateurs plus certains. Cependant, ce n'est pas ce qui est observé, la campagne définie comme plus simple (Figure 3.14b) obtient les résultats escomptés, les classes sont distinctes, certains utilisateurs se trompent, mais dans l'ensemble, les utilisateurs certains et précis possèdent la vraie réponse. L'inconsistance est liée à la première campagne, définie comme plus difficile (Figure 3.14a), ici les utilisateurs se trompent beaucoup plus et n'arrivent pas à représenter correctement leur incertitude et leur imprécision.

Lorsqu'on entraîne sur ces données un modèle capable de prendre en compte ces labels



(a) Colley (b) Berger des Shetland

(c) Beagle (d) Epagneul Breton

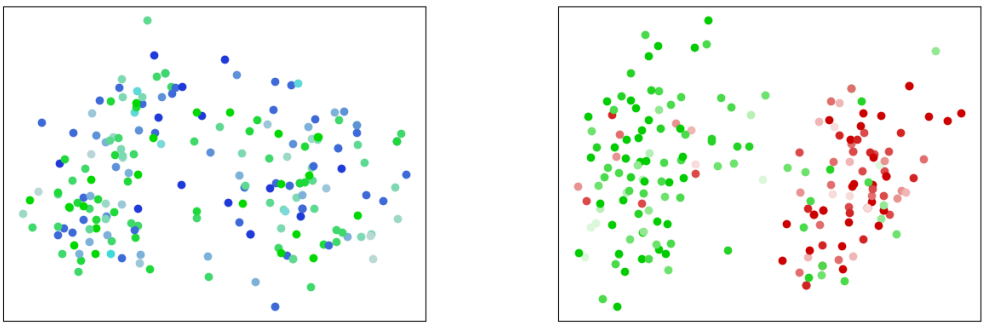
FIGURE 3.13 – Différence entre les deux campagnes *Credal Dog-2* réalisées, la première avec (a) et (b), la seconde avec (c) et (d).

riches, les forêts aléatoires crédibilistes [11] par exemple, on obtient 54.4% de prédictions correctes en moyenne pour la campagne difficile et 93.8% de prédictions correctes pour la campagne plus simple [12]. Sachant qu'un modèle complètement aléatoire obtient 50% sur un jeu de données à deux classes distribuées identiquement, les 54.4% de la campagne difficile montrent qu'il n'est pas évident d'exploiter l'information donnée par les utilisateurs sur cette expérience, alors qu'avec la connaissance issue de la seconde campagne, les performances sont bien plus importantes.

3.4 Problématique des sources non fiables

Ces travaux se placent sous une hypothèse de fiabilité des sources. Ce qui ne veut pas dire que les contributeurs possèdent la bonne réponse mais qu'ils sont capables de représenter correctement leur ignorance. Un utilisateur qui est ignorant aura peu d'impact lors

11. Les forêts aléatoires crédibilistes sont introduites dans ce document à la section 4.3.
12. Sur des labels parfaits, les performances sont de 95% pour la campagne difficile et 99% pour la campagne plus simple.



(a) (b)

FIGURE 3.14 – Différence après labellisation imparfaite entre les deux campagnes *Credal Dog-2* réalisées, la première entre *Colley* et *Berger des shetland* (a) et la seconde entre *Beagle* et *Epagneul breton* (b).

de la phase de classification pour un jeu de données. Cependant, lorsqu'un contributeur ment ou se trompe dans la représentation de son incertitude et de son imprécision, il peut induire un biais important. Le cas évoqué dans la section précédente en est un exemple particulier, où le problème est trop complexe pour que les utilisateurs puissent représenter correctement leur ignorance. Les travaux de Thierry, Martin et al. 2023 ont été réalisés en ce sens, avec pour objectif de qualifier les contributeurs lors de campagnes de production participative. Quatre profils sont retenus, et permettent pas la suite, par exemple, d'affaiblir les réponses d'utilisateurs non fiables. L'*expert* est le contributeur le plus qualifié, qui possède la connaissance pour un sujet donné. Le *bon contributeur* a moins de connaissance, mais prend le temps nécessaire pour répondre et est capable de représenter son ignorance. Le *contributeur moyen* a lui aussi une connaissance limitée, mais peut représenter un doute plus fort lors de sa réponse sous forme d'une certitude plus faible. Le *mauvais contributeur* n'est pas forcément le moins qualifié, mais peut perdre patience et se hâter pour répondre en espérant finir le questionnaire au plus vite. Il peut avoir un profile malveillant, une perception biaisée ou encore être trop sûr de lui.

3.5 Conclusion du chapitre

Cinq jeux de données imparfaitement labellisés *Credal Dog-7*, *Credal Dog-4*, *Credal Dog-2*, *Credal Bird-10* et *Credal Bird-2* ont été présentés et seront utilisés lors des ex-

périences menées dans le reste de ce document. Permettre à une source de répondre de manière incertaine et imprécise offre la possibilité à ceux qui sont certains de leur réponse, d'avoir un impact plus important lors de la phase de classification. Deux étapes permettent d'obtenir plusieurs degrés d'imprécision sur des labels modélisés avec la théorie des fonctions de croyance. L'utilisation de la théorie des fonctions de croyance donne également une grande flexibilité pour la représentation des labels, rendant possible le parallèle avec d'autres outils de raisonnement avec l'incertain, comme les sous-ensembles flous, les possibilités, ou encore les probabilités. Enfin, ces jeux de données, gratuits et mis à disposition de la communauté scientifique, seront également un critère supplémentaire pour comparer les performances de différents modèles.

Annexe du chapitre : remerciements

Les images des jeux de données *Credal Dog* sont pour la majorité des images redimensionnées issues d'ImageNet, publié par Deng, Dong et al. 2009, et complétées par des apports personnels. Ces jeux de données sont fortement inspirés du jeu de données *Sanford Dogs*, de Khosla, Jayadevaprakash et al. 2011. Pour ceux sur les oiseaux, certaines images sont de notre équipe, nous souhaitons aussi remercier les contributeurs de Pixabay et de Wikimedia, tous les alias sont listés dans le répertoire.

Chapitre 4

Modèles crédibilistes d'apprentissage

Toujours avec l'objectif de travailler sur la qualité des données, on s'intéresse ici aux modèles capables de prendre en compte des labels riches. Plusieurs modèles d'apprentissage utilisant les fonctions de croyance ont été introduits (voir Côme, Oukhellou et al. 2009 ; Denœux 1995 ; Elouedi, Mellouli et al. 2001 ; Z. Tong, Xu et al. 2021), dont une version crédibiliste [1] des arbres de décision par Elouedi, Mellouli et al. 2001 et plus récemment, un modèle présenté par Z. Tong, Xu et al. 2021 permettant d'associer la théorie des fonctions de croyance aux réseaux de neurones convolutifs. Dans ce chapitre, une nouvelle version des K plus proches voisins crédibiliste est proposée afin de pouvoir travailler avec des données imparfaitement labellisées tout en gardant une équivalence avec le modèle originel. L'équivalence sera démontrée théoriquement et expérimentalement. Un nouveau modèle d'arbres de décision crédibiliste est aussi présenté, plus compétitif et robuste au surapprentissage. Enfin, pour pallier la variance élevée de ce modèle, nous introduisons les forêts aléatoires crédibilistes, capables à la fois de traiter des labels riches et de faire des prédictions incertaines et imprécises, c'est-à-dire sous forme de fonctions de croyance.

Publications

— Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2022), « Imperfect Labels with Belief Functions for Active Learning », in : *Belief Functions : Theory and Applications*, p. 44-53

— Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2023), « Evidential Random Forests », in : *Expert Systems with Applications* 230

1. *Modèle crédibiliste* fait ici référence à un modèle qui fait intervenir la théorie des fonctions de croyance. *Evidential* en anglais.

4.1 Modèle des K plus proches voisins crédibilistes

La première contribution n'est pas un nouveau modèle, mais la modification d'un paramètre des K plus proches voisins crédibiliste. L'objectif, comme pour tous les modèles présentés dans ce chapitre, est d'avoir un modèle capable de faire une prédiction incertaine et imprécise mais aussi de le rendre compatible avec des labels riches. Pour cette contribution, il s'agit en fait de l'introduction d'un nouvel outil, la mesure de distance moyenne entre les éléments d'une même classe, lorsqu'il n'y a plus de labels durs, mais des labels riches, modélisés avec la théorie des fonctions de croyance. Cette distance moyenne, appliquée aux K plus proches voisins, permet d'en faire un modèle crédibiliste compatible avec nos exigences.

4.1.1 Modèle des K plus proches voisins

Dans un contexte de données parfaitement labellisées, un modèle de discrimination non paramétrique connu sous le nom des K plus proches voisins [2], noté K-NN, a été introduit par Fix et Hodges 1951. C'est un modèle populaire d'apprentissage supervisé où la classe d'une observation est prédite en fonction de ses K plus proches voisins (voir figure 4.1 [3]). La dispersion des voisins autour de l'observation peut avoir un impact sur la classification. Une proposition est alors faite par Dudani 1976 pour ajouter un poids aux voisins en fonction de leur distance à l'observation. L'hypothèse suivante est alors faite : il serait raisonnable de pondérer plus fortement l'information issue d'un voisin proche de l'observation à classer que celle d'un voisin avec lequel la distance est plus importante. L'auteur propose pour le j-ème voisin la pondération w_j suivante :

$$w_j = \begin{cases} \dfrac{d_K - d_j}{d_K - d_1}, & d_K \neq d_1 \\ 1 & d_K = d_1 \end{cases} \tag{4.1}$$

Avec d_K la distance entre l'observation à classer et son voisin le plus éloigné (parmi ses K plus proches voisins), d_1 la distance avec son voisin le plus proche et d_j la distance avec le voisin j. Cette version de K-NN pondérée par distance sera la référence utilisée pour le reste de ce chapitre.

2. K-Nearest Neighbors en anglais (K-NN)
3. Antti Ajanki AnAj, CC BY-SA 3.0, *via* Wikimedia Commons

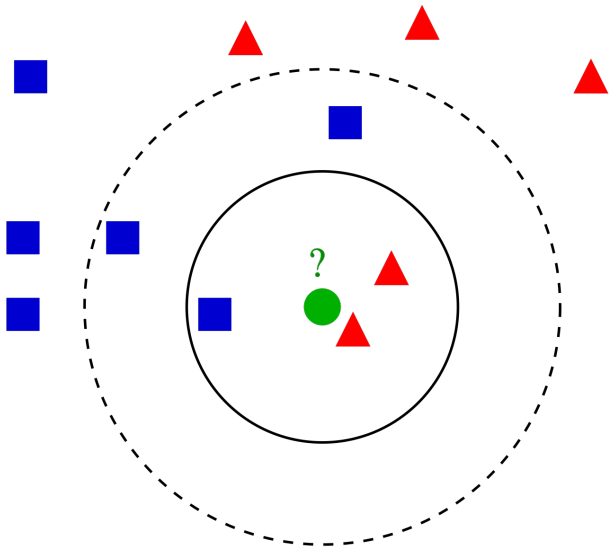


FIGURE 4.1 – Exemple de classification avec les K plus proches voisins.

4.1.2 EK-NN : versions crédibilistes

Une version crédibiliste de K-NN est introduite par Denœux 1995 : EK-NN [4]. Elle utilise la théorie des fonctions de croyance pour assigner un label à une observation. Dans l'article originel, ce modèle est présenté comme utilisant des données labellisées parfaitement. Plus tard, plusieurs travaux ont permis de le rendre compatible avec des données imparfaitement labellisées. Les auteurs Denœux et Zouhal 2001 proposent une version de EK-NN utilisant des données labellisées avec la théorie des possibilités et T. Denoeux, Kanjanatarakul et al. 2019 proposent de dissoudre l'ambiguïté présente dans le calcul d'un des paramètres quand il s'agit de données imparfaitement labellisées avec la théorie des fonctions de croyance, cependant l'équivalence avec le modèle originel est alors perdue.

4.1.3 γ_i-EKNN : nouvelle version

Présentation du modèle de Denœux

Soit $\mathcal{X}$ une collection de N observations décrites par une collection de variables de taille P telle que $\mathcal{X} = \{x^n = (x_1^n, \ldots, x_P^n) | n = 1, \ldots, N\}$, et Ω un ensemble de C classes tel que $\Omega = \{\omega_1, \ldots, \omega_C\}$. La distance $d^{s,i}$ représente la distance entre x^s et x^i. Avec x^s l'observation à classer en utilisant les informations contenues dans le jeu d'entraînement

4. EK-NN pour "Evidential K-Nearest Neighbors".

et x^i un de ses K plus proches voisins. Classer x^s signifie lui attribuer un élément de Ω. L'ensemble Φ^s représente les K plus proches voisins de x^s dans $\mathcal{X}$, et m_i est la fonction de masse associée au voisin x^i.

Dans la première publication du modèle par Denœux 1995, l'auteur introduit l'équation d'une fonction de masse en fonction d'une observation x^s et un voisin x^i pour des données imparfaitement labellisées.

Proposition de Denœux

Si x^s est une observation à classer, la croyance sur la classe de x^s induite par la connaissance $x^i \in \Phi^s$, peut être représentée par une fonction de masse $m_{s,i}$ déduite de m_i et $d^{s,i}$.

$$\begin{aligned} m_{s,i}(A) &= \alpha_0 \varphi(d^{s,i}) m_i(A) \\ m_{s,i}(\Omega) &= 1 - \sum_{A \in 2^\Omega \setminus \Omega} m_{s,i}(A) \end{aligned} \tag{4.2}$$

Avec φ une fonction monotone décroissante et :

$$\begin{aligned} 0 < \alpha_0 < 1 \\ \varphi(0) = 1 \\ \lim_{d \to \infty} \varphi(d) = 0 \end{aligned} \tag{4.3}$$

Pour la fonction décroissante φ, l'auteur suggère de choisir :

$$\varphi(d) = e^{-\gamma d^\beta} \tag{4.4}$$

Avec $\gamma > 0$ et $\beta \in \{1, 2, \ldots\}$. β possiblement fixé à une valeur faible.
Quand φ_q est d'abord introduite, elle fait référence à la fonction qui augmente l'ignorance de $m_{s,i}$ à mesure que la distance entre x^s et x^i augmente. La fonction φ_q dépend de la classe q du voisin x^i et il y a autant de φ_q et donc de γ_q que de classes différentes. Sachant qu'en se plaçant dans un contexte de données imparfaitement labellisées, le voisin x^i n'a plus un unique label, mais une fonction de masse qui lui est attribuée, γ_q et donc φ_q ne peut être calculé. Cette spécificité force le modèle à différer d'un modèle utilisant des données parfaitement labellisées. Cette différence est traitée dans la partie suivante.

Toutes les fonctions de masse sont alors combinées en utilisant la règle de combinaison conjonctive, donnée par l'équation (2.11), nous avons donc :

$$m_s^{'}(A) = \sum_{B_1 \cap \ldots \cap B_K = A} \prod_{x^i \in \Phi^s} \alpha_0 \varphi(d^{s,i}) m_i(B_i) \qquad \forall A \in 2^\Omega \tag{4.5}$$

En considérant un monde fermé, la masse sur l'ensemble vide doit être forcée à nulle. La nouvelle combinaison normalisée, donnée par l'équation (2.12), notée m_s est obtenue par :

$$\begin{cases} m_s(A) = \dfrac{1}{1-\kappa} m_s^{'}(A), & A \neq \emptyset \\ m_s(\emptyset) = 0 \end{cases} \tag{4.6}$$

Avec κ l'inconsistance de la fusion (voir chapitre 2).

Chaque nouvelle observation peut alors être classée en maximisant la probabilité pignistique.

Optimisation des paramètres

Trois paramètres sont retenus : K, α_0, et β. Le nombre K de plus proches voisins peut être optimisé identiquement à K-NN, en utilisant par exemple une validation croisée. De plus, l'utilisation d'un jeu de données de taille évolutive au sein de l'apprentissage actif a un impact sur la valeur optimale de K. La valeur $0,8$ est affectée au paramètre α_0 mais peut dépendre de la connaissance relative aux sources, donnant des résultats légèrement différents. Le dernier paramètre donne des résultats satisfaisants pour $\beta = 2$, avec un très faible impact lorsqu'il est changé.

Lorsqu'il s'agit de données imparfaitement labellisées, l'utilisation d'un γ par classe devient impossible puisqu'il n'y a plus de classes, mais uniquement des fonctions de masse associées à des observations. Plusieurs options ont été proposées par Denœux 1995 puis par Denœux et Zouhal 2001 et par T. Denoeux, Kanjanatarakul et al. 2019 et comparées dans la publication de T. Denoeux, Kanjanatarakul et al. 2019.

— Dans sa première version, introduite par Denœux 1995, et renommée ici γ_q-EKNN, le modèle est présenté avec un paramètre γ_q dépendant de la classe ω_q du voisin x^i. La formule utilisée pour calculer γ_q est alors $1/\mathsf{d}_q^\beta$ avec d_q la distance moyenne entre

deux vecteurs d'entraînement appartenant à la classe ω_q.

— Une version avec un unique paramètre γ, introduite par Denœux et Zouhal 2001, γ-EKNN, est ensuite proposée dans un environnement possibiliste et compatible avec des données imparfaitement labellisées. L'utilisation d'un unique paramètre γ induit une perte d'équivalence avec le modèle initial des K plus proches voisins crédibilistes.

— Enfin, un modèle fondé sur un affaiblissement contextuel, proposé par T. Denoeux, Kanjanatarakul et al. 2019, avec M paramètres γ_q optimisables, est introduit et sera noté CD-EKNN.

γ_i-EKNN ou la distance moyenne entre éléments d'une même classe pour des labels riches

Nous proposons ici la notion de distance moyenne entre les éléments d'une même classe pour des labels riches. Il est courant de calculer la distance moyenne entre les éléments d'une même classe pour des labels durs, il suffit de prendre tous les éléments appartenant à la classe en question, de faire une somme deux à deux de leurs distances et de la diviser par le nombre total sommé. On a donc la formule suivante de la distance moyenne d_q entre les éléments de classe q :

$$\mathsf{d}_q = \frac{\sum_{\nu=1}^{N_q} \sum_{\mu=1}^{N_q} d^{\nu,\mu}}{N_q^2 - N_q}, \tag{4.7}$$

avec N_q le nombre total d'observations. La distance moyenne entre les éléments d'une même classe est notamment utilisée dans le modèle originel des K plus proches voisins crédibilistes (*cf.* Denœux 1995). Cependant, lorsqu'on travaille avec des labels riches, plus aucune observation *n'appartient* à une classe, puisqu'elles sont toutes représentées par des fonctions de masse à valeurs dans 2^Ω. Nous proposons donc le modèle γ_i-EKNN suivant, compatible avec des labels riches et équivalent au modèle originel dans le cas de données labellisées parfaitement.

Modèle γ_i-EKNN : Pour maintenir l'équivalence avec le modèle introduit par Denœux 1995 quand il s'agit de données parfaitement labellisées, il est proposé d'utiliser un paramètre γ pour chaque voisin, en se référant à sa ressemblance. Autrement dit, chaque observation appartient à sa propre classe, elle-même plus ou moins ressemblante aux classes des autres observations. Quand il s'agit de données imparfaitement labellisées,

γ est calculé en relation avec la distance aux autres observations en prenant en compte la ressemblance (la distance de Jousselme est alors utilisée, voir équation (2.8)). Plus la labellisation est parfaite, plus le modèle se rapproche de l'original, avec un γ par classe :

$$\gamma_i = \frac{1}{\mathsf{d}_i^{\beta}}, \qquad \mathsf{d}_i = \frac{\sum_{\nu=1}^{N}\sum_{\mu=1}^{N}(1-d_J^{i,\nu})(1-d_J^{i,\mu})d^{\nu,\mu}}{[\sum_{\nu=1}^{N}(1-d_J^{i,\nu})]^2 - \sum_{\nu=1}^{N}(1-d_J^{i,\nu})^2}, \tag{4.8}$$

avec N le nombre total d'observations et $d_J^{i,\nu}$ la distance de Jousselme entre m_i et m_ν.

Ce modèle sera choisi pour étudier la pertinence de l'utilisation des données imparfaitement labellisées en comparant un modèle crédibiliste et non-crédibiliste, au coût de sa complexité calculatoire, car il maintient l'équivalence avec le modèle originel.

Équivalence dans le cas de la labellisation parfaite

L'équivalence entre le modèle présenté et le modèle original dans le cas de données parfaitement labellisées est prouvée dans cette partie. La masse normalisée présentée dans le document originel de Denœux 1995, notée $m_s^{\mathcal{D}}$, d'un élément focal ω_q, pour son modèle crédibiliste de K-NN, est développée comme suit :

$$m_s^{\mathcal{D}}(\{\omega_q\}) = \frac{[1 - \prod_{x^i\in\Phi_q^s}(1-\alpha)]\prod_{r\neq q}\prod_{x^i\in\Phi_r^s}(1-\alpha)}{\sum_{q=1}^{M}[(1-\prod_{x^i\in\Phi_q^s}(1-\alpha))\prod_{r\neq q}\prod_{x^i\in\Phi_r^s}(1-\alpha)] + \prod_{q=1}^{M}\prod_{x^i\in\Phi_q^s}(1-\alpha)}, \tag{4.9}$$

avec Φ_q^s l'ensemble des K plus proches voisins de x^s appartenant à ω_q et $\alpha = \alpha_0\varphi(d^{s,i})$ (pour des raisons de simplicité d'écriture).

Dans le modèle présenté ici, permettant le traitement de données imparfaitement labellisées, la masse sur l'un des singletons est :

$$m_s(\{\omega_q\}) = \frac{\sum_{B_1\cap\ldots\cap B_K=\omega_q}\prod_{x^i\in\Phi^s}\alpha m_i(B_i)}{\prod_{x^i\in\Phi^s}(1-\sum_{A\in 2^\Omega\setminus\Omega}\alpha m_i(A)) + \sum_{A\in 2^\Omega\setminus\Omega}\sum_{B_1\cap\ldots\cap B_K=A}\prod_{x^i\in\Phi^s}\alpha m_i(B_i)}. \tag{4.10}$$

Pour tout $\omega_q \in \Omega$, sous condition d'équivalence des φ (prouvée ci-après) et dans le cas

de fonctions de masse à supports simples [5] sur singletons, nous avons :

$$\begin{aligned}\prod_{x^i\in\Phi^s}(1-\sum_{A\in 2^\Omega\setminus\Omega}\alpha m_i(A)) &= \prod_{x^i\in\Phi^s}(1-\alpha)\\ &= \prod_{q=1}^{M}\prod_{x^i\in\Phi^s_q}(1-\alpha).\end{aligned} \tag{4.11}$$

Et :

$$\sum_{B_1\cap\ldots\cap B_K=\Omega_q}\prod_{x^i\in\Phi^s}\alpha m_i(B_i) = [1-\prod_{x^i\in\Phi^s_q}(1-\alpha)]\prod_{r\neq q}\prod_{x^i\in\Phi^s_r}(1-\alpha). \tag{4.12}$$

On a donc pour tout ω_q :

$$\begin{aligned}m_s(\{\omega_q\}) =& \frac{\sum_{B_1\cap\ldots\cap B_K=\omega_q}\prod_{x^i\in\Phi^s}\alpha m_i(B_i)}{\prod_{x^i\in\Phi^s}(1-\sum_{A\in 2^\Omega\setminus\Omega}\alpha m_i(A))+\sum_{A\in 2^\Omega\setminus\Omega}\sum_{B_1\cap\ldots\cap B_K=A}\prod_{x^i\in\Phi^s}\alpha m_i(B_i)}\\ =& \frac{[1-\prod_{x^i\in\Phi^s_q}(1-\alpha)]\prod_{r\neq q}\prod_{x^i\in\Phi^s_r}(1-\alpha)}{\prod_{x^i\in\Phi^s}(1-\sum_{A\in 2^\Omega\setminus\Omega}\alpha m_i(A))+\sum_{q=1}^{M}[[1-\prod_{x^i\in\Phi^s_q}(1-\alpha)]\prod_{r\neq q}\prod_{x^i\in\Phi^s_r}(1-\alpha)]}\\ =& \frac{[1-\prod_{x^i\in\Phi^s_q}(1-\alpha)]\prod_{r\neq q}\prod_{x^i\in\Phi^s_r}(1-\alpha)}{\sum_{q=1}^{M}[[1-\prod_{x^i\in\Phi^s_q}(1-\alpha)]\prod_{r\neq q}\prod_{x^i\in\Phi^s_r}(1-\alpha)]+\prod_{q=1}^{M}\prod_{x^i\in\Phi^s_q}(1-\alpha)}\\ =& m_s^{\mathcal{D}}(\{\omega_q\}).\end{aligned} \tag{4.13}$$

On a également pour la masse du cadre de discernement :

$$\begin{aligned}m_s(\Omega) &= 1-\sum_{q=1}^{M}m_s(\{\omega_q\})\\ &= 1-\sum_{q=1}^{M}m_s^{\mathcal{D}}(\{\omega_q\})\\ &= m_s^{\mathcal{D}}(\Omega).\end{aligned} \tag{4.14}$$

5. Le modèle oringinel n'utilise que des fonctions de masses à support simple.

Il reste à prouver l'équivalence entre les deux fonctions φ utilisées dans les équations (4.9) et (4.10). Cela revient à comparer les deux γ :

> Denœux utilise γ_q déterminé séparément pour chaque classe avec $1/\mathsf{d}_q^\beta$, où d_q est la distance moyenne entre deux observations appartenant à la classe ω_q.

$$\gamma_q = \frac{1}{\mathsf{d}_q^\beta}, \qquad \mathsf{d}_q = \frac{\sum_{\mu=1}^{N_q}\sum_{\nu=1}^{N_q} d^{\nu,\mu}}{N_q^2 - N_q} \tag{4.15}$$

Avec N_q le nombre d'observations appartenant à la classe ω_q. Dans le modèle présenté, nous avons :

$$\gamma_i = \frac{1}{\mathsf{d}_i^\beta}, \qquad \mathsf{d}_i = \frac{\sum_{\nu=1}^{N}\sum_{\mu=1}^{N}(1-d_J^{i,\nu})(1-d_J^{i,\mu})d^{\nu,\mu}}{[\sum_{\nu=1}^{N}(1-d_J^{i,\nu})]^2 - \sum_{\nu=1}^{N}(1-d_J^{i,\nu})^2}, \tag{4.16}$$

avec N le nombre total d'observations et d_J la distance de Jousselme entre deux masses.

Puisque les observations sont labellisées parfaitement, les masses sont catégoriques et dogmatiques. Le membre $(1-d_J^{i,\nu})(1-d_J^{i,\mu})d^{\nu,\mu}$ devient $d^{\nu,\mu}$ quand x^ν, x^μ et x^i sont de même classe, et zéro autrement. De même manière, le diviseur devient égal à $N_q^2 - N_q$.

Pour des données parfaitement labellisées, les γ sont égaux, la fonction φ est la même et il y a équivalence entre les deux modèles. Le théorème proposé par l'auteur dans le document originel est alors vérifié.

Théorème de Denœux

Si les K plus proches voisins d'un point x^s sont à la même distance de x^s, et si $\varphi_1 = \varphi_2 = \ldots = \varphi_C$. Alors la règle de décision produit la même décision que le vote majoritaire.

Expérimentations sur les différentes valeurs de γ

Le tableau 4.1 est une comparaison entre le modèle K-NN et certaines des approches liées à EK-NN présentées dans la partie précédente. Les expériences sont réalisées avec

Dataset	K-NN	γ_q-EKNN	γ-EKNN	γ_i-EKNN
Iris	**0.965** ± 0.006	0.963 ± 0.006	0.964 ± 0.006	0.963 ± 0.006
Wine	**0.737** ± 0.013	0.696 ± 0.012	0.704 ± 0.012	0.696 ± 0.012
Breast Cancer	0.927 ± 0.004	**0.928** ± 0.004	**0.928** ± 0.004	**0.928** ± 0.004
Credal Bird-10	0.383 ± 0.015	0.389 ± 0.014	0.411 ± 0.014	**0.412** ± 0.014

TABLEAU 4.1 – Exactitude moyenne (± intervalle de confiance à 95%) après 100 itérations sur jeux de données parfaitement labellisées (Iris, Wine, Breast Cancer) et sur jeu de données imparfaitement labellisées (Credal Bird-10).

7 plus proches voisins et le résultat est présenté au travers de l'exactitude moyenne [6] des modèles sur 100 itérations, auquel s'ajoute un intervalle de confiance [7] à 95%. La version pondérée par distances de K-NN est comparée avec la version originale γ_q-EKNN, la version avec un gamma unique γ-EKNN en utilisant $\gamma = 1/\mathsf{d}^\beta$ avec d la distance moyenne entre deux vecteurs d'entraînement, et la version γ_i-EKNN proposée ici. Les jeux de données sont séparés en deux, les jeux de données parfaitement labellisées (Iris, Wine et Breast Cancer) et le jeu de données imparfaitement labellisées (Credal Bird-10 public).

Comme on peut l'observer dans le tableau 4.1, il y a équivalence entre γ_q-EKNN et la version γ_i-EKNN proposée sur les jeux de données parfaitement labellisées (Iris, Wine et Breast Cancer). Quand il s'agit de données imparfaitement labellisées, ce même modèle γ_i-EKNN a des performances supérieures aux modèles non-crédibilistes et équivalentes aux modèles prenant en compte des données imparfaitement labellisées. Il n'est cependant pas possible d'affirmer à 95% de meilleures performances que le modèle avec un γ unique. L'objectif est cependant atteint car il ne s'agit pas d'avoir de meilleures performances, mais de représenter au mieux l'incertitude et l'imprécision du modèle et de prendre en compte l'imperfection présente dans les labels riches tout en gardant une équivalence.

4.2 Arbres de décision crédibilistes

Nous proposons dans cette partie un nouveau modèle d'arbre de décision, capable à la fois de faire une prédiction incertaine et imprécise, mais également d'utiliser l'information présente dans les labels riches. La notion de conflit au sein des fonctions de croyance est

6. On parle ici de l'exactitude moyenne du modèle et non de la précision, pour éviter toute confusion avec la précision liée à l'imperfection des données. Il s'agit de la proportion d'observations classées correctement par le modèle.

7. La formule utilisée est : $[\bar{x} - 1{,}96\frac{\mathcal{S}}{\sqrt{n}}; \bar{x} + 1{,}96\frac{\mathcal{S}}{\sqrt{n}}]$ avec $\bar{x}$ la moyenne de l'échantillon de taille n et $\mathcal{S}$ l'écart-type de la série de mesures.

utilisée pour regrouper des éléments de réponse dans un même nœud. Ce modèle s'avère plus robuste à l'imprécision que les modèles existants et offre de bonnes performances sur les jeux de données labellisés de manière incertaine et imprécise. Un rappel sur les arbres de décision est d'abord fait avant de présenter le modèle proposé.

4.2.1 Arbres de décision

En classification, les arbres de décision sont utilisés pour prédire la classe d'une observation grâce à une structure de nœuds et de divisions. Il s'agit d'une des approches les plus connues en apprentissage automatique. Parmi les différentes méthodes de classification, ils ont l'avantage d'être facilement compréhensibles, et leur interprétation est à la portée d'un plus grand nombre de personnes (voir Quinlan 1987). L'arbre se construit au préalable sur un jeu d'entraînement et les différentes divisions des nœuds définissent le chemin à suivre pour une prédiction. Un nœud contient des observations et peut avoir un nœud parent et au moins deux nœuds enfants [8]. Une division sépare un nœud en plusieurs nœuds enfants grâce aux différentes valeurs possibles d'un attribut. Pour diviser un nœud il convient de choisir l'attribut qui maximise une fonction de gain. Lors de la première étape, toutes les observations du jeu d'entraînement se trouvent dans un seul nœud, appelé racine de l'arbre.

L'efficacité de ces modèles est reconnue, ils sont simples à définir, ont une bonne interprétabilité et peuvent être utilisés en analyse exploratoire (*cf.* Siciliano 1998). Cependant, les arbres de décision sont sujets au sur-entraînement, qui se produit lorsqu'un processus d'apprentissage sur-optimise l'erreur sur l'ensemble d'apprentissage au détriment de la généralisation (voir Bramer 2013).

Les arbres de décision les plus connus sont C4.5, de Quinlan 1993 et CART, de Leo Breiman, Friedman et al. 1984. On parle alors d'arbres à induction descendante [9], ils sont définis par un critère de sélection utilisé pour définir le meilleur attribut pour une division, une stratégie de partitionnement qui divise un nœud en utilisant le critère de sélection, et plusieurs conditions d'arrêt pour stopper les divisions et faire du nœud enfant une feuille.

8. Le nombre de nœuds enfants pour un arbre dépend de l'architecture choisie.
9. *Top down induction decision trees*, en anglais.

Stratégie de partitionnement

L'attribut sélectionné pour diviser un nœud est celui qui maximise le gain suivant. Soit S un noeud qui contient une collection d'observations [10], et $\Omega = \{\omega_1, ..., \omega_C\}$ l'ensemble de toutes les classes possibles pour les éléments de S. Soit $\mathcal{A}$ un attribut [11] du domaine fini $\mathcal{D}_\mathcal{A}$. Le gain d'information $Gain(S, \mathcal{A})$ de la division de S en $\mathcal{A}$ est défini par Quinlan tel que :

$$Gain(S, \mathcal{A}) = Info(S) - Info_\mathcal{A}(S), \tag{4.17}$$

où $Info(S)$ est l'information du nœud S selon le critère de sélection, et $Info_\mathcal{A}(S)$ est la somme pondérée de l'information des nœuds issus de la division sur l'attribut $\mathcal{A}$:

$$Info_\mathcal{A}(S) = \sum_{v \in \mathcal{A}} \frac{|S_v|}{|S|} Info(S_v), \tag{4.18}$$

avec S_v le sous-ensemble de S d'éléments pour lesquels l'attribut $\mathcal{A}$ vaut v, autrement dit, le nœud enfant issu de la division sur $\mathcal{A}$ prenant comme valeur v. Une division est réalisée sur chaque nœud en partant de la racine et en maximisant le gain d'information, de manière récursive, jusqu'à ce qu'une condition d'arrêt soit rencontrée.

Critère de sélection

Maximiser le *Gain* revient à choisir le meilleur attribut $\mathcal{A}$ pour diviser S. Pour ce faire, Quinlan propose d'utiliser l'entropie de Shannon 1948 comme critère de sélection [12] :

$$Info(S) = -\sum_{\omega \in \Omega} p_\omega(S) \log_2 p_\omega(S), \tag{4.19}$$

avec $p_\omega(S)$ la proportion d'observations dans S appartenant à la classe ω. Le critère de Gini est aussi communément utilisé comme critère de sélection :

$$Info(S) = 1 - \sum_{\omega \in \Omega} p_\omega(S)^2. \tag{4.20}$$

Conditions d'arrêt

La construction de l'arbre est stoppée lorsque l'une de ces conditions est atteinte :

10. Si toutes les observations d'entraînement sont dans S, alors il s'agit de la racine de l'arbre.
11. Dans le chapitre 3 on parle de variables descriptives, il s'agit des attributs de l'observation.
12. Le logarithme en base 2 est utilisé pour le cas binaire.

— le nœud actuel ne compte qu'une seule observation.

— toutes les observations possèdent la même classe.

— les attributs restants ont un *Gain* inférieur ou égal à zéro.

L'élagage est une technique de compression qui vise à réduire la taille d'un arbre. Le pré-élagage [13] sera légèrement abordé dans ce chapitre, mais, lorsque cela n'est pas spécifié et par compatibilité avec les forêts aléatoires, tous les arbres seront considérés comme complètement développés.

Dans le cas de variables quantitatives, la méthode utilisée est la suivante ; pour un attribut $\mathcal{A}$, les observations sont triées par ordre croissant selon cet attribut et pour chaque paire consécutive c_1 et c_2, un seuil v est défini tel que $v = c_1 + (c_2 - c_1)/2$. Le seuil maximisant le gain d'information est utilisé pour diviser un nœud en deux nœuds enfants, l'un avec des valeurs strictement inférieures à v et l'autre avec des valeurs supérieures ou égales à v.

Prédiction

Une fois le processus de création de l'arbre terminé, les observations non labellisées vont parcourir l'arbre en partant de la racine et en fonction de leurs attributs. Quand une feuille est atteinte, l'observation se voit attribuer une probabilité d'appartenance à chaque classe en fonction des proportions représentées dans le nœud. La classe maximisant cette probabilité est la classe prédite ; par exemple, si une nouvelle observation atteint une feuille I, composée de 9 observations de classe ω_1 et 1 observation de classe ω_2, la probabilité prédite associée à ω_1 vaut 0.9 et celle associée à ω_2 vaut 0.1.

4.2.2 Motivation et arbres crédibilistes

Lorsque les observations ne sont plus labellisées avec des labels durs, mais avec des fonctions de masse, la proportion p_ω d'observations appartenant à la classe ω n'existe plus. Le *Gain* proposé par l'équation (4.17) n'est plus calculable, que ce soit avec l'entropie (4.19) ou avec le critère de Gini (4.20) puisqu'une observation n'est plus caractérisée par une classe, mais par une fonction de masse. Pour rendre à nouveau possible le calcul du gain, un autre critère d'information doit être utilisé. Cette partie en présente plusieurs.

13. L'élagage pendant la construction de l'arbre.

Approche par incertitude

Un nouvel arbre de décision est introduit par Denoeux et Bjanger 2000. Il utilise la théorie des fonctions de croyance et l'incertitude de Klir et Wierman 1998 pour regrouper des observations dans un même nœud en réduisant autant que possible l'incertitude présente dans le nœud. Deux informations, la non-spécificité et la discorde, respectivement données par les équations (2.16) et (2.17) sont utilisées pour calculer l'incertitude de Klir présente dans un nœud. Le critère d'information est défini par :

$$Info(S) = \lambda N(\bar{m}^S) + (1 - \lambda)D(\bar{m}^S), \tag{4.21}$$

avec un coefficient positif $\lambda = [0, 1]$ et une fonction de masse $\bar{m}^S$ associée au nœud S. Nous avons choisi de prendre $\bar{m}^S$ comme étant la fonction de masse moyenne des observations du nœud S. La valeur de $\bar{m}^S$ est alors la combinaison de chaque label riche [14] présent dans le nœud d'après la moyenne des masses. La règle de combinaison de Dempster, donnée par l'équation (2.12), n'est pas utilisée puisque certains jeux de données possèdent des labels issus de sources non indépendantes. Le paramètre λ agit comme un curseur sur le degré de non-spécificité et de discorde souhaité.

Cette méthode est ici nommée Uncertainty-EDT [15], avec $\lambda = 0.5$. Pour les expériences réalisées, plus de non-spécificité en critère (*i.e.* $\lambda > 0.5$) dégrade les performances du modèle tandis que plus de discorde (*i.e.* $\lambda < 0.5$) augmente les performances jusqu'à un seuil atteint en $\lambda = 0.5$.

Approche par distance euclidienne

Une autre approche, celle de Elouedi, Mellouli et al. 2001, propose de s'appuyer sur la distance euclidienne entre fonctions de masse [16]. L'objectif est alors de minimiser la distance intra masses dans les nœuds enfants. Les fonctions de masse proches au sens de la distance euclidienne sont regroupées pour former les divisions. La distance euclidienne $d(m_i, m_j)$ entre deux fonctions de masse m_i et m_j est définie comme suit :

$$d(m_i, m_j) = \sqrt{\sum_{A \subseteq \Omega} (m_i(A) - m_j(A))^2}. \tag{4.22}$$

14. C'est à dire labellisé avec une fonction de masse et non une classe.
15. *Uncertainty* pour *incertitude* et *EDT* pour *Evidential Decision Tree*.
16. Les auteurs proposent dans l'article d'utiliser un vecteur de correspondance au lieu d'utiliser directement les masses.

Le critère d'information $Info(S)$ pour le nœud S est donc la distance moyenne entre les fonctions de masse de toutes les observations du nœud S et ce modèle est ici nommé Euclidean-EDT. Pour rappel, les fonctions de masse sont les labels des observations et sont données dans le jeu de données.

Approche par distance de Jousselme

Dans la publication de Trabelsi, Elouedi et al. 2019, un arbre de décision utilisant la distance de Jousselme, Grenier et al. 2001 est introduit. Cette approche est similaire à celle de la distance euclidienne à l'exception de la distance utilisée. La distance de Jousselme considère les fonctions de masse comme des éléments de réponse et non comme de simples vecteurs. Le critère d'information $Info(S)$ pour le nœud S est donc la distance de Jousselme moyenne, calculée à partir de l'équation (2.8), entre les fonctions de masse de toutes les observations du nœud S. La méthode diffère de celle utilisant la distance euclidienne en assignant une distance plus faible entre deux éléments de réponse qui contiennent en partie la même information, ce modèle est nommé Jousselme-EDT.

Toutes ces méthodes ont l'avantage d'être compatibles avec des données imparfaitement labellisées, mais sont notamment vulnérables au surapprentissage [17]. Ceci conduit à notre proposition d'arbre de décision crédibiliste avec un critère plus robuste.

4.2.3 Conflit : nouveaux arbres de décision crédibilistes

Peu de modèles sont capables de prendre en compte l'incertitude et l'imprécision. Cette imperfection dans des labels plus riches peut être utilisée pour approximer ce qu'une source pense réellement, et même augmenter les performances d'un modèle (voir Hoarau, Martin, J.-C. Dubois et Le Gall 2022). Parmi les modèles d'arbres de décision crédibilistes présentés, deux propriétés peuvent être discutées. La première est que le degré de discorde, donné par l'équation (2.17), peut être supérieur à zéro pour une seule réponse d'un seul utilisateur, et ne peut donc pas représenter au mieux la contradiction entre plusieurs réponses (*i.e.* $D(\bar{m}^S) \geq 0$ avec $|S| = 1$). La seconde propriété est que pour l'utilisation des distances, l'impureté et donc l'information dans un nœud, est non nulle pour des éléments de réponses différents, même s'ils sont inclus les uns dans les autres (*i.e.* $m_i \neq m_j \iff d(m_i, m_j) \neq 0$). Pour aborder ces deux problématiques, nous utilisons la notion de conflit dans la théorie des fonctions de croyance.

17. L'effet de surapprentissage est montré dans la partie 4.2.4.

Approche par conflit

Nous proposons d'utiliser une mesure de conflit comme critère d'information. Ce conflit, composé d'un degré d'inclusion et d'une distance est introduit par Martin 2019. Deux inclusions sont définies, l'inclusion stricte, définie par l'équation (2.19), qui inclut une masse dans une autre lorsque tous les éléments focaux de l'une sont inclus un à un dans chaque élément focal de l'autre et l'inclusion légère, définie par l'équation (2.20), qui inclut une masse dans une autre lorsque tous les éléments focaux de l'une sont inclus dans au moins un élément focal de l'autre.

Nous proposons et introduisons $\delta^{i\subseteq j}(m_i, m_j)$, un degré d'inclusion moins stricte que celui défini à l'équation (2.19), sans prendre en compte l'inclusion sur l'ignorance.

Inclusion souple

Une fonction de masse m_i est incluse dans m_j si tous les éléments focaux de m_i sur $2^\Omega \backslash \Omega$ sont inclus un à un dans chaque élément focal de m_j sur $2^\Omega \backslash \Omega$.

Le degré d'inclusion souple $\delta^{i\subseteq j}(m_i, m_j)$ est donné comme suit :

$$\delta^{i\subseteq j}(m_i, m_j) = \frac{1}{|\mathcal{L}_i||\mathcal{L}_j|} \sum_{A\in\mathcal{L}_i} \sum_{B\in\mathcal{L}_j} Inc(A, B), \tag{4.23}$$

avec $\mathcal{L}_i$ et $\mathcal{L}_j$ respectivement la collection d'éléments focaux sur $2^\Omega \backslash \Omega$ de m_i et m_j. Cette équation est utilisée au lieu de l'inclusion stricte, car l'ignorance, selon l'inclusion stricte, est uniquement incluse dans elle-même.

Exemple : Soit y_1, y_2 et y_3 des labels riches avec $\Omega = \{chien,\ chat,\ oiseau\}$, tels que :

y_1 **:** $m_1(\{chien\}) = 1$, *"C'est un chien"*.

y_2 **:** $m_2(\{chat\}) = 1$, *"C'est un chat"*.

y_3 **:** $m_3(\{chat,\ oiseau\}) = 1$, *"C'est un chat ou un oiseau"*.

Les degrés d'inclusion souple pour différents couples de réponses sont : $\delta^{1\subseteq 1}(y_1, y_1) = 1$, $\delta^{1\subseteq 2}(y_1, y_2) = 0$, $\delta^{1\subseteq 3}(y_1, y_3) = 0$, $\delta^{2\subseteq 3}(y_2, y_3) = 1$ et $\delta^{3\subseteq 2}(y_3, y_2) = 0$.

Calcul du critère d'information

Grâce au degré d'inclusion souple, on introduit $\delta(m_i, m_j)$, un degré d'inclusion de m_i et m_j où l'ordre n'a plus d'importance. Il est défini par :

$$\delta(m_i, m_j) = \max(\delta^{i \subseteq j}(m_i, m_j), \delta^{j \subseteq i}(m_j, m_i)). \tag{4.24}$$

La mesure de conflit $\mathcal{C}(m_i, m_j)$, utilisée pour calculer le critère d'information dans le modèle proposé est défini comme suit :

$$\mathcal{C}(m_i, m_j) = (1 - \delta(m_i, m_j)) d_J(m_i, m_j), \tag{4.25}$$

avec $d_J(m_j, m_i)$ la distance de Jousselme, donnée par l'équation (2.8), entre m_i et m_j. Le calcul de $\mathcal{C}$ donne le conflit, et l'information $Info(S)$ est alors le conflit moyen deux à deux dans le nœud S :

$$Info(S) = \frac{\sum_{x_i \in S} \sum_{x_j \in S} \mathcal{C}(m_i, m_j)}{|S|^2 - |S|}. \tag{4.26}$$

Le gain est calculé identiquement aux arbres de décision avec l'équation (4.17). Cette approche diffère de Jousselme-EDT en autorisant deux observations à appartenir au même nœud, sans perte de gain, si une réponse est incluse dans l'autre. Le modèle proposé, utilisant la notion de conflit, est nommé Conflict-EDT.

Prédiction

Une fois l'arbre construit, une nouvelle observation traverse l'arbre en partant de la racine et d'après la valeur de ses attributs. Quand une feuille est atteinte, l'observation se voit attribuer une fonction de masse égale à la moyenne [18] des fonctions de masse présentes dans le nœud. Par exemple, si une nouvelle observation atteint une feuille composée de 10 éléments de masses $\{m_1, \ldots, m_{10}\}$, alors la fonction de masse prédite est $\bar{m}$ la fonction de masse moyenne des 10 fonctions de masse. La classe maximisant la probabilité pignistique, calculée avec l'équation (2.15), de cette fonction de masse est la classe prédite.

Exemple : Soit $\Omega = \{\omega_1, \omega_2\}$ l'ensemble des classes possibles dans un problème de classification. Soit S une feuille atteinte par l'observation x et composée de deux fonctions

18. D'autres combinaisons ont été étudiées mais n'ont pas été aussi performantes et satisfaisantes que la moyenne.

de masse m_1 et m_2 telles que :

m_1 : $m_1(\{\omega_1\}) = 0.3$, $m_1(\Omega) = 0.7$,

m_2 : $m_2(\{\omega_1\}) = 0.2$, $m_2(\Omega) = 0.8$.

La fonction de masse $\bar{m}$ pour le nœud S est le label riche prédit pour x par l'arbre de décision crédibiliste, on a : $\bar{m}(\{\omega_1\}) = 0.25$ et $\bar{m}(\Omega) = 0.75$. Au niveau décisionnel, la classe maximisant la probabilité pignistique est ω_1 et elle est choisie comme label dur.

Robustesse au surapprentissage

Quand on manipule des observations parfaitement labellisées et des arbres de décision classiques, plusieurs observations possèdent la même classe et l'arbre arrête de grandir lorsque toutes les observations d'un nœud sont de même classe. Il n'y a plus de meilleure division possible, réduisant l'impureté du nœud et augmentant le *Gain*.

Ici, on s'intéresse aux données imparfaitement labellisées avec la théorie des fonctions de croyance, presque aucune observation n'a un label identique. L'arbre de décision a donc plus de mal à s'arrêter de grandir, formant un arbre surentraîné qui a plus de mal à généraliser. Avec le modèle proposé Conflict-EDT, un nœud de plusieurs fonctions de masse peut avoir un conflit nul si les fonctions de masse sont incluses les unes dans les autres. Cela permet, comme pour le cas des données parfaitement labellisées, d'avoir un *Gain* égal à zéro et d'arrêter la pousse de l'arbre.

Exemple : Soit $\Omega = \{\omega_1, \omega_2\}$ un cadre de discernement. Soit trois observations, x_1, x_2 et x_3 respectivement labellisées par m_1, m_2 et m_3 faisant partie d'un nœud racine tel que :

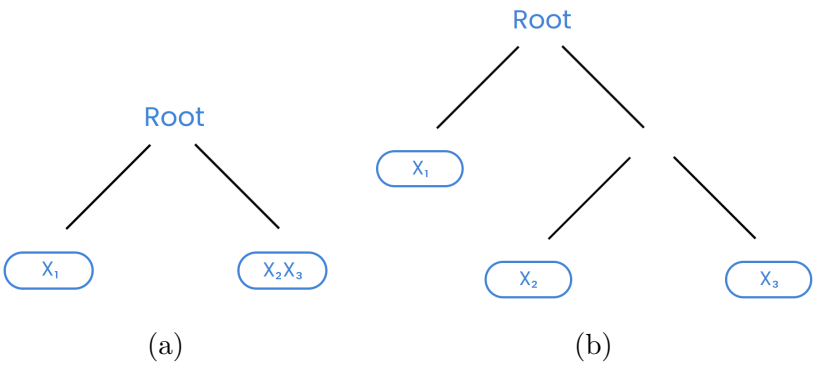
m_1 : $m_1(\omega_1) = 0.9$, $m_1(\Omega) = 0.1$,

m_2 : $m_2(\omega_2) = 0.8$, $m_2(\Omega) = 0.2$,

m_3 : $m_3(\omega_2) = 0.9$, $m_3(\Omega) = 0.1$.

La fonction de masse m_1 peut être vue comme une croyance forte que l'observation x_1 appartient à ω_1. Les deux fonctions de masse m_2 et m_3 peuvent être vues comme des croyances fortes que x_2 et x_3 appartiennent à ω_2 avec une plus faible ignorance pour x_3.

En utilisant le modèle Uncertainty-EDT, et s'il y a suffisamment d'attributs pour séparer les nœuds, la discorde calculée en combinant les masses m_2 et m_3 sera non nulle, créant ainsi l'arbre le plus profond possible, visible sur la figure 4.2b. La même propriété est vérifiée pour Euclidean-EDT et Jousselme-EDT, chacune des deux distances utilisées entre m_2 et m_3 est non nulle, formant le même arbre (voir figure 4.2b). Cependant, avec



(a) (b)

FIGURE 4.2 – Un arbre de décision moins développé (a) et un arbre de décision plus profond (b) pour trois observations x_1, x_2 et x_3.

Conflict-EDT, m_2 est incluse dans m_3 et le conflit d'équation (4.25) entre m_2 et m_3 est nul. Un arbre plus petit avec des feuilles plus grosses est alors créé, visible sur la figure 4.2a.

Pour résumer, les autres modèles crédibilistes ne regroupent pas x_2 et x_3 dans le même nœud à cause de leur label différent, tandis que le modèle proposé les regroupe parce que les réponses sont incluses l'une dans l'autre.

4.2.4 Expériences

Dans cette section, des expériences de comparaison avec le modèle proposé d'arbre de décision crédibiliste sont présentées. Pour toutes les situations, l'objectif est de montrer la robustesse du modèle à l'imperfection, à la fois sur données bruitées et sur des jeux de données qui ont été réellement labellisés de manière incertaine et imprécise.

Lorsque cela n'est pas précisé, chaque expérience est réalisée 100 fois pour obtenir une estimation de l'exactitude du modèle pour chaque jeu de données. Une itération correspond à un tirage aléatoire de 20% du jeu de données comme jeu de test, le reste est utilisé pour l'entraînement. La plupart des jeux de données utilisés ici sont disponibles sur UCI Machine Learning Repository (*cf.* Dua et Graff 2017). Les autres jeux de données sont effectivement labellisés de manière incertaine et imprécise par des contributeurs. De tels jeux de données où les utilisateurs ont eu la possibilité de représenter une imperfection ne sont pas nombreux. Nous avons utilisé ceux présentés à la section 3.2. Les détails de ces jeux de données sont présentés dans le tableau 4.2.

On s'intéresse ici aux observations labellisées avec incertitude et imprécision en utilisant la théorie des fonctions de croyance et certains des jeux de données utilisés sont de cette nature, les autres sont bruités comme suit.

TABLEAU 4.2 – Description des jeux de données, avec le nombre total d'observations, le nombre de classes et le nombre de caractéristiques.

Dataset	Observations	Classes	Caractéristiques
Breast cancer	569	2	30
Ionosphere	351	2	34
Post-operative	86	2	8
Sonar	208	2	60
Liver	345	2	6
Balance scale	625	3	4
Iris	150	3	4
Wine	178	3	13
Glass	214	6	9
Ecoli	336	8	7
Credal Dog-2	200	2	42
Credal Dog-4	400	4	47
Credal Dog-7	700	7	43
Credal Bird-2	40	2	17
Credal Bird-10	200	10	30

Bruit par imprécision

Une observation est choisie aléatoirement et le label correspondant perd un degré de précision (*i.e.* la cardinalité du sous-ensemble contenant la réponse augmente de 1), avec une autre classe choisie aléatoirement sur Ω. Par exemple, dans le cas du jeu de données Iris de Fisher, si une source labellise une observation *Virginica*, alors le label bruité devient soit *Virginica ou Setosa* soit *Virginica ou Versicolor*. Un jeu de données bruité à 50% veut dire que la moitié des labels ont perdu un degré de précision. Pour les modèles non crédibilistes, le label dur est celui maximisant la probabilité pignistique, calculée à partir de l'équation (2.15).

Performances globales

La première partie se concentre sur l'intérêt global de la méthode. Le tableau 4.3 présente les performances des arbres de décision, des modèles Euclidean-EDT, Uncertainty-EDT, Jousselme-EDT et du modèle proposé Conflict-EDT. Les jeux de données utilisés possèdent des labels riches et les dix premiers sont bruités à 50%, la moitié de chaque jeu

TABLEAU 4.3 – Exactitude moyenne sur jeux bruités à 50% (± un intervalle de confiance à 95% pour l'estimation de la moyenne). Un arbre de décision (DT) et quatre arbres de décision crédibilistes (EDT) sont utilisés pour comparaison, le modèle proposé est Conflict (la significativité d'un t-test de Welch à la p-valeur < 0.05 est indiquée par un *).

Dataset	DT	EDT			
		Euclidean	Uncertainty	Jousselme	Conflict
Breast cancer	70.1 ± 0.9	72.4 ± 0.8	71.4 ± 0.7	72.8 ± 0.9	**91.1*** ± 0.5
Ionosphere	67.6 ± 1.2	71.0 ± 0.9	68.4 ± 1.1	71.9 ± 1.0	**87.4*** ± 0.8
Post-operative	55.7 ± 2.4	**60.4** ± 2.4	56.7 ± 2.2	57.8 ± 2.4	59.9 ± 2.3
Sonar	61.1 ± 1.6	59.5 ± 1.5	60.2 ± 1.6	59.2 ± 1.4	**66.8*** ± 1.4
Liver	53.8 ± 1.2	55.4 ± 1.2	54.8 ± 1.1	56.0 ± 1.2	**58.0*** ± 1.1
Balance scale	59.4 ± 1.0	69.8 ± 0.8	57.5 ± 0.9	69.9 ± 0.7	**75.1*** ± 0.6
Iris	63.6 ± 2.0	74.0 ± 1.8	68.5 ± 1.7	73.9 ± 1.7	**90.4*** ± 1.2
Wine	70.4 ± 1.9	70.1 ± 1.4	68.3 ± 1.6	68.3 ± 1.7	**88.1*** ± 1.3
Glass	50.7 ± 1.5	51.4 ± 1.5	51.8 ± 1.5	53.1 ± 1.6	**60.5*** ± 1.5
Ecoli	56.8 ± 1.3	58.8 ± 1.1	57.1 ± 1.2	59.0 ± 1.2	**69.9*** ± 1.0
Credal Dog-2	82.9 ± 1.3	81.2 ± 1.4	82.4 ± 1.2	81.8 ± 1.3	**83.0** ± 1.2
Credal Dog-4	57.7 ± 1.3	58.0 ± 1.2	57.7 ± 1.0	58.2 ± 1.2	**59.2** ± 1.1
Credal Dog-7	50.1 ± 0.9	50.1 ± 0.9	48.8 ± 0.9	50.0 ± 0.8	**53.1*** ± 1.0
Credal Bird-2	50.8 ± 3.6	**62.8** ± 3.7	57.3 ± 3.7	59.8 ± 3.3	52.4 ± 3.4
Credal Bird-10	42.0 ± 1.6	43.1 ± 1.5	45.0 ± 1.7	42.7 ± 1.6	**45.4** ± 1.6

de données a donc perdu un degré de précision. Les jeux de données *Credal* n'ont pas eu besoin d'être bruités, puisqu'ils possèdent de manière inhérente des labels riches.

Le modèle proposé obtient de meilleurs résultats à la fois sur des données bruitées et sur des jeux de données imparfaitement labellisés. Conflict-EDT a de meilleures performances, et parfois avec un écart important, comme sur les jeux de données Breast cancer, Iris ou Wine, exception faite pour les jeux de données Post-operative et Credal Bird-2. L'explication pour cet écart de performance est donnée dans l'expérience suivante.

Robustesse au bruit

Cette partie s'intéresse particulièrement à la robustesse à l'imprécision et à l'évolution des performances du modèle avec l'augmentation de bruit. La figure 4.3 présente les résultats de cette expérience sur un bruit par imprécision. Les jeux de données Iris (Figure 4.3a), Wine (Figure 4.3b), Glass Identification (Figure 4.3c), Balance Scale (Figure 4.3d), Ionosphere (Figure 4.3e) et Ecoli (Figure 4.3f) sont bruités de 0% à 100% et l'exactitude moyenne des modèles est présentée.

Sur les jeux de données Iris, Balance Scale et Ionosphere, les modèles utilisant une distance, Euclidean-EDT, Jousselme-EDT et Conflict-EDT, ont de meilleures performances que les arbres de décision et Uncertainty-EDT. Cependant, parmi ces modèles utilisant

des distances, le modèle proposé Conflict-EDT est le plus robuste au bruit avec environ 90% de bonnes prédictions sur le jeu de données Iris bruité de moitié, contre 75% de bonnes prédictions pour le second meilleur modèle. Sur les trois autres jeux de données Wine, Glass Identification et Ecoli, Euclidean-EDT et Jousselme-EDT ne montrent pas de meilleures performances que les modèles qui n'utilisent pas de distances, mais Conflict-EDT est toujours aussi performant comparé à tous les autres modèles présentés.

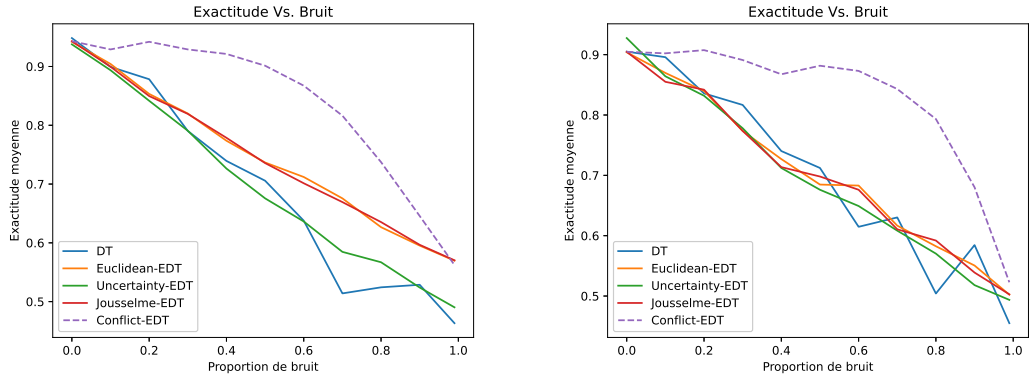
Pour un bruit par imprécision, le modèle proposé Conflict-EDT obtient de meilleurs résultats sur tous les jeux de données présentés et à tous niveaux de bruit. Le modèle n'est en fait pas plus robuste à l'imprécision que les autres modèles crédibilistes, mais au surapprentissage, grâce à l'équation (4.25) qui permet de regrouper plus d'observations dans un même nœud en utilisant la notion de conflit. Une autre expérience a montré qu'en pré-élaguant tous les arbres, les performances des autres modèles augmentent et l'écart de performance avec le modèle proposé disparaît. Un exemple est présenté en figure 4.4, il s'agit de la même expérience que la précédente pour le jeu de données Balance Scale, mais cette fois les modèles ont été élagués.

Le pré-élagage empêche l'arbre de grandir trop profondément durant la phase d'apprentissage. Le nombre d'observations est limité et l'arbre ne peut pas créer de feuilles avec moins de 5 éléments. Ici, les modèles Uncertainty-EDT, Euclidean-EDT, Jousselme-EDT et Conflict-EDT sont tous robustes au bruit par imprécision. Conflict-EDT n'est plus le modèle le plus performant puisque son avantage est de limiter le surapprentissage. L'élagage protège les arbres de cet effet de surapprentissage, et la différence de performances n'est plus notable. Seul l'arbre de décision classique a de très faibles performances, parce que les données sont labellisées avec imprécision, une information qui n'est pas prise en compte par ce modèle non crédibiliste.

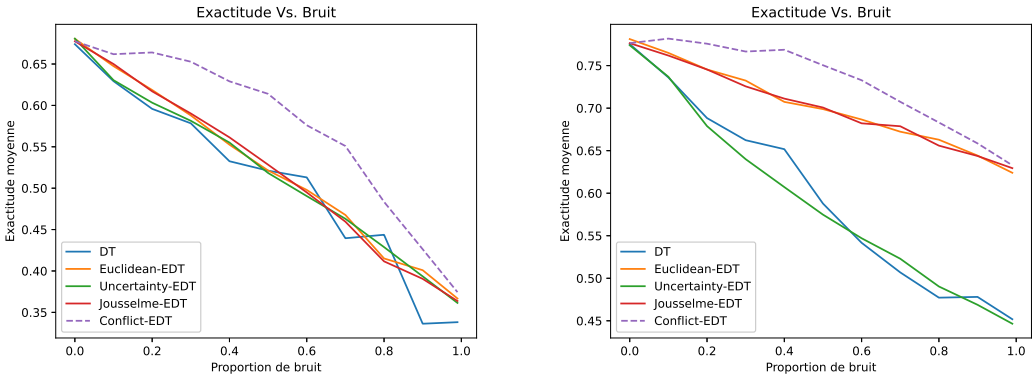
En l'absence d'élagage, Conflict-EDT offre une meilleure robustesse au surapprentissage quand les données sont labellisées avec imprécision.

Pousse de l'arbre

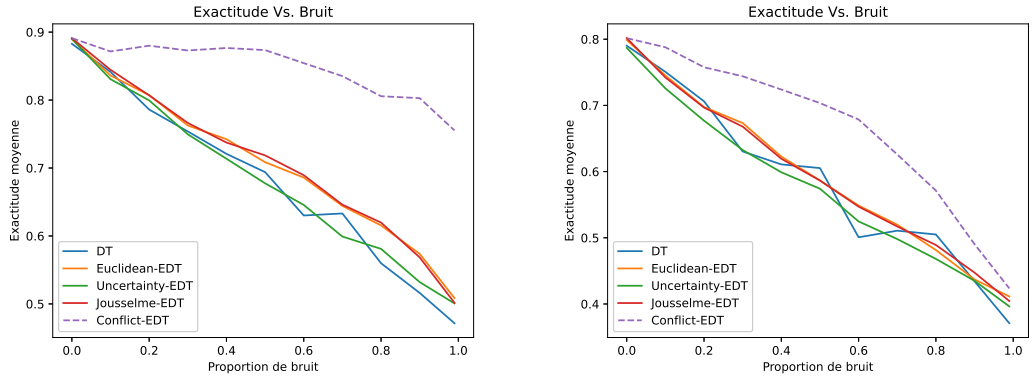
Dans l'expérience précédente, la robustesse du modèle au surapprentissage a été déduite de ses performances. Dans cette expérience, deux critères additionnels sont présentés pour montrer les bénéfices de la méthode proposée. Avec des données labellisées de manière incertaine et imprécise, les modèles d'arbres de décision crédibilistes ont tendance à sur apprendre sur le jeu d'entraînement et à délivrer de moins bonnes performances sur le jeu de test. Il en résulte de larges arbres avec de petites feuilles. La *profondeur* de l'arbre



(a) Iris (b) Wine



(c) Glass Identification (d) Balance Scale



(e) Ionosphere (f) Ecoli

FIGURE 4.3 – Exactitude moyenne par niveau de bruit sur plusieurs jeux de données, pour un arbre de décision (DT) et des arbres de décision crédibilistes (EDT).

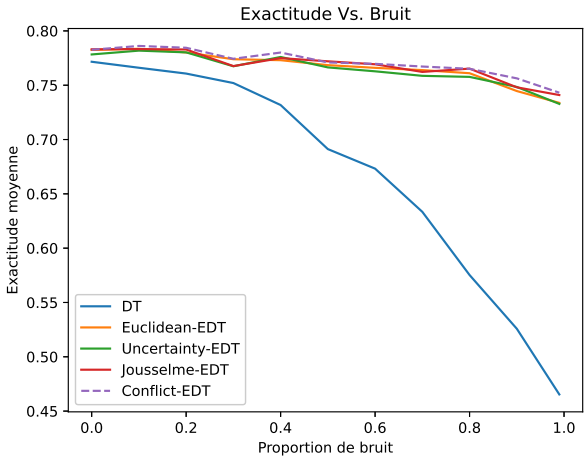


FIGURE 4.4 – Exactitude moyenne en fonction du bruit sur des modèles pré-élagués et pour le jeu de données Balance Scale.

correspond au nombre maximum de divisions entre la racine de l'arbre et une feuille. Plus l'arbre est profond, plus il est entraîné sur les données. La *taille des feuilles* est le nombre moyen d'observations présentes dans une feuille de l'arbre. Plus les feuilles sont grandes, plus l'arbre peut généraliser.

Le tableau 4.4 présente la profondeur moyenne de chaque arbre (non élagué) de décision crédibiliste étudié. Les dix premiers jeux de données sont arbitrairement bruités à 30%. Comme vu dans l'expérience précédente, n'importe quelle valeur de bruit peut être utilisée pour montrer la pertinence d'utiliser le modèle proposé. Le bruit par imprécision est utilisé et les valeurs présentées sont des moyennes sur 50 expériences, arrondies à l'unité. Les jeux de données Credal n'ont pas été bruités puisqu'ils ont déjà été labellisés de manière incertaine et imprécise par des contributeurs. Pour les jeux de données Iris et Wine, Euclidean-EDT, Uncertainty-EDT et Jousselme-EDT ont une profondeur moyenne entre 12 et 18 alors que Conflict-EDT a une profondeur moyenne de 8. Cette différence est présente pour tous les jeux de données et particulièrement notable sur les jeux de données avec un grand nombre de classes.

Le tableau 4.5 représente la taille moyenne des feuilles en suivant les mêmes spécifications. La tendance est la même pour l'ensemble des jeux de données, Conflict-EDT obtient de plus grosses feuilles et l'écart semble être plus important pour les jeux de données à deux classes.

De manière générale, Conflict-EDT a grandi moins profondément et a obtenu de plus

TABLEAU 4.4 – Profondeur moyenne des arbres arrondie à l'unité, avec un bruit à 30% (excepté pour les jeux de données Credal), le modèle proposé est Conflict.

Dataset	Euclidean	Uncertainty	Jousselme	Conflict
Breast cancer	19	33	19	**7**
Ionosphere	17	24	18	**10**
Post-operative	11	13	11	**9**
Sonar	12	17	12	**7**
Liver	18	18	17	**13**
Balance scale	13	14	13	**12**
Iris	13	14	13	**8**
Wine	12	18	13	**8**
Glass	21	16	20	**14**
Ecoli	26	19	26	**17**
Credal Dog-2	15	36	15	**9**
Credal Dog-4	25	41	23	**20**
Credal Dog-7	48	42	36	**24**
Credal Bird-2	12	18	11	**6**
Credal Bird-10	26	24	20	**14**

TABLEAU 4.5 – Nombre d'observations moyen dans les feuilles, arrondi à l'unité et sur des jeux de données bruités à 30% (excepté pour Credal), le modèle proposé est Conflict.

Dataset	Euclidean	Uncertainty	Jousselme	Conflict
Breast cancer	5	3	5	**36**
Ionosphere	5	2	5	**18**
Post-operative	3	2	3	**5**
Sonar	5	3	5	**12**
Liver	3	2	3	**6**
Balance scale	3	1	3	**5**
Iris	3	2	3	**7**
Wine	4	2	4	**9**
Glass	2	2	2	**3**
Ecoli	2	2	2	**3**
Credal Dog-2	2	1	2	**8**
Credal Dog-4	1	1	1	**2**
Credal Dog-7	1	1	1	**2**
Credal Bird-2	1	1	1	**5**
Credal Bird-10	1	1	1	**2**

larges feuilles que les autres modèles d'arbres de décision crédibilistes. Cette robustesse au surapprentissage sur des données incertaines et imprécises permet d'obtenir de meilleurs résultats de généralisation.

4.3 Forêts aléatoires crédibilistes

Un nouveau modèle de forêt aléatoire est introduit, pouvant produire des prédictions incertaines et imprécises et utilisant l'information présente dans les labels riches. Ce modèle s'appuie sur les arbres de décision crédibilistes introduits précédemment. Un rappel sur les forêts aléatoires est fait avant de détailler la méthode et de présenter les expériences réalisées (voir Hoarau, Martin, J.-C. Dubois et Le Gall 2023).

4.3.1 Forêts aléatoires

Introduit par L. Breiman 1996, le *bagging*[19] est la première étape vers les forêts aléatoires, définies plus tard par Leo Breiman 2001 en y ajoutant une sélection aléatoire de variables. Le principe est de combler la faiblesse des arbres de décision, la variance élevée, en combinant les prédictions d'un nombre important d'arbres, la forêt. Cette section explique les opérations de *bagging* et de sélection aléatoire de variables, toutes deux utilisées dans la version la plus répandue des forêts aléatoires de Leo Breiman 2001.

Bagging

Cette définition est largement reprise de la publication de L. Breiman 1996 où le bagging sur des arbres de décision est introduit. Soit $\Omega = \{\omega_1, \ldots, \omega_M\}$ une collection de M différentes classes et $\mathcal{L} = \{x_k|1, \ldots, K\}$ un jeu d'entraînement de K observations où chaque élément est associé à un label $y_k \in \Omega$. Étant donné un estimateur $\varphi(x, \mathcal{L})$, l'objectif est de créer une séquence $\{\mathcal{L}_n\}$ de nouveaux jeux d'entraînement pour améliorer les performances de l'estimateur sur le jeu unique. Une nouvelle séquence d'estimateurs est alors introduite $\{\varphi(x, \mathcal{L}_n)\}$. Le bagging fait référence au ***b**ootstrap **agg**regat**ing*** où le bootstrap permet de créer les $\{\mathcal{L}_n\}$ jeux d'entraînement, le résultat est alors donné suite à une agrégation des estimateurs $\{\varphi(x, \mathcal{L}_n)\}$.

19. L'anglicisme *bagging* sera ici adopté pour plus de simplicité de compréhension.

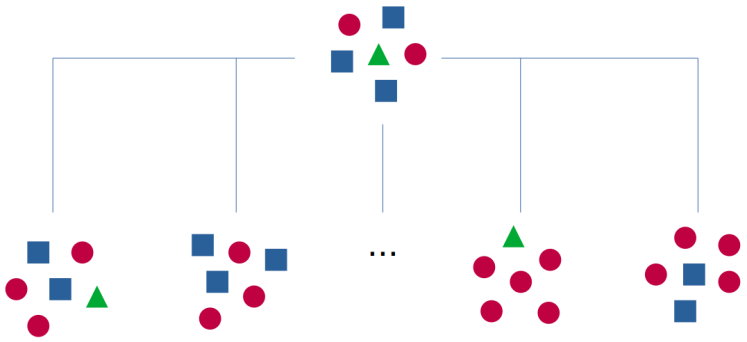


FIGURE 4.5 – Bootstrap. Le jeu d'entraînement est divisé en N jeux d'entraînement de bootstrap.

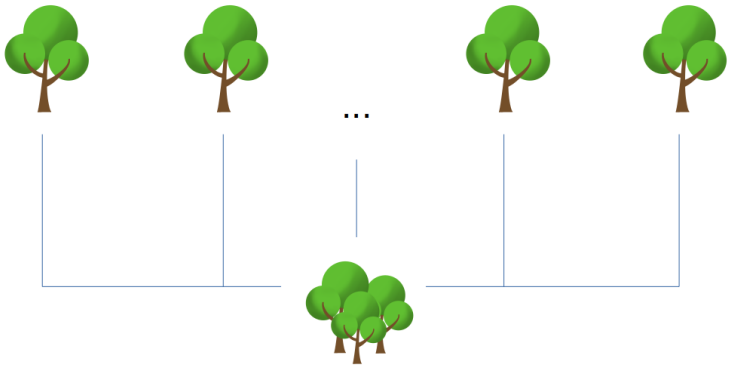


FIGURE 4.6 – Aggregating. Après entraînement sur chaque ensemble, les prédictions des estimateurs sont agrégées par vote. Le résultat est la prédiction de la forêt aléatoire.

Bootstrap : Les $\{\mathcal{L}_n\}$ jeux d'entraînement sont composés de K éléments, en bagging la taille des ensembles $\{\mathcal{L}_n\}$ est identique à $\mathcal{L}$. Chaque élément est tiré aléatoirement avec remise, dans $\mathcal{L}$. Ce qui veut dire qu'une observation x_k peut apparaître plusieurs fois ou aucune dans n'importe quel sous-ensemble $\mathcal{L}_n$. Ce processus est présenté sur la figure 4.5.

Aggregating : Lorsque $\varphi(x, \mathcal{L})$ prédit une classe (voir la publication de L. Breiman 1996 pour des prédictions continues), une méthode d'agrégation des $\{\varphi(x, \mathcal{L}_n)\}$ prédictions est le vote majoritaire, ou un vote pondéré. Les résultats des N arbres sont alors agrégés, pour former la prédiction de l'unique modèle de forêt aléatoire, comme présenté sur la figure 4.6.

Sélection aléatoire de variables

Inspiré par Amit et Geman 1997, Leo Breiman 2001 introduit la sélection aléatoire de variables pour les forêts aléatoires. Le premier objectif est d'être compétitif avec le modèle Adaboost. Le principe est d'utiliser uniquement un nombre réduit de variables, choisies aléatoirement à chaque division lors de l'entraînement de l'arbre de décision. Pour les modèles présentés, $\sqrt{p}$ variables sont aléatoirement sélectionnées à chaque division, avec p le nombre de variables et la dimension du vecteur x. Lors d'une division, l'arbre de décision utilisé dans les forêts aléatoires a uniquement accès à un nombre réduit de variables pour définir sa meilleure division.

4.3.2 Nouvelle forêt aléatoire crédibiliste

Les arbres de décision crédibilistes souffrent des mêmes problèmes que les arbres de décision classiques, il s'agit notamment d'un modèle à variance élevée (voir L. Breiman 1996). Nous proposons une forêt aléatoire crédibiliste, qui utilise les arbres de décision crédibilistes et qui tire avantage à la fois des performances élevées sur données imparfaitement labellisées et de l'augmentation de performance due à la réduction de variance.

Le bagging et la sélection de variables aléatoires sont utilisés sur les arbres de décision crédibilistes pour modéliser les forêts aléatoires crédibilistes.

L'étape de *bootstrap* dans le modèle proposé implique des labels riches : soit $\Omega = \{\omega_1, \ldots, \omega_M\}$ une collection de M différentes classes et $\mathcal{L} = \{x_k|1, \ldots, K\}$ un ensemble d'entraînement de K observations où chaque élément est associé à une fonction de masse m_k sur 2^Ω interprétée comme un label. Soit l'estimateur $\varphi(x, \mathcal{L})$ la forêt aléatoire crédibiliste et $\{\varphi(x, \mathcal{L}_n)\}$ la séquence d'arbres de décision crédibilistes. Les $\{\mathcal{L}_n\}$ ensembles sont les nouveaux ensembles bootstrap d'apprentissage de K éléments tirés aléatoirement avec remplacement dans $\mathcal{L}$. Le nombre N (avec $n \in N$) d'ensembles de bootstrap est discuté dans les expériences et fixé à 50.

Lors de l'agrégation, chaque estimateur $\varphi(x, \mathcal{L}_n)$ prédit une fonction de masse m_n pour une observation non labellisée. Nous proposons de combiner toutes les fonctions de masse par moyenne :

$$\bar{m}(A) = \frac{1}{N} \sum_{n=1}^{N} m_n(A), \quad A \in 2^\Omega, \tag{4.27}$$

avec $\bar{m}$ la prédiction de la forêt aléatoire crédibiliste.

Prise de décision

La motivation est de représenter une fonction de masse dans la sortie du modèle ERF (*Evidential Random Forest*), mais une décision peut être prise sur l'ensemble des classes Ω. La classe ω_{ERF} maximisant la probabilité pignistique, calculée à partir de l'équation (2.15), de $\bar{m}$ est la classe prédite :

$$\omega_{ERF} = \underset{\omega \in \Omega}{\operatorname{argmax}}(BetP(\omega,\ \bar{m})), \tag{4.28}$$

avec $BetP(\omega,\ m)$ la probabilité pignistique de ω d'après m.

Exemple : Soit $\Omega = \{\omega_1, \omega_2\}$ l'ensemble des classes possibles pour un problème de classification. Soit $\varphi(x, \mathcal{L})$ une forêt aléatoire crédibiliste de $N = 3$ estimateurs, des arbres de décision crédibilistes, prédisant respectivement m_1, m_2 et m_3 pour une nouvelle observation non labellisée telle que :

m_1 **:** $m_1(\omega_1) = 0.8$, $m_1(\Omega) = 0.2$,

m_2 **:** $m_2(\omega_1) = 0.9$, $m_2(\Omega) = 0.1$,

m_3 **:** $m_2(\omega_2) = 0.2$, $m_2(\Omega) = 0.8$.

Les estimateurs $\varphi(x, \mathcal{L}_1)$ et $\varphi(x, \mathcal{L}_2)$ supportent fortement la classe ω_1 tandis que $\varphi(x, \mathcal{L}_3)$ supporte très légèrement ω_2. La prédiction $\bar{m}$ pour la forêt aléatoire crédibiliste est alors :

$\bar{m}$ **:** $\bar{m}(\omega_1) = 0.57$, $\bar{m}(\omega_2) = 0.07$, $\bar{m}(\Omega) = 0.36$.

Au niveau de la prise de décision, on calcule les probabilités pignistiques pour les deux classes :

$BetP(\omega_1) = 0.75$,

$BetP(\omega_2) = 0.25$.

La classe ω_1 est choisie comme label dur, car elle maximise la probabilité pignistique sur $\bar{m}$. C'est donc la prédiction finale du modèle.

4.3.3 Expériences

Cette partie compare le modèle proposé de forêt aléatoire crédibiliste avec d'autres modèles et montre l'amélioration apportée aux arbres de décision crédibilistes. Les forêts aléatoires et forêts aléatoires crédibilistes sont entraînées avec 50 estimateurs, c'est-à-dire 50 arbres de décision entraînés sur des jeux de bootstrap. L. Breiman 1996 montre que

les performances du modèle atteignent un sommet aux alentours de 25 estimateurs, et n'augmentent plus au-delà de 50. Nous obtenons des résultats similaires, avec un arrêt de l'augmentation des performances au-delà de 30 arbres de décision. Une sélection aléatoire de variables est également réalisée à chaque nœud et quand cela n'est pas précisé, les paramètres des modèles sont ceux par défaut présent dans la bibliothèque scikit-learn (*cf.* Pedregosa, Varoquaux et al. 2011).

Les spécifications des expériences sont les mêmes que pour les arbres de décision crédibilistes. Chaque expérience est réalisée 100 fois pour obtenir une estimation de l'exactitude moyenne du modèle pour tous les jeux de données. Une itération correspond à un tirage aléatoire de 20% du jeu de données comme jeu de test, le reste est utilisé pour l'entraînement. Les détails concernant les jeux de données sont présents dans le tableau 4.2. Le bruit utilisé est le bruit par imprécision introduit dans la partie précédente 4.2.4.

Gain en performance sur les arbres de décision crédibilistes

Les forêts aléatoires sont des modèles à faible variance, définis pour réduire l'erreur des arbres de décision. Cette expérience propose de calculer le gain apporté par les forêts aléatoires crédibilistes proposées comparées aux arbres de décision crédibilistes, le modèle utilisé dans le processus de création de la forêt. Ce gain en performance [20] est présenté dans le tableau 4.6, l'augmentation brute en performance ainsi que le gain en pourcentage sont donnés.

Les moyennes sont estimées sur 100 itérations et le bruit utilisé est celui sur l'imprécision à hauteur de 50%, sauf pour les jeux de données Credal qui sont déjà labellisés de manière incertaine et imprécise. Le modèle augmente en effet les performances des arbres de décision crédibilistes en regroupant les prédictions de multiples estimateurs. Sur des jeux de données avec peu de classes (Iris, Wine, Balance, Breast cancer, Inosphere) les forêts aléatoires crédibilistes augmentent moins, mais toujours significativement, les performances des arbres de décision crédibilistes. Quand le nombre de classes augmente (Glass, Ecoli, Credal Dog-7, Credal Bird-10) le gain en performance devient très important, avec une augmentation de plus de 20%. Quand les jeux de données sont imparfaitement labellisés, sans utiliser de bruit, et sur un nombre de classes important (Credal Bird-10 et Credal Dog-7), le gain en performance est le plus impressionnant (avec une augmentation respective de 35% et de 49%).

20. Le gain correspond à la différence entre les exactitudes moyennes des deux modèles.

TABLEAU 4.6 – Gain en performance des forêts aléatoires crédibilistes comparées aux arbres de décision crédibilistes sur jeux de données bruités et labellisés imparfaitement. La différence d'exactitude entre les deux modèles, ainsi que le pourcentage gagné sont présentés.

Dataset	Gain d'exactitude	Gain en pourcentage
Breast cancer	3.5	3.8
Ionosphere	5.2	6.0
Post-operative	11.1	18.6
Sonar	10.0	15.0
Liver	8.5	14.7
Balance scale	9.5	12.7
Iris	5.3	5.8
Wine	9.5	10.8
Glass	14.5	23.9
Ecoli	15.6	22.3
Credal Dog-2	10.8	13.1
Credal Dog-4	18.0	30.4
Credal Dog-7	26.1	49.2
Credal Bird-2	0.2	0.5
Credal Bird-10	15.8	34.9

Comparaison avec les forêts aléatoires

Le modèle proposé réduit la variance élevée des arbres de décision crédibilistes, comme prévu théoriquement. Cependant, il reste à montrer que les performances des forêts aléatoires crédibilistes sont compétitives avec les forêts aléatoires. Les exactitudes des deux modèles sont comparées dans cette expérience, pour démontrer les bénéfices de la proposition. Les exactitudes moyennes, avec les intervalles de confiance, sont présentées dans le tableau 4.7.

Pour tous les jeux de données, bruités ou labellisés de manière incertaine et imprécise, les performances des forêts aléatoires sont significativement améliorées par les forêts aléatoires crédibilistes proposées. Cette augmentation atteint jusqu'à 10 points pour les jeux de données Post-operative, Iris ou encore Balance scale. Quand l'imperfection liée à une labellisation humaine est représentée dans les données, un gain d'exactitude est également notable.

Autres forêts prudentes et modèles crédibilistes

Dans cette partie, un modèle récent de forêt aléatoire prudente introduit par H. Zhang, Quost et al. 2023, ainsi qu'un autre modèle crédibiliste, les K plus proches voisins crédibilistes introduit par Denœux 1995, sont utilisés pour comparaison.

TABLEAU 4.7 – Exactitude moyenne pour les forêts aléatoires (RF) et les forêts aléatoires crédibilistes (ERF) sur jeux de données bruités à 50% (± un intervalle de confiance à 95% pour l'estimation de la moyenne, la significativité d'un t-test de Welch à la p-valeur $<$ 0.05 est indiquée par un *).

Dataset	RF	ERF
Breast cancer	90.5 ± 0.5	**94.5*** ± 0.4
Ionosphere	84.4 ± 1.0	**92.6*** ± 0.6
Post-operative	60.5 ± 2.3	**71.0*** ± 2.0
Sonar	72.0 ± 1.3	**76.8*** ± 1.2
Liver	58.2 ± 1.2	**66.5*** ± 0.9
Balance scale	75.1 ± 0.6	**84.5*** ± 0.5
Iris	84.4 ± 1.3	**95.3*** ± 0.7
Wine	91.5 ± 1.0	**97.5*** ± 0.5
Glass	68.2 ± 1.5	**75.1*** ± 1.3
Ecoli	77.6 ± 0.9	**85.5*** ± 0.8
Credal Dog-2	91.4 ± 1.0	**93.8*** ± 0.9
Credal Dog-4	72.3 ± 1.0	**77.1*** ± 0.9
Credal Dog-7	77.4 ± 0.7	**79.1*** ± 0.8
Credal Bird-2	45.0 ± 3.2	**52.6*** ± 3.2
Credal Bird-10	52.8 ± 1.4	**61.2*** ± 1.5

Les forêts aléatoires prudentes utilisent le modèle de Dirichlet imprécis au niveau de la prédiction des estimateurs et la théorie des fonctions de croyance de Dempster 1967 et de Shafer 1976 pour l'agrégation. Ce modèle possède donc des similarités avec le modèle proposé de forêts aléatoires crédibilistes. Il est autorisé à faire des prédictions prudentes en combinant des intervalles de probabilités. Cependant, ce modèle n'est défini que pour des jeux de données à deux classes. Chaque arbre de décision dans la forêt produit un intervalle de probabilité pour la classe positive, grâce au modèle de Dirichlet imprécis. La forêt aléatoire prudente prédit ensuite la classe d'une nouvelle observation soit précisément [21] ou de manière imprécise, sur $\Omega = \{\omega_1, \omega_2\}$. Nous avons utilisé un calcul d'exactitude classique pour les prédictions précises (la proportion de prédictions correctes), et pris la réponse la plus plausible (en utilisant le score de plausibilité défini par les auteurs) quand le modèle retourne une prédiction prudente. Pour s'adapter à la capacité du modèle à donner des réponses prudentes, les auteurs proposent d'utiliser le critère u_{65} qui récompense une prédiction imprécise sur $\{\omega_1, \omega_2\}$ par 0.65 (au lieu de 1 pour une prédiction précise).

Les K plus proches voisins crédibilistes ont été introduits plus tôt dans ce chapitre, il s'agit d'une version des K plus proches voisins capable à la fois de prendre en compte

21. Une prédiction précise est réalisée sur un singleton, ω_1 ou ω_2 ici dans le cas de problèmes à deux classes.

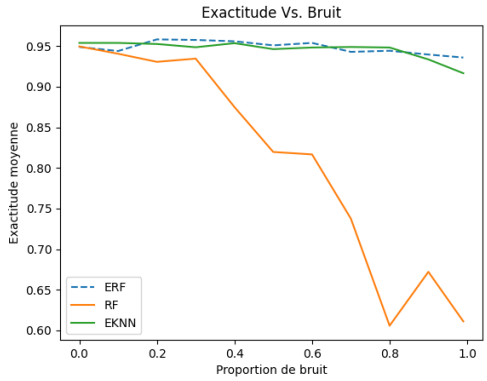
TABLEAU 4.8 – Score moyen d'exactitude pour les forêts aléatoires prudentes (CRF), les K plus proches voisins crédibilistes (EK-NN) et les forêts aléatoires crédibilistes proposées (ERF) sur jeux de données bruités à 50%, excepté pour les jeux de données Credal (± un intervalle de confiance à 95%). Le score u_{65} est aussi présent pour CRF et la significativité d'un t-test de Welch à la p-valeur < 0.05 est indiqué par un *.

Dataset	CRF		EK-NN	ERF
	Acc	u_{65}	*Acc*	*Acc*
Breast cancer	90.4 ± 0.5	91.7 ± 0.5	**95.4*** ± 0.4	94.5 ± 0.5
Ionosphere	84.3 ± 0.9	84.3 ± 0.8	83.2 ± 0.9	**92.6*** ± 0.6
Post-operative	62.9 ± 2.0	59.2 ± 1.9	**71.7** ± 2.0	71.0 ± 2.0
Sonar	72.4 ± 1.3	74.4 ± 1.2	**77.5** ± 1.3	76.8 ± 1.2
Liver	57.6 ± 1.2	61.2 ± 1.0	60.8 ± 1.1	**66.5*** ± 0.9
Balance scale			**88.2*** ± 0.5	84.5 ± 0.5
Iris			94.6 ± 0.8	**95.3** ± 0.7
Wine			95.9 ± 0.7	**97.5*** ± 0.5
Glass			64.2 ± 1.5	**75.1*** ± 1.3
Ecoli			84.8 ± 0.8	**85.5** ± 0.8
Credal Dog-2	91.1 ± 1.0	92.8 ± 0.8	73.8 ± 1.5	**93.8** ± 0.9
Credal Dog-4			69.3 ± 1.0	**77.1*** ± 0.9
Credal Dog-7			75.8 ± 0.7	**79.1*** ± 0.8
Credal Bird-2	45.3 ± 3.1	47.3 ± 3.1	**58.9*** ± 3.4	52.6 ± 3.2
Credal Bird-10			60.6 ± 1.5	**61.2** ± 1.5

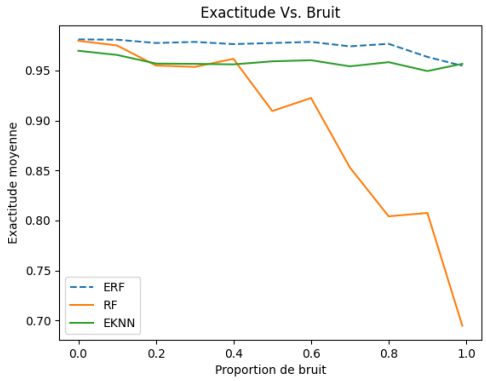
des labels riches et de produire une prédiction incertaine et imprécise en utilisant la théorie des fonctions de croyance. Ce modèle est devenu une référence dans l'apprentissage crédibiliste. Le meilleur nombre K de voisins est estimé grâce à une validation croisée à 5 plis et les paramètres utilisés sont ceux définis pour la version γ-EKNN présentée par Hoarau, Martin, J.-C. Dubois et Le Gall 2022 et plus tôt dans ce document.

Le tableau 4.8 présente l'exactitude moyenne pour les forêts aléatoires prudentes, pour les K plus proches voisins crédibilistes et pour le modèle proposé des forêts aléatoires crédibilistes. Le score u_{65} est aussi présent pour les forêts aléatoires prudentes pour récompenser les réponses prudentes du modèle. Également, ce modèle n'est compatible qu'avec des jeux de données à 2 classes, d'où l'absence de résultats dans le tableau pour les jeux de données avec plus de deux classes. Comme pour les autres expériences, les 10 premiers jeux de données sont bruités et les jeux de données Credal possèdent directement des labels riches.

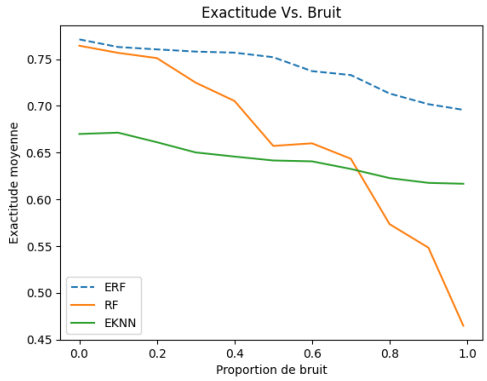
La capacité du modèle de forêts aléatoires prudentes à émettre des prédictions prudentes ne permet pas de compenser la baisse de performance comparée aux modèles crédibilistes et liée aux labels riches. Le modèle des K plus proches voisins crédibilistes et les forêts aléatoires crédiblistes proposées bénéficient tous les deux des labels riches et sont



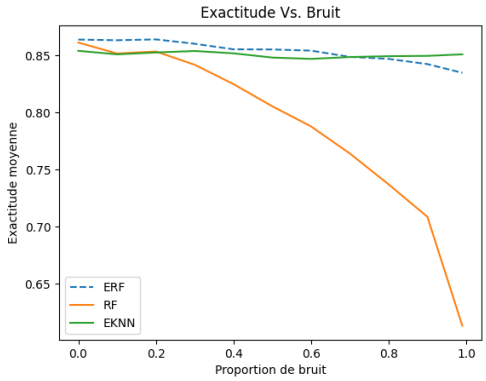
(a) Iris



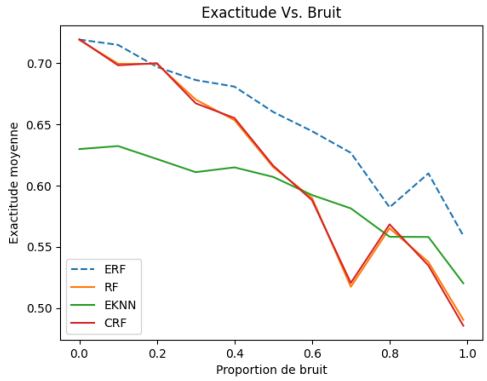
(b) Wine



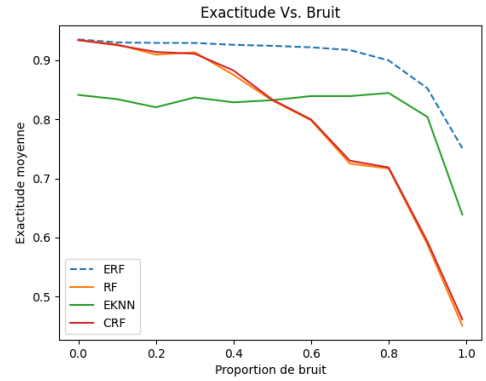
(c) Glass Identification



(d) Ecoli



(e) Liver



(f) Ionosphere

FIGURE 4.7 – Exactitude moyenne des modèles sur plusieurs jeux de données en fonction du bruit, pour les forêts aléatoires (RF), les forêts aléatoires prudentes (CRF), les K plus proches voisins crédibilistes (EKNN) et le modèle proposé des forêts aléatoires crédibilistes (ERF).

de ce fait plus compétitif. Sur les 10 jeux de données bruités, les forêts aléatoires crédibilistes offrent de meilleures performances pour 6 jeux de données. Sur les jeux de données Credal, le modèle proposé obtient de meilleurs résultats (sauf pour Credal Bird-2).

La figure 4.7 montre la robustesse des modèles étudiés au bruit. Les forêts aléatoires semblent toujours commencer avec un score d'exactitude élevé, quand il n'y a pas de bruit, et chutent en performances très rapidement avec l'augmentation du bruit. Les forêts aléatoires prudentes, évaluées ici sur l'exactitude (et non le critère u_{65}) donnent des résultats très proches des forêts aléatoires (sur les jeux de données à 2 classes). Les modèles crédibilistes (Les K plus proches voisins crédibilistes et les forêts aléatoires crédibilistes) évoluent identiquement. Les deux modèles commencent à un niveau de performances différent sur des données non bruitées, dépendamment de leur affinité avec le jeu de données, mais obtiennent tous deux des performances qui déclinent très lentement avec l'ajout de bruit. Pour conclure, les performances du modèle proposé sont meilleures sur les jeux de données présentés.

4.4 Conclusion du chapitre

Une nouvelle version des K plus proches voisins crédibilistes a été présentée, permettant la classification d'observations labellisées imparfaitement. Une proposition a été faite quant au paramètre γ au travers de γ_i-EKNN permettant de retrouver une équivalence avec le modèle original. Cette équivalence a été démontrée théoriquement puis appuyée expérimentalement. Les résultats de l'expérience ne nous permettent cependant pas d'affirmer de meilleures performances que l'état de l'art sur des données imparfaitement labellisées.

Un nouveau modèle d'arbres de décision crédibiliste a aussi été introduit, utilisant une mesure de conflit. Ce critère permet de regrouper dans le même nœud les observations avec des éléments de réponse inclus les uns dans les autres. Il en résulte des arbres moins profonds avec un risque plus faible de surentraînement.

Enfin, un modèle de forêt aléatoire crédibiliste a été proposé, permettant de combler la variance élevée des arbres de décision crédibilistes. Le modèle a été comparé avec un autre modèle récent et avec la référence en classification crédibiliste. Les résultats montrent des performances compétitives à la fois sur des données bruitées et sur des jeux de données qui ont été labellisés de manière incertaine et imprécise.

La motivation sous-jacente n'est pas tant de travailler sur les performances du modèle,

mais d'étudier le processus de labellisation et de comprendre si en ajoutant de l'information à ce niveau et en représentant au mieux la connaissance, de meilleurs résultats peuvent être obtenus. Nous essayons de combler le manque de modèles qui sont à la fois capables de prendre en compte des labels riches et faire une prédiction modélisée par la théorie des fonctions de croyance. Ce travail s'inscrit donc dans une première étape introduisant le prochain chapitre traitant de ces mêmes données en utilisant l'apprentissage actif pour réduire les coûts de labellisation.

Chapitre 5

Apprentissage actif sur données imparfaites

Après avoir travaillé sur la qualité des labels, d'abord lors de la phase de labellisation, puis en classification, il convient de travailler sur la quantité de labels. La méthode d'apprentissage actif a été choisie, car elle permet de limiter le nombre de données à labelliser, mais également parce que la capacité du modèle à retranscrire une incertitude peut s'avérer utile lors de l'échantillonnage. Coupler la théorie des fonctions de croyance à l'apprentissage actif permet de répondre à une problématique où les données sont labellisées imparfaitement, mais également où le nombre de labellisations est limité.

Le travail de recherche réalisé pour travailler à la fois sur la qualité et la quantité de labels est présenté dans ce chapitre, d'abord en s'intéressant à l'échantillonnage par incertitude, très utilisé en apprentissage actif, puis en le couplant avec les travaux présentés dans les chapitres précédents. Une comparaison entre différentes versions des K plus proches voisins crédibilistes est également réalisée. Ensuite, avec l'objectif de réduire encore plus le nombre de données labellisées en utilisant la nouvelle information d'incertitude et d'imprécision, de nouvelles méthodes d'échantillonnages sont envisagées au travers d'entropies crédibilistes.

Publications

— Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2022), « Imperfect Labels with Belief Functions for Active Learning », in : *Belief Functions : Theory and Applications*, p. 44-53

— Arthur Hoarau, Vincent Lemaire, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2024a), « Evidential uncertainties and rich labels for active learning », in : *Machine Learning*

5.1 Apprentissage actif et labels riches

Cette première partie étudie les propositions faites dans les chapitres précédents en y ajoutant un échantillonnage par incertitude en apprentissage actif, c'est-à-dire aux jeux de données possédant des labels riches et aux modèles crédibilistes présentés, en y appliquant un apprentissage actif, et en se limitant à l'état de l'art sur ce sujet. L'objectif est de travailler sur des jeux de données labellisés de manière incertaine et imprécise avec des modèles capables de prendre en compte cette information, tout en réduisant les coûts de labellisation. L'apprentissage actif est un bon candidat puisqu'il permet de limiter le nombre d'observations qui vont être labellisées.

5.1.1 Pertinence des modèles crédibilistes sur labels riches

Dans cette expérience, les jeux de données Iris et Wine ont été bruités pour moitié. Le bruit est souvent utilisé pour simuler l'imperfection des données, cette expérience est donc notre première étape de comparaison entre un modèle crédibiliste (EK-NN) [1] et sa version non-crédibiliste (K-NN).

Bruitage des données

L'utilisation de bruit est souvent représentée dans la littérature. La première des expériences liée à l'apprentissage actif a donc été réalisée sur des jeux de données connus, auxquels du bruit a été ajouté sur les labels. Le même bruit qu'au chapitre 4 est utilisé. Pour rappel : une observation est choisie aléatoirement et le label correspondant perd un degré de précision (*i.e.* la cardinalité du sous-ensemble contenant la réponse augmente de 1), avec une autre classe choisie aléatoirement sur Ω. Un jeu de données bruité à 50% veut dire que la moitié des labels ont perdu un degré de précision.

La répartition d'entraînement et de test est faite aléatoirement, avec 20% du jeu total utilisé comme jeu de test et les 80% restants comme jeu d'entraînement. Les comparaisons se font principalement entre K-NN et EK-NN, d'autres modèles sont également ajoutés à titre informatif. La version de K-NN utilisée est celle présentée dans la section 4.1.1 avec

1. Les modèles EK-NN et K-NN ont été introduits dans la partie 4.1.

une pondération sur la distance et en utilisant 7 plus proches voisins [2]. Sont également utilisées, une Régression Logistique avec l'algorithme Newton-cg (voir Fletcher 2000) pour l'optimisation et une Forêt Aléatoire. Les différents modèles utilisent les paramètres par défaut de la bibliothèque scikit-learn, mise à disposition par Pedregosa, Varoquaux et al. 2011.

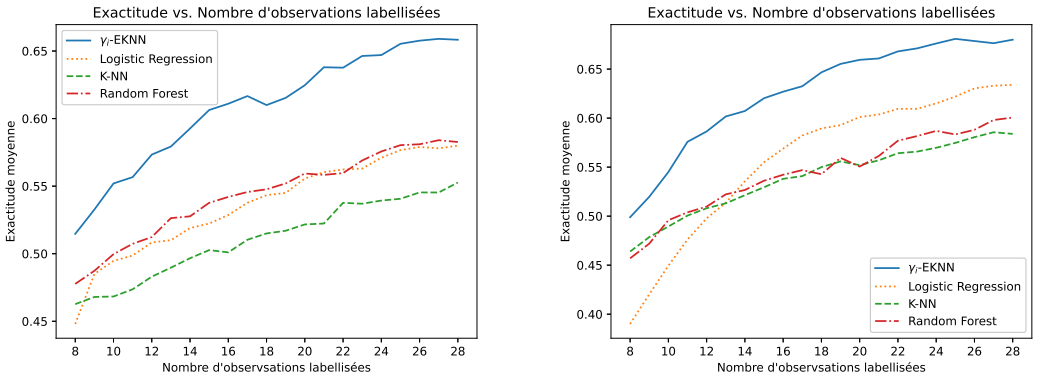
L'expérience est réalisée 100 fois et la moyenne des exactitudes est présentée pour chaque modèle. La figure 5.1 représente l'évolution de l'exactitude moyenne de γ_i-EKNN, d'une Régression Logistique, de K-NN et d'une Forêt Aléatoire en fonction du nombre de requêtes labellisées par apprentissage actif. Les performances de γ_i-EKNN atteignent environ 0.65 d'exactitude moyenne sur le jeu de données Iris, avec seulement 28 observations labellisées. Cela représente un gain de 20% de performance sur K-NN, dû à une meilleure prise en compte du bruit permettant de moins altérer le véritable label. La distance entre les exactitudes moyennes de γ_i-EKNN et K-NN augmente également avec le nombre de requêtes, ce qui signifie que l'étape d'échantillonnage sélectionne de meilleures instances ou des instances permettant d'augmenter l'exactitude du modèle plus rapidement. On comprend ici la pertinence d'utiliser de l'apprentissage actif, et, malgré le fait que les deux modèles utilisent le même échantillonnage par incertitude (2.23), γ_i-EKNN obtient de meilleurs résultats. Les résultats sur le jeu Wine sont moins optimistes, mais montrent toujours une dominance de γ_i-EKNN sur sa version non-crédibiliste.

L'interprétation de ces résultats est cependant biaisée par le bruit, car même si la distribution du bruit est identique, les labels utilisés pour les classifieurs non-crédibilistes ne contiennent pas la même information que ceux utilisés pour l'apprentissage de γ_i-EKNN. Afin de rendre une comparaison possible, il est proposé dans les expériences suivantes de ne pas s'intéresser au bruit mais d'utiliser des jeux de données labellisées imparfaitement. L'objectif est de comprendre si l'ajout d'information, durant la phase de labellisation, permet d'obtenir de meilleurs résultats lors de la phase d'apprentissage.

5.1.2 Expérience sur jeux de données labellisées imparfaitement

Jusqu'à présent, seuls des jeux de données parfaitement labellisées ont été utilisés avec de l'apprentissage actif dans ce chapitre. Dans cette partie, une procédure est présentée permettant cette fois-ci de comparer des modèles crédibilistes et non-crédibilistes,

2. Ce nombre est choisi car il donne de bons résultats à la fois pour K-NN et EK-NN, une validation croisée pour choisir K n'est pas toujours optimale en apprentissage actif, puisque la taille du jeu de données est évolutive.



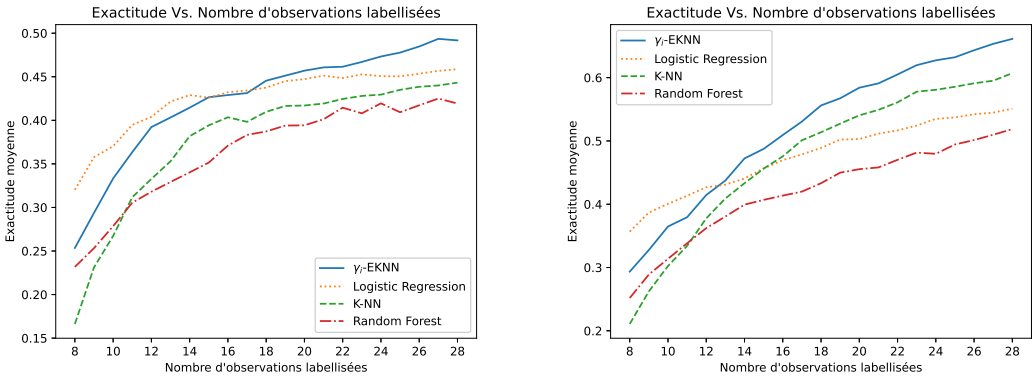
(a) (b)

FIGURE 5.1 – Exactitude moyenne des modèles en fonction du nombre d'observations labellisées avec 50% de bruit, 8 observations sont labellisées aléatoirement et 20 requêtes sont faites. Jeux de données *Iris* (*a*) et *Wine* (*b*).

à l'aide de données fondamentalement imparfaites et non bruitées. En utilisant les jeux de données présentés dans la section 3.2 avec le vecteur large de 512 variables, EK-NN est comparé à sa version non crédibiliste. Il s'agit d'une expérience similaire à celle réalisée précédemment, la différence réside dans les jeux de données, imparfaitement labellisés par des contributeurs. La figure 5.2 représente l'évolution de l'exactitude moyenne (sur 100 expériences) de γ_i-EKNN, d'une Régression Logistique, de K-NN et d'une Forêt Aléatoire en fonction du nombre de requêtes labellisées par apprentissage actif. Les performances de γ_i-EKNN atteignent environ 0.48 d'exactitude moyenne sur le jeu de données *Credal Bird-10*, avec seulement 28 observations labellisées, contre respectivement 0.44 d'exactitude moyenne pour K-NN. On note donc une augmentation des performances grâce à l'utilisation de la version crédibiliste des K plus proches voisins.

5.1.3 Expérience de comparaison entre différentes versions de EK-NN

Cette expérience a pour objectif de comparer différentes versions de EK-NN présentées dans la section 4.1.3. Les modèles γ_i-EKNN, γ-EKNN, γ_q-EKNN, K-NN sont comparés selon les mêmes modalités que l'expérience précédente. La figure 5.3 représente l'évolution de l'exactitude moyenne (sur 100 expériences) de ces modèles en fonction du nombre de requêtes labellisées par apprentissage actif. Les résultats montrent que les versions de EK-NN prenant en compte des données labellisées imparfaitement ont des résultats très



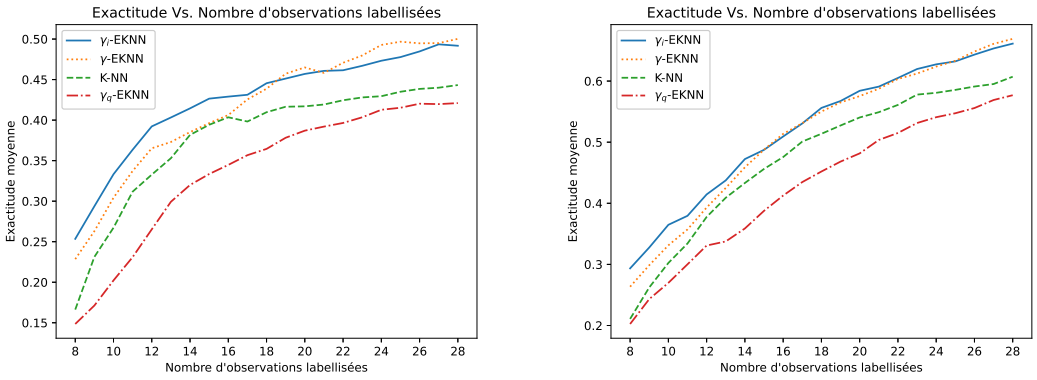
(a) (b)

FIGURE 5.2 – Exactitude moyenne des modèles en fonction du nombre d'observations labellisées, 8 observations sont labellisées aléatoirement et 20 requêtes sont faites. Jeux de données *Credal Bird-10* (*a*) et *Credal Dog-7* (*b*).

similaires (γ_i-EKNN et γ-EKNN). La version γ_q-EKNN (avec un paramètre γ_q par classe) a quant à elle des résultats plus proches de K-NN, puisqu'elle aussi utilise des labels durs.

5.1.4 Bilan des premières expériences

Ces premières expériences s'inscrivent dans l'état de l'art de l'apprentissage actif sur données imparfaitement labellisées. Cependant, les résultats présentés n'apportent pas de nouveauté fondamentale. Les modèles crédibilistes qui ont été présentés au chapitre précédent ont démontré de meilleures performances sur des labels riches, qu'ils soient bruités ou labellisés imparfaitement par des contributeurs. En observant de meilleurs résultats sur les jeux de données complets, on pouvait s'attendre à retrouver de meilleurs résultats sur une fraction du jeu de données, et c'est ce qui est observé ici sur un nombre très faible de données labellisées, ne dépassant pas 30 observations. Les modèles ont été comparés, mais les méthodes d'apprentissage actif utilisées sont celles qui ont été mises en place pour des modèles classiques, dont l'échantillonnage par incertitude qui est le plus répandu. Maintenant que les jeux de données possèdent des labels riches et que certains modèles sont capables de les interpréter, il serait pertinent de travailler en apprentissage actif à réduire d'autant plus les coûts de labellisation. Il est envisageable à ce niveau d'utiliser l'information qui a été ajoutée en sortie des modèles pour améliorer l'échantillonnage et sélectionner les observations qui vont permettre au modèle de gagner en performance rapidement.



(a) (b)

FIGURE 5.3 – Comparaison de différentes versions de EK-NN avec l'exactitude moyenne des modèles en fonction du nombre d'observations labellisées, 8 observations sont labellisées aléatoirement pour 20 requêtes. Jeux de données *Credal Bird-10* (a) et *Credal Dog-7* (b).

5.2 Utilisation d'entropies crédibilistes

La première tentative pour améliorer l'échantillonnage en apprentissage actif est de s'intéresser aux entropies crédibilistes. Si l'échantillonnage par incertitude utilise le plus souvent une mesure de confiance, donnée par l'équation (2.23), ou l'entropie de Shannon, donnée par l'équation (2.24), d'autres entropies ont été définies avec la théorie des fonctions de croyance. Ces entropies peuvent prendre en compte une information supplémentaire sur l'incertitude et l'imprécision de la prédiction du modèle, maintenant représentée par une fonction de masse.

Entropies crédibilistes

Les entropies comparées lors d'échantillonnages par incertitude dans les expériences suivantes sont présentées ici, elles sont listées par Zhu, Martin et al. 2021 et par Pan, Zhou et al. 2019 et utilisent cette fois-ci les propriétés des fonctions de croyance et la capacité du modèle crédibiliste à prédire de manière incertaine et imprécise.
L'*entropie de Höhle* $H_h(m)$ utilise la crédibilité Cr.

$$H_h(m) = -\sum_{A \in 2^\Omega} m(A)\, log_2(Cr(A)) \tag{5.1}$$

L'*entropie de Yager* $H_y(m)$ utilise la plausibilité Pl.

$$H_y(m) = -\sum_{A \in 2^\Omega} m(A)\, log_2(Pl(A)) \tag{5.2}$$

L'*entropie de Nguyen* $H_y(m)$ n'utilise que la fonction de masse m.

$$H_n(m) = -\sum_{A \in 2^\Omega} m(A)\, log_2(m(A)) \tag{5.3}$$

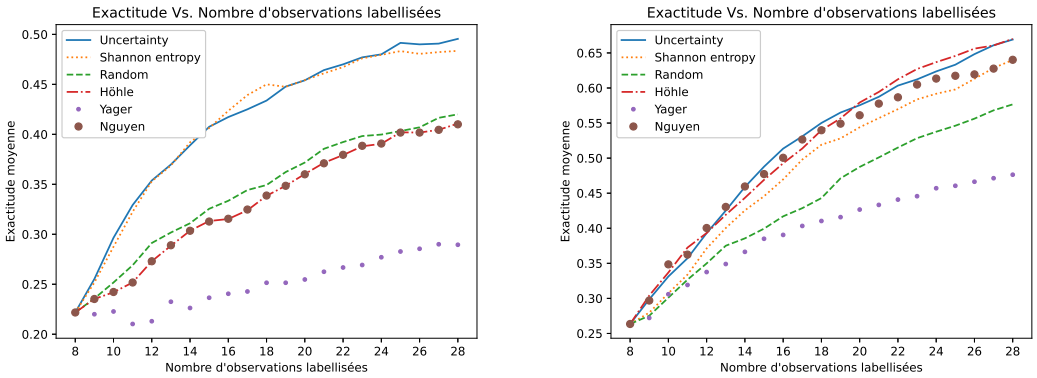
Expérience de comparaison entre différents échantillonnages

Pour toutes les expériences précédentes, l'échantillonnage par incertitude a été utilisé lors de l'apprentissage actif avec la formule (2.23). Cette expérience vise à comparer les performances du modèle EK-NN en fonction de l'entropie utilisée lors de l'échantillonnage. La version γ-EKNN, présentée dans la section 4.1.3, a été utilisée comme référence pour cette expérience. L'échantillonnage par incertitude utilisé lors des expériences précédentes est comparé à l'échantillonnage aléatoire, à l'entropie de Shannon et aux entropies crédibilistes de Höhle, Yager et Nguyen. La figure 5.4 représente l'évolution de l'exactitude moyenne (sur 100 expériences) en fonction du nombre de requêtes labellisées par apprentissage actif avec ces différents échantillonnages. En s'intéressant uniquement aux jeux de données *Credal Bird-10* et *Credal Dog-7* on constate que les formules non crédibilistes obtiennent de bonnes ou meilleures performances. L'entropie de Shannon et la formule utilisée dans les expériences précédentes offrent des résultats performants et très similaires. Les entropies crédibilistes de Höle et Nguyen ont également toutes deux des résultats presque identiques.

Utiliser les entropies crédibilistes nous permet au mieux d'égaler les performances de l'échantillonnage par incertitude. L'objectif étant d'améliorer les performances et de réduire les coûts de labellisation, une autre approche a été envisagée en se concentrant sur l'information représentée lors de l'échantillonnage par incertitude. Cette méthode est présentée au chapitre suivant.

5.3 Conclusion du chapitre

Ce chapitre s'intéresse à la fois à la qualité des labels, avec l'utilisation de modèles crédibilistes et de labels riches, mais également à la quantité de labels, puisque l'utili-



(a) (b)

FIGURE 5.4 – Comparaison de différents échantillonnages avec l'exactitude moyenne des modèles en fonction du nombre d'observations labellisées, 8 observations sont labellisées aléatoirement et 20 requêtes sont faites. Jeux de données *Credal Bird-10* (*a*) et *Credal Dog-7* (*b*).

sation d'apprentissage actif permet de réduire le nombre d'observations labellisées. Les premières expériences offrent des résultats optimistes quant aux performances de classifieurs crédibilistes en apprentissage actif. Une approche par entropie crédibiliste est aussi envisagée mais ne permet pas de conclure sur un gain en performance comparé aux méthodes classiques de l'apprentissage actif.

Dans le chapitre suivant, une toute autre approche est envisagée en s'intéressant à la décomposition des incertitudes du modèle et en cartographiant les meilleures observations à requêter. La capacité du modèle à définir ces zones d'incertitude et à les catégoriser peut s'avérer être une information importante en apprentissage actif et les résultats obtenus sont bien plus optimistes.

Chapitre 6

Échantillonnage par incertitudes crédibilistes

Ce chapitre introduit deux nouvelles méthodes d'échantillonnage, selon une incertitude crédibiliste et une incertitude épistémique crédibiliste, qui permettent cette fois-ci d'arriver à des résultats plus prometteurs et à une représentation plus riche de l'incertitude du modèle. Les récents travaux sur la décomposition de l'incertitude en incertitudes réductible et irréductible sont repris, et en plus de la nouvelle proposition de représentation visuelle des incertitudes du modèle, les méthodes proposées ne se cantonnent plus uniquement aux labels riches, mais proposent même d'améliorer les performances de l'apprentissage actif sur des jeux de données classiques. Des résultats expérimentaux en apprentissage actif sont présentés et montrent des résultats très satisfaisants. Enfin, un bilan examinant les limites et les perspectives de ces méthodes est proposé, une conclusion clôture le chapitre.

Publications

— Arthur Hoarau, Vincent Lemaire, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2024b), « Méthode crédibiliste pour l'extraction d'incertitudes sans dépendance aux observations », in : *Extraction et Gestion des Connaissances (EGC)*

— Arthur Hoarau, Vincent Lemaire, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2024a), « Evidential uncertainties and rich labels for active learning », in : *Machine Learning*

— Arthur Hoarau et Vincent Lemaire (2024), « DEMAU : Decompose, Explore, Model and Analyse Uncertainties », in : *Soumis à ECMLPKDD*

6.1 Méthodes d'échantillonnage

Jusqu'à présent, l'objectif a été de réduire les coûts de labellisation en utilisant des modèles crédibilistes. Utiliser les entropies crédibilistes en apprentissage actif n'a pas permis d'aboutir à des résultats satisfaisants.

Dans cette partie, l'intérêt est porté vers la représentation de l'incertitude du modèle lors de l'échantillonnage par incertitude. Deux nouvelles approches sont introduites, l'échantillonnage par incertitude crédibiliste et celui par incertitude épistémique crédibiliste. Ces deux méthodes permettent de représenter l'incertitude du modèle en prenant en compte l'incertitude déjà présente dans les labels. Même dans le cas où aucune incertitude n'est présente dans les labels, ces méthodes exploitent la capacité du modèle à faire une prédiction incertaine et imprécise. La première stratégie utilise la discorde et la non-spécificité, déjà utilisées dans ce document et la seconde est une extension de l'incertitude épistémique au raisonnement crédibiliste et à plusieurs classes, tout en simplifiant la phase calculatoire.

L'échantillonnage par incertitude est d'abord rappelé plus en détails, afin d'introduire les méthodes proposées. Davantage de détails concernant les expériences de cette partie ont été détaillées en annexe du chapitre 6.5.

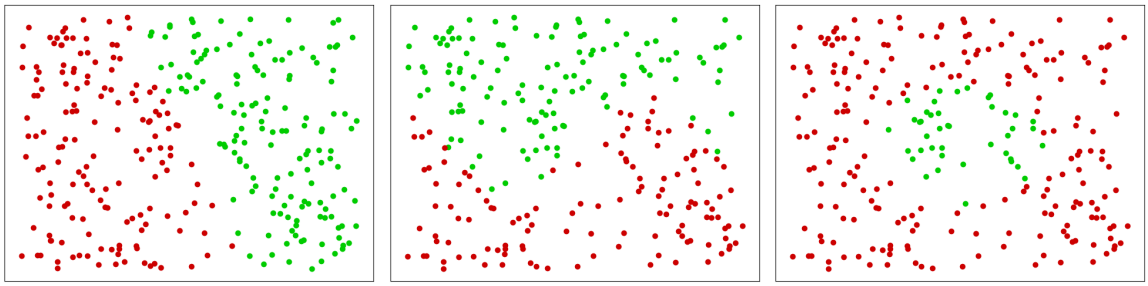
6.1.1 Échantillonnage par incertitude

Étant donné un modèle d'apprentissage et un jeu de données, l'objectif est de représenter l'incertitude du modèle pour définir les zones de l'espace où l'ajout de nouvelles observations serait pertinent pour améliorer les performances. Les critères d'échantillonnage par incertitude souvent utilisés sont la mesure de confiance, donnée par l'équation (2.23) et l'entropie de Shannon, donnée par l'équation (2.24). On a donc $\mathcal{U}$ l'incertitude liée à la labellisation de x pour un modèle donné θ et $\Omega = \{\omega_1, \ldots, \omega_M\}$ l'ensemble des M classes possibles. L'incertitude $\mathcal{U}$ peut donc être calculée avec l'entropie de Shannon :

$$\mathcal{U}(x) = -\sum_{\omega \in \Omega} p(\omega|x)\log[p(\omega|x)], \tag{6.1}$$

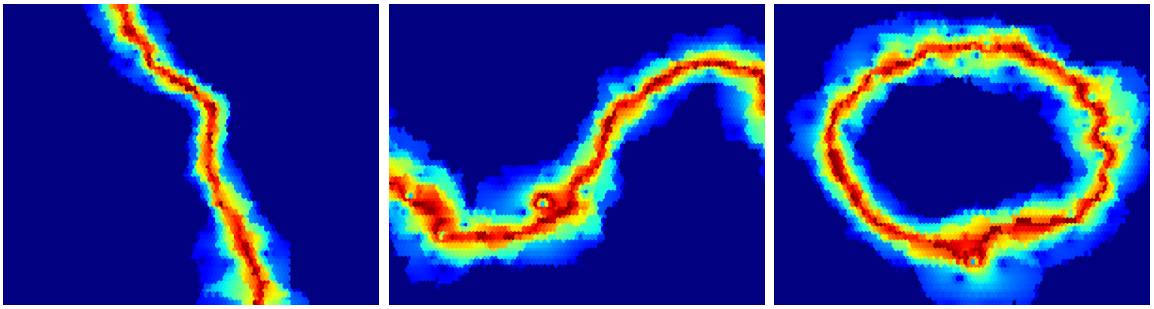
ou encore avec une mesure de confiance :

$$\mathcal{U}(x) = 1 - \max_{\omega \in \Omega}[p(\omega|x)]. \tag{6.2}$$



(a) (b) (c)

FIGURE 6.1 – Trois jeux de données (Figure 6.1a), (Figure 6.1b) et (Figure 6.1c) à deux classes et sur deux dimensions.



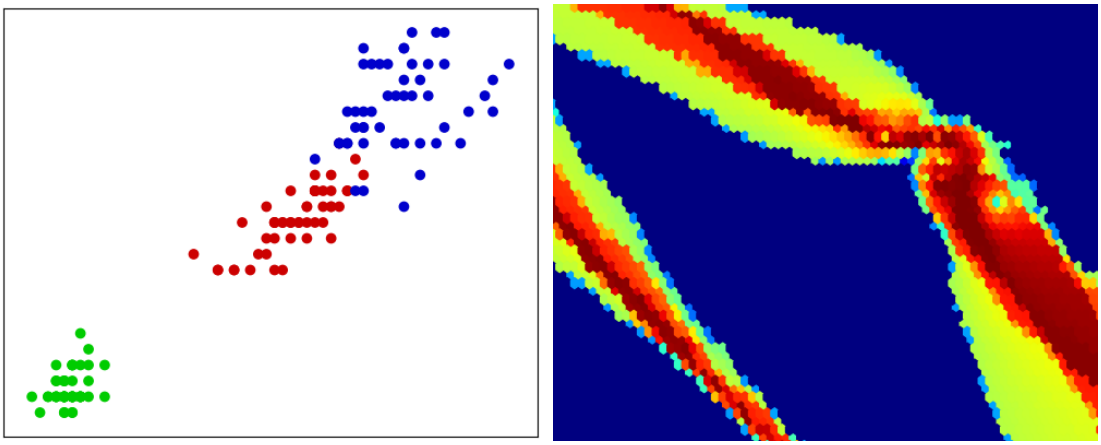
(a) (b) (c)

FIGURE 6.2 – De gauche à droite, les zones d'incertitude correspondant aux jeux de données (Figure 6.1a), (Figure 6.1b) et (Figure 6.1c) avec un échantillonnage par incertitude.

Mesurer l'incertitude d'un modèle lors de la phase de prédiction peut s'avérer utile pour l'apprentissage actif, en permettant au modèle de requêter de nouveaux labels dans ces zones d'incertitude. La figure 6.1 représente trois jeux de données à deux dimensions, possédant deux classes parfaitement séparables.

Étant donné un modèle et un critère d'échantillonnage, il est possible de calculer l'incertitude en n'importe quel point de l'espace (voir l'annexe du chapitre pour tous les détails expérimentaux). Ce résultat est présenté en figure 6.2 et pour chaque jeu de données, la zone d'incertitude du modèle est représentée. Plus la couleur tend vers le rouge, plus le modèle est incertain. Ces zones d'incertitude sont souvent confondues avec la frontière de décision du modèle, puisque c'est à cet endroit que le modèle hésite le plus entre plusieurs classes. Généralement, plus une observation est proche de la frontière de décision, moins le modèle est certain de sa prédiction.

Un exemple sur données réelles est présenté sur la figure 6.3. Le jeu de données Iris de Fisher est présenté sur deux variables. Les trois classes d'Iris sont en figure 6.3a et les



(a) (b)

FIGURE 6.3 – Sur la gauche, le jeu de données Iris de Fisher avec deux variables descriptives, la longueur (horizontale) et la largeur (verticale) des pétales. Trois classes sont représentées, les Setosas sont en vert, les Versicolors en rouge et les Virginicas en bleu. Sur la droite, les zones d'incertitude du modèle sont représentées.

zones d'incertitude du modèle sont en figure 6.3b, ce sont les régions où le modèle serait le plus incertain lors d'une prédiction sur une nouvelle observation. Autrement dit, si une nouvelle observation se trouve dans une zone rouge, le modèle aura plus de difficulté à prédire sa classe.

L'échantillonnage par incertitude consiste à choisir les observations pour lesquelles le modèle est le moins certain de sa prédiction. Il s'agit d'un des principes de base de l'apprentissage actif, cependant, d'autres méthodes permettent d'extraire plus d'information concernant l'incertitude du modèle. Comme proposé dans de récents articles (voir Hüllermeier et Waegeman 2021, Kendall et Gal 2017 et Charpentier, Zügner et al. 2020), l'incertitude peut être décomposée en deux notions distinctes : l'incertitude épistémique et l'incertitude aléatoire. L'incertitude aléatoire découle de la propriété stochastique de l'événement et n'est donc pas réductible, tandis que l'incertitude épistémique est liée à un manque de connaissances et peut être réduite.

6.1.2 Incertitudes épistémique et aléatoire

L'incertitude $\mathcal{U}(x)$ peut être séparée en deux incertitudes (*cf.* Hora 1996), l'une réductible et l'autre irréductible. L'exemple de la figure 6.4 [1] présente ces deux types d'in-

1. Cet exemple est repris de l'intervention de Eyke Hüllermeier "Representation and Quantification of Uncertainty in Machine Learning" à la conférence LFA2022. Dans notre exemple, le mot *Pile* s'écrit

(a) Incertitude aléatoire (b) Incertitude épistémique

FIGURE 6.4 – Représentation des incertitudes épistémique et aléatoire au travers du lancer d'une pièce et du mot *pile* ou *face* écrit en Finnois.

certitude, sur la figure 6.4a le résultat d'un lancer de pièce est incertain et il n'est pas possible d'augmenter sa connaissance pour prédire le résultat, cette ignorance est appelée incertitude aléatoire. Sur la figure 6.4b l'un des deux mots *pile* ou *face* est écrit en finnois, c'est une incertitude qui peut être levée en apprenant cette langue, c'est l'incertitude épistémique.

Être capable de modéliser ces deux incertitudes peut permettre de délimiter où il est plus pertinent d'apporter de la connaissance et où cela est inutile. L'incertitude $\mathcal{U}(x)$ est souvent notée comme la somme de l'incertitude épistémique $\mathcal{U}_e(x)$ et de l'incertitude aléatoire $\mathcal{U}_a(x)$:

$$\mathcal{U}(x) = \mathcal{U}_e(x) + \mathcal{U}_a(x). \tag{6.3}$$

Pour un problème à deux classes $\Omega = \{0, 1\}$, une proposition est faite par Senge, Bösner et al. 2014 pour modéliser ces incertitudes [2] en calculant la plausibilité π d'appartenance à chaque classe avec la formule suivante, et selon un modèle probabiliste θ :

$$\begin{aligned} \pi(1|x) &= \sup_{\theta \in \Theta} \min[\pi_\Theta(\theta), p_\theta(1|x) - p_\theta(0|x)], \\ \pi(0|x) &= \sup_{\theta \in \Theta} \min[\pi_\Theta(\theta), p_\theta(0|x) - p_\theta(1|x)], \end{aligned} \tag{6.4}$$

avec $\pi_\Theta(\theta)$ qui dépend de la vraisemblance $L(\theta)$ et le maximum de vraisemblance $L(\hat{\theta})$:

$$\pi_\Theta(\theta) = \frac{L(\theta)}{L(\hat{\theta})}. \tag{6.5}$$

L'incertitude épistémique est donc la plus élevée quand les deux classes sont très plausibles alors que l'incertitude aléatoire est plus grande quand les deux classes sont peu

Kruuna en finnois.

2. Le formalisme décrit par Nguyen, Shaker et al. 2022 est utilisé ici.

plausibles :

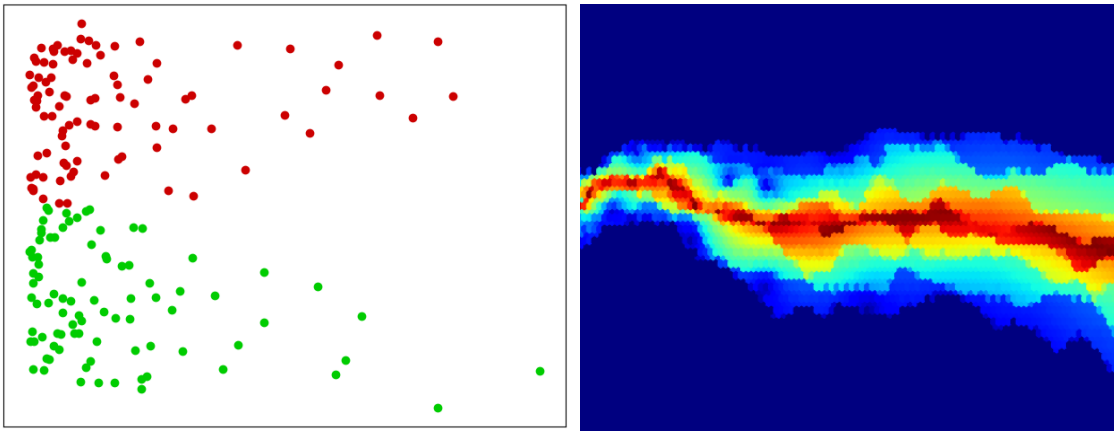
$$\begin{aligned} \mathcal{U}_e(x) &= \min[\pi(1|x), \pi(0|x)], \\ \mathcal{U}_a(x) &= 1 - \max[\pi(1|x), \pi(0|x)]. \end{aligned} \tag{6.6}$$

Cette étape calculatoire dépend à la fois de la prédiction du modèle, mais aussi des observations. Pour résumer, moins d'observations sont présentes dans une région, ou moins d'éléments de décision sont disponibles pour prédire fortement une classe, plus la plausibilité des deux classes est grande, et plus l'incertitude est réductible (et donc épistémique) en ajoutant de la connaissance. Un exemple est présenté sur la figure 6.5 [3] montrant un jeu de données à deux classes (voir figure 6.5a) et la zone d'incertitude du modèle (voir figure 6.5b) selon l'échantillonnage par incertitude, discuté dans la section précédente. Une ligne horizontale peut être distinguée où l'incertitude du modèle est maximale. Cependant, l'échantillon représenté sur la figure 6.5a montre qu'une partie de l'incertitude peut être levée plus aisément en ajoutant de nouvelles observations. Sur la figure 6.6, trois différents jeux de données montrent la possible évolution de l'échantillon en ajoutant des observations. Peu importe la distribution finale, l'incertitude à gauche n'est presque pas réductible, alors que l'incertitude sur la droite peut être modifiée en ajoutant des observations. Ces deux incertitudes peuvent être calculées avec l'équation (6.6). L'incertitude épistémique, l'incertitude aléatoire et l'incertitude totale sont présentées sur la figure 6.7. L'incertitude aléatoire, et donc irréductible, est visible sur la figure 6.7b et l'incertitude épistémique, réductible, est visible sur la figure 6.7a. L'incertitude totale est alors la somme des deux (voir figure 6.7c). L'objectif est d'utiliser l'incertitude épistémique pour exploiter les zones où le modèle peut apprendre à nouveau et où cela a de l'impact.

Utiliser l'incertitude épistémique comme méthode d'échantillonnage n'est pas régressif. Lorsque les incertitudes épistémique et aléatoire sont indiscernables, les zones d'incertitudes sont similaires à celles produites par la méthode étudiée précédemment.

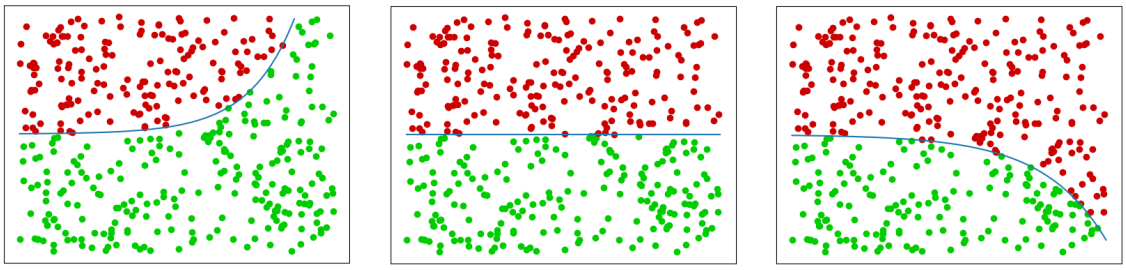
Une telle information peut être utilisée pour trouver les zones d'incertitude réductible, mais cela n'est pas compatible avec des labels riches qui contiennent eux aussi de l'incertitude. Le calcul de cette incertitude épistémique est aussi dépendant des observations, ces contraintes nous conduisent à la section suivante où un échantillonnage par incertitude pour labels riches est proposé, le calcul de cet échantillonnage est aussi étendu à plusieurs classes.

3. Voir l'annexe du chapitre pour les détails expérimentaux.



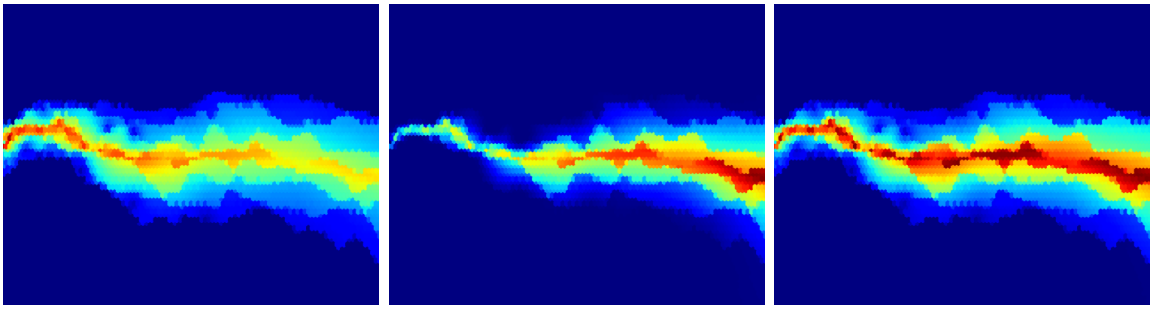
(a) (b)

FIGURE 6.5 – Sur la gauche, l'échantillon d'un jeu de données, et sur la droite, la zone d'incertitude du modèle selon l'échantillonnage par incertitude.



(a) (b) (c)

FIGURE 6.6 – Trois jeux de données possibles, selon les observations présentes sur la figure 6.5a.



(a) (b) (c)

FIGURE 6.7 – Les zones d'incertitude épistémique et aléatoire, selon la figure 6.5a. De gauche à droite, l'incertitude aléatoire (Figure 6.7a), l'incertitude épistémique (Figure 6.7b) et l'incertitude totale (Figure 6.7c).

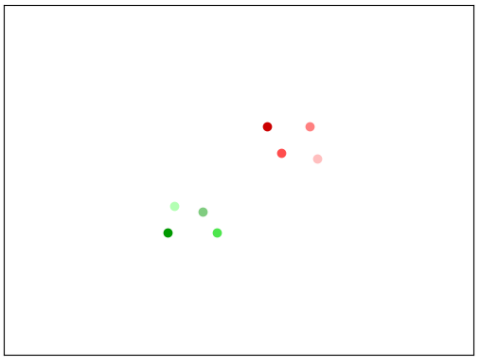


FIGURE 6.8 – Représentation de plusieurs observations sur deux dimensions avec leur label riche. Plus le point est foncé, plus le label est incertain. Plus le point est clair, plus le label est ignorant.

6.1.3 Échantillonnage par incertitude crédibiliste

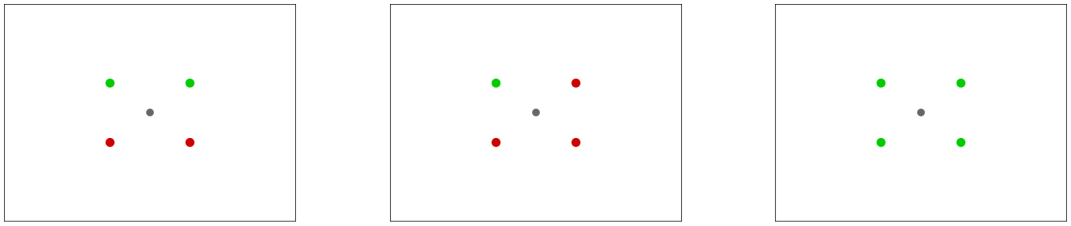
Dans cette section, un échantillonnage par incertitude capable de prendre en compte des labels riches est présenté. Cet échantillonnage ne dépend plus des observations, mais uniquement de la prédiction du modèle [4]. La phase calculatoire est aussi grandement simplifiée [5] et la méthode est naturellement étendue à un nombre de classes supérieur à deux.

Les labels peuvent donc à présent être incertains et imprécis, ce qui veut dire qu'une information supplémentaire sur l'ignorance peut être représentée. La figure 6.8 montre comment ces labels sont représentés dans cette section. Plus le point est foncé, moins il y a d'ignorance dans le label [6], et plus le point est clair, plus le label est ignorant [7].

Discorde et non-spécificité

Dans le cadre de la théorie des fonctions de croyance, la discorde et la non-spécificité, respectivement données par les équations (2.17) et (2.16) sont des outils permettant de modéliser l'incertitude. Nous proposons d'utiliser la représentation de Klir et Wierman 1998 pour un échantillonnage par incertitude, des ponts peuvent être faits avec les incertitudes épistémique et aléatoire.

4. L'incertitude n'est plus directement dépendante des observations, mais le modèle l'est toujours.
5. Voir la section 6.5.
6. Par exemple : *Je suis sûr qu'il s'agit d'une Setosa.*
7. Par exemple : *Je n'ai aucune idée entre Setosa et Versicolor.*



(a) (b) (c)

FIGURE 6.9 – Représentation de la discorde (au point central), le maximum de discorde est atteint sur la figure 6.9a, la figure 6.9b montre moins de discorde et sur la figure 6.9c il n'y a pas de discorde.

Discorde : Elle est ici appliquée en sortie d'un modèle capable de faire une prédiction incertaine et imprécise [8]. Elle représente la quantité d'information conflictuelle dans la prédiction du modèle et se calcule avec la formule (2.17). La figure 6.9 représente trois différents cas où la discorde varie, de très élevée quand les labels autour du point central (l'observation à labelliser) sont fortement en désaccord (Figure 6.9a) à très faible lorsque les labels sont en accord (Figure 6.9c).

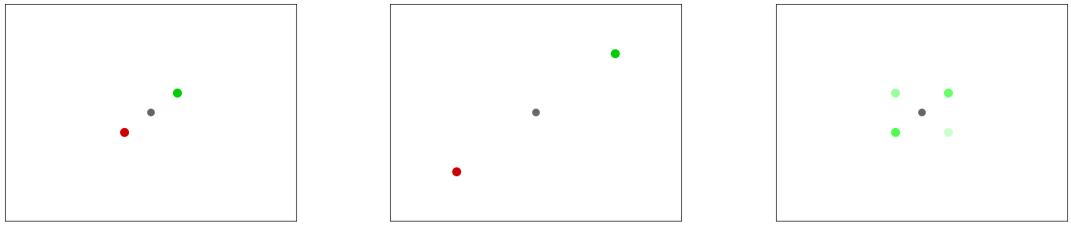
Non-spécificité : Elle permet de quantifier le degré d'ignorance du modèle, au plus elle est élevée, au plus la réponse du modèle est imprécise, elle se calcule avec la formule (2.16). La figure 6.10 représente trois cas où la non-spécificité varie. Sur la figure 6.10a la non-spécificité est basse, comme il y a des sources d'informations pertinentes proches de l'observation à labelliser, pour la figure 6.10b la non-spécificité augmente au plus les sources d'information s'éloignent de l'observation et sur la figure 6.10c la non-spécificité est aussi élevée puisque les sources proches sont elles-mêmes ignorantes.

Incertitude crédibiliste : Elle est alors dérivée de la discorde et de la non-spécificité en additionnant les deux formules précédentes :

$$\mathcal{U}_m(x) = N(x) + D(x), \tag{6.7}$$

avec $N(x)$ et $D(x)$ respectivement la non-spécificité et la discorde du modèle en x. Klir propose d'utiliser le même poids pour la discorde et la non-spécificité, mais un paramètre

8. Le modèle des K plus proches voisins crédibilistes de Denœux 1995 est ici considéré pour illustrer les exemples, ces résultats peuvent varier en fonction du modèle utilisé.



(a) (b) (c)

FIGURE 6.10 – Représentation de la non-spécificité (au point central), sur la figure 6.10a la non-spécificité est basse alors que sur la figure 6.10b elle est plus importante. Sur la figure 6.10c la non-spécificité est aussi élevée.

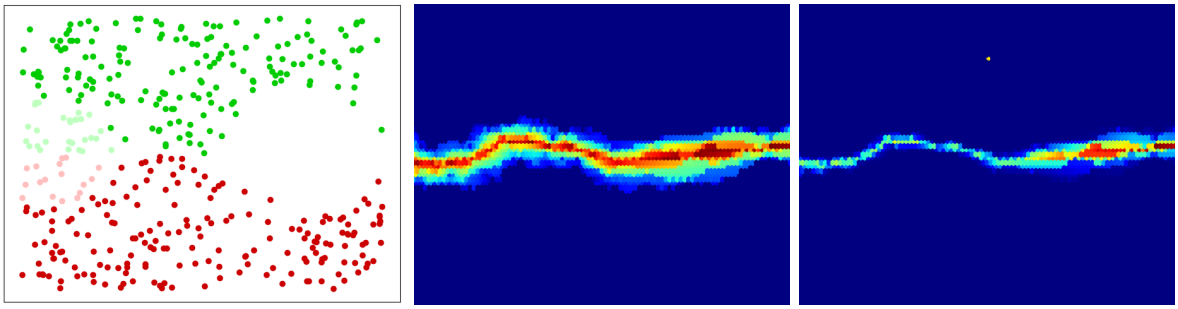
$\lambda \in [0, 1]$ est introduit dans Denoeux et Bjanger 2000 permettant d'apporter plus de poids à la non-spécificité (pour plus d'exploration [9]) ou à la discorde (pour plus d'exploitation [10]) :

$$\mathcal{U}_m(x) = \lambda N(x) + (1 - \lambda)D(x). \tag{6.8}$$

Il est important de noter que cette incertitude est naturellement étendue à un nombre de classes $|\Omega| \geq 2$ supérieur à deux. Cette forme présente l'avantage d'identifier l'incertitude réductible, mais aussi de prendre en compte l'incertitude déjà présente dans les labels et d'être ajustable pour plus d'exploration ou d'exploitation (voir le chapitre 2.2.6 pour le dilemme exploration-exploitation). La figure 6.11 montre un jeu de données avec deux zones d'incertitude (Figure 6.11a), sur la droite une zone où il manque des observations et sur la gauche une zone où les labels sont ignorants. L'échantillonnage par incertitude, en utilisant l'entropie de Shannon donnée par l'équation (2.24), ou la mesure de confiance donnée par l'équation (2.23), n'est capable de distinguer aucune de ces deux zones (Figure 6.11b). L'incertitude épistémique, calculée à partir de l'équation (6.6), est capable de distinguer l'incertitude liée au placement des observations dans l'espace (l'incertitude de droite), mais pas l'incertitude liée à l'ignorance des sources (Figure 6.11c) [11].

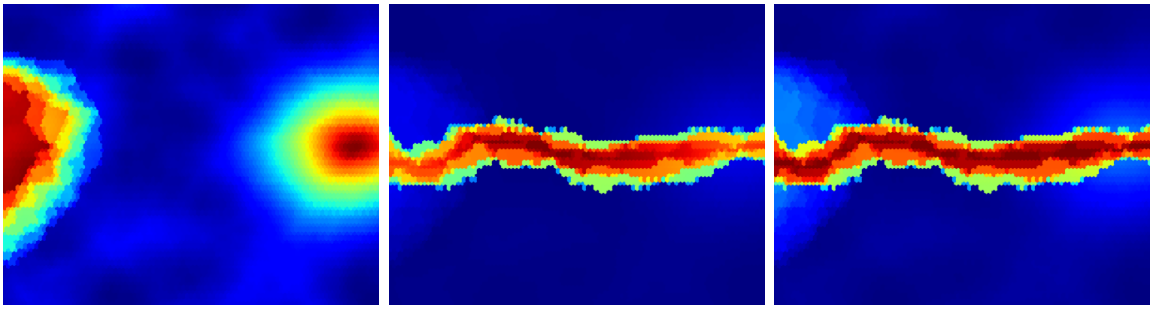
La proposition d'utiliser l'incertitude crédibiliste (somme de la discorde et la non-spécificité) permet de représenter l'ensemble des incertitudes présentes dans le jeu de données. La figure 6.12 montre les zones de non-spécificité (Figure 6.12a), de discorde

9. Nous proposons d'utiliser cette représentation pour traiter le problème d'exploration-exploitation en apprentissage actif.
10. Voir le chapitre 2.2.6 pour le dilemme exploration-exploitation.
11. Voir l'annexe du chapitre pour les détails expérimentaux.



(a) (b) (c)

FIGURE 6.11 – Un jeu de données avec labels riches (Figure 6.11a) accompagné des zones d'incertitudes selon l'échantillonnage par incertitude (Figure 6.11b) et l'échantillonnage par incertitude épistémique (Figure 6.11c).

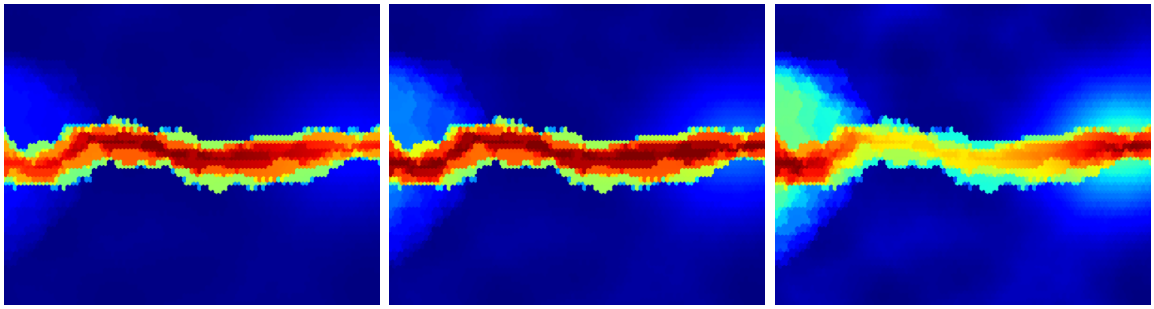


(a) (b) (c)

FIGURE 6.12 – Zones d'incertitude pour le jeu de données présenté sur la figure 6.11a, selon la non-spécificité (Figure 6.12a), la discorde (Figure 6.12b) et l'incertitude crédibiliste totale (Figure 6.12c).

(Figure 6.12b) et d'incertitude crédibiliste (Figure 6.12c). C'est l'incertitude crédibiliste qui est alors utilisée pour l'échantillonnage, il est également possible de faire varier les résultats vers plus d'exploration ou plus d'exploitation en modifiant λ. La figure 6.13 montre les zones d'incertitude pour différentes valeurs de λ, plus de discorde sur la figure 6.13a à plus de non-spécificité sur la figure 6.13c.

Même si la discorde peut rappeler l'incertitude aléatoire (non réductible) et la non-spécificité peut rappeler l'incertitude épistémique (réductible), ces notions ne sont pas complètement équivalentes et interchangeables, l'information représentée peut différer. C'est pourquoi nous proposons également dans la section suivante, une extension de l'incertitude épistémique (et de l'incertitude aléatoire) pour des labels riches et pour un nombre de classes supérieur à deux.



(a) (b) (c)

FIGURE 6.13 – Différentes zones d'incertitude crédibiliste, modifiant la quantité de non-spécificité et de discorde. Avec $\lambda < 0.5$ (ici 0.2), plus de discorde est prise en compte (Figure 6.13a), avec $\lambda = 0.5$, la discorde et la non-spécificité sont autant utilisées (Figure 6.13b) et avec $\lambda > 0.5$ (ici 0.8), plus de non-spécificité est prise en compte (Figure 6.13c).

6.1.4 Echantillonnage par incertitude épistémique crédibiliste

L'incertitude épistémique peut être étendue aux labels riches en utilisant la notion de plausibilité dans le cadre des fonctions de croyance. Elle représente la quantité totale de croyance qui ne supporte pas l'événement complémentaire pour une classe ω ou plus généralement pour un élément $A \in 2^{\Omega}$. La plausibilité $Pl(A)$, calculée à partir de l'équation (2.5), définit la croyance qui pourrait être attribuée en A.

La plausibilité représentant une croyance cohérente, la crédibilité $Cr(A)$, calculée à partir de l'équation (2.4), est la croyance totale qui supporte directement A. En monde fermé (*i.e.* $m(\emptyset) = 0$), il est possible de noter $Pl(A) = 1 - Cr(\bar{A})$.

De manière analogue à l'équation (6.6) et pour deux classes $\Omega = \{\omega_1, \omega_2\}$, l'incertitude épistémique crédibiliste est maximale quand les deux classes sont fortement plausibles. Les incertitudes épistémique et aléatoire crédibilistes proposées sont donc définies comme suit :

$$\begin{aligned} \mathcal{U}_e(x) &= \min[Pl(\omega_1|x), Pl(\omega_2|x)], \\ \mathcal{U}_a(x) &= 1 - \max[Pl(\omega_1|x), Pl(\omega_2|x)]. \end{aligned} \tag{6.9}$$

L'équation de l'incertitude aléatoire crédibiliste peut être réécrite pour dépendre de la crédibilité Cr :

$$\mathcal{U}_a(x) = \min[Cr(\omega_1|x), Cr(\omega_2|x)]. \tag{6.10}$$

La somme des incertitudes épistémique et aléatoire est l'incertitude crédibiliste totale :

$$\mathcal{U}(x) = \mathcal{U}_e(x) + \mathcal{U}_a(x). \tag{6.11}$$

Pour l'extension à un nombre de classes supérieur à deux, l'équation de l'incertitude épistémique ne peut être simplifiée au minimum de plausibilité :

$$\begin{aligned} \mathcal{U}_e(x) &\neq \min([Pl(\omega|x)|\omega \in \Omega]), \\ \mathcal{U}_a(x) &\neq 1 - \max([Pl(\omega|x)|\omega \in \Omega]). \end{aligned} \tag{6.12}$$

Il est donc préférable de définir dans un premier temps l'incertitude relative à une classe ω, réécrite avec la crédibilité Cr pour éviter d'avoir à manipuler $\bar{\omega}$:

$$\begin{aligned} \mathcal{U}_e(\omega|x) &= \min[Pl(\omega|x), Pl(\bar{\omega}|x)] \\ &= \min[Pl(\omega|x), 1 - Cr(\omega|x)]. \end{aligned} \tag{6.13}$$

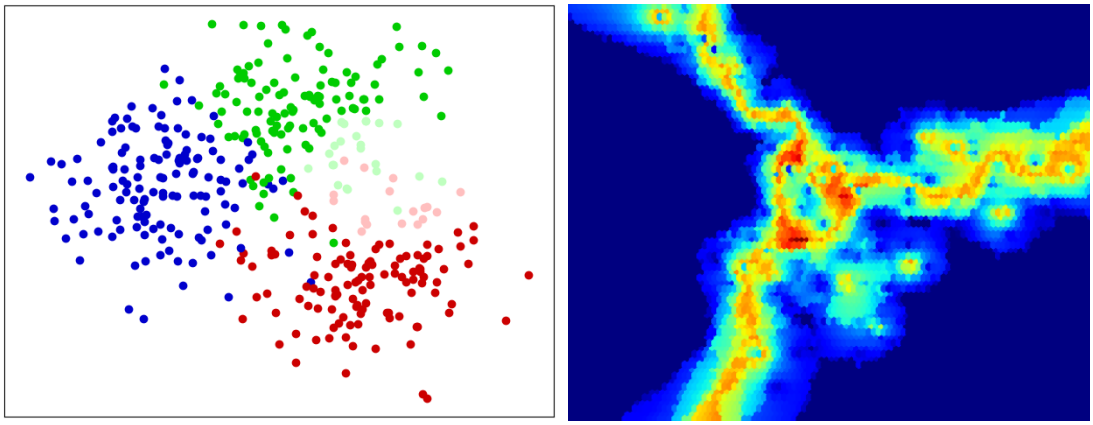
L'extension de l'incertitude épistémique et aléatoire pour un nombre de classes $|\Omega| \geq 2$ est alors la somme des incertitudes pour chaque classe :

$$\begin{aligned} \mathcal{U}_e(x) &= \sum_{\omega \in \Omega} \min[Pl(\omega|x), 1 - Cr(\omega|x)], \\ \mathcal{U}_a(x) &= \sum_{\omega \in \Omega} \min[Cr(\omega|x), 1 - Pl(\omega|x)]. \end{aligned} \tag{6.14}$$

L'exemple de la figure 6.14 montre un jeu de données à trois classes comprenant une zone d'ignorance pour certains labels (entre la classe rouge et verte). Que ce soit l'échantillonnage probabiliste classique, défini par les équations (2.23) et (2.24) ou épistémique, défini par l'équation (6.6), les incertitudes ne peuvent pas modéliser l'imprécision présente dans les labels, cette zone d'incertitude moins complète est présentée sur la figure 6.14b.

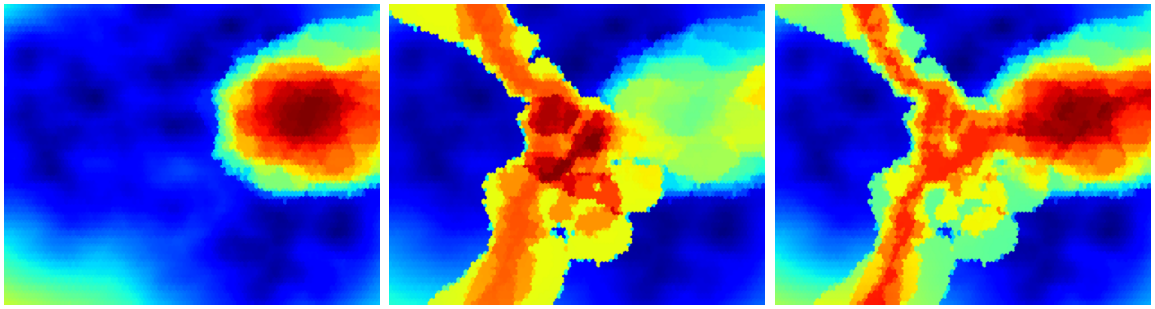
L'incertitude présentée précédemment, résultante de la discorde et de la non-spécificité, est présentée sur la figure 6.15. Elle parvient à capturer l'information à la fois pour une exploration (Figure 6.15a) et pour une exploitation (Figure 6.15b) afin de donner une meilleure représentation de l'incertitude (Figure 6.15c).

L'autre proposition, l'extension de l'incertitude épistémique, est présentée dans l'expérience suivante. Dans un premier temps, la zone d'incertitude épistémique pour chacune des trois classes est présentée sur la figure 6.16. Ensuite, l'incertitude épistémique résultante pour le modèle est déduite de l'équation (6.14) sur la figure 6.17 avec les incertitudes



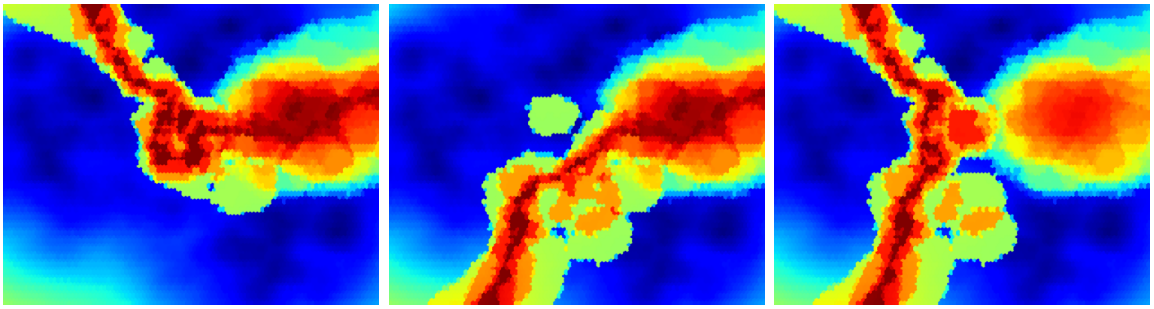
(a) (b)

FIGURE 6.14 – Sur la gauche, un jeu de données à trois classes avec une zone d'ignorance (labellisée avec imprécision) et sur la droite la zone d'incertitude selon l'échantillonnage par incertitude (mesure de confiance).



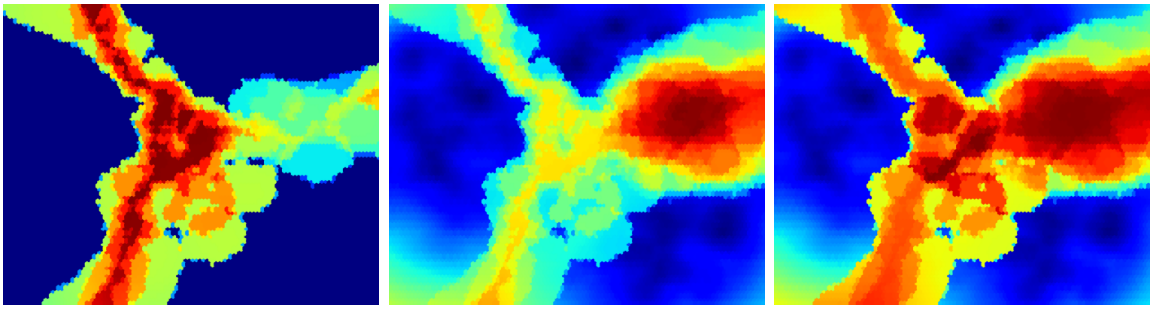
(a) (b) (c)

FIGURE 6.15 – Zones d'incertitude correspondant au jeu de données présenté sur la figure 6.14a selon la non-spécificité (Figure 6.15a), la discorde (Figure 6.15b) et l'incertitude de Klir totale (Figure 6.15c).



(a) (b) (c)

FIGURE 6.16 – Zones d'incertitude correspondant au jeu de données présenté sur la figure 6.14a selon l'incertitude épistémique crédibiliste pour la classe verte (Figure 6.16a), pour la classe rouge (Figure 6.16b) et pour la classe bleue (Figure 6.16c).



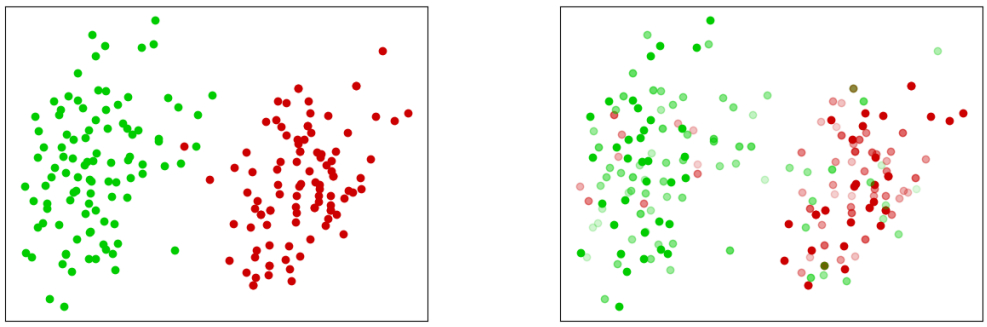
(a) (b) (c)

FIGURE 6.17 – Zones d'incertitude pour les incertitudes épistémique et aléatoire crédibilistes pour le jeu présent sur la figure (6.14a). De gauche à droite, l'incertitude aléatoire (Figure 6.17a), l'incertitude épistémique (Figure 6.17b) et l'incertitude crédibiliste totale (Figure 6.17c).

aléatoire et totale.

6.2 Echantillonage sur données réelles

Cette partie applique les deux méthodes d'échantillonnage proposées à un jeu de données réellement labellisé de manière incertaine et imprécise. Les méthodes conventionnelles pour calculer l'incertitude du modèle ne prennent pas en compte les degrés d'imprécision présents dans ces labels riches. Les deux méthodes proposées sont illustrées sur *Credal Dog-2*, un des jeux de données introduits dans la section 3.2. La figure 6.18 représente le jeu de données sur le premier plan factoriel d'une ACP, les vrais labels sont représentés sur la figure 6.18a et les labels incertains et imprécis donnés par des contributeurs sur la



(a) (b)

FIGURE 6.18 – Sur la gauche (Figure 6.18a), le jeu de données *Credal Dog-2* avec les vrais labels, la race Épagneul Breton est en vert et la race Beagle en rouge. Sur la droite (Figure 6.18b), le jeu de données labellisé de manière incertaine et imprécise par des contributeurs.

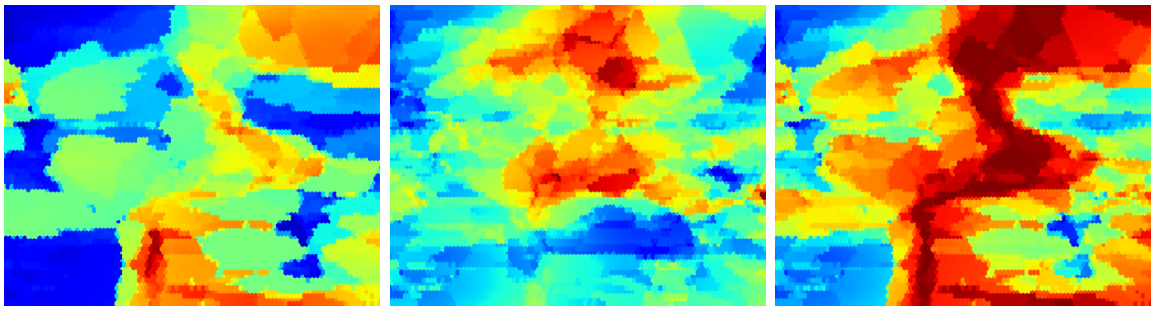
figure 6.18b. Plus les points sont foncés, plus les labels sont certains, et vice-versa.

La figure 6.19 montre les résultats de la première méthode proposée, l'échantillonnage par incertitude crédibiliste, sur le jeu de données à labels riches. La non-spécificité est présentée sur la figure 6.19a et peut être interprétée comme les zones d'imprécision du modèle. La discorde est aussi représentée sur la figure 6.19b et l'incertitude totale (Figure 6.19c) est la somme des deux. En apprentissage actif, c'est l'incertitude totale qui devra être utilisée pour échantillonner.

La seconde méthode proposée, l'extension de l'incertitude épistémique, une incertitude réductible appliquée au raisonnement crédibiliste, est présentée sur la figure 6.20. L'incertitude aléatoire crédibiliste, irréductible est présentée (Figure 6.20a) ainsi que l'incertitude épistémique crédibiliste (Figure 6.20b), réductible. L'incertitude totale (Figure 6.20c) est la somme des incertitudes réductible et irréductible. Pour l'échantillonnage, ce n'est pas l'incertitude totale, mais l'incertitude épistémique, et donc réductible, qui est utilisée.

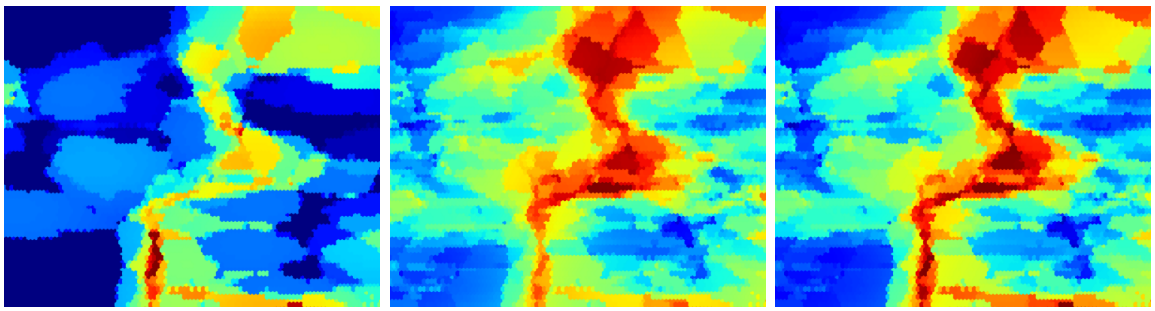
6.3 Application à l'apprentissage actif

L'échantillonnage par incertitude crédibiliste (6.8) a été choisi pour cette série d'expériences, et le seul paramètre de la méthode : λ, est au cœur de cette étude. Une valeur de λ qui tend vers 0 implique plus d'exploitation, alors qu'une valeur de λ qui tend vers 1 implique plus d'exploration.



(a) (b) (c)

FIGURE 6.19 – Zones d'incertitudes correspondantes au jeu de données présenté sur la figure 6.18b selon la non-spécificité (Figure 6.19a), la discorde (Figure 6.19b) et l'incertitude crédibiliste totale (Figure 6.19c).



(a) (b) (c)

FIGURE 6.20 – Zones d'inertitude pour les incertitudes épistémique et aléatoire crédibilistes, correspondantes à la figure 6.18b. De gauche à droite, l'incertitude aléatoire (Figure 6.20a), l'incertitude épistémique (Figure 6.20b) et l'incertitude totale (Figure 6.20c).

TABLEAU 6.1 – Description des jeux de données, avec le nombre d'observations, de classes, de caractéristiques et l'entropie pour la distribution des classes.

Nom du jeu	Observations	Classes	Caractéristiques	Entropie
Breast Cancer	569	2	30	0.95
Ionosphere	351	2	34	0.94
Heart	303	2	7	1.00
Parkinson	195	2	22	0.81
Sonar	208	2	60	1.00
Liver	345	2	6	0.98
Dog-2	200	2	42	1.00
Seeds	210	3	7	1.00
Iris	150	3	4	1.00
Wine	178	3	13	0.99
Glass	214	6	9	0.83
Ecoli	336	8	7	0.73

Pour ces expériences, le paramètre λ est arbitrairement positionné à 0.2, ce qui veut dire que le modèle opte pour davantage d'exploitation que d'exploration. Cette valeur permet d'obtenir de bons résultats, mais nos études pour comprendre quand l'exploration est plus pertinente que l'exploitation (et *vive versa*) ne nous permettent pas pour l'instant de conclure autre chose que ce qui est présenté dans les expériences suivantes (une partie du chapitre 7 est consacré à cette question).

Les résultats sont comparés avec l'échantillonnage aléatoire (voir chapitre 2.2.4) et le très populaire échantillonnage par incertitude 2.23. Les expériences sont réalisés sur des jeux de données possèdant de 2 à 8 classes avec un nombre d'observations variant dans différentes échelles. Les jeux de données sont disponibles sur *UCI Machine Learning Repository* (*cf.* Dua et Graff 2017) et sont très souvent utilisés en apprentissage actif. Le jeu de données *Dog-2*, présenté au chapitre 3 et labellisé de manière imparfaite est également ajouté. Le tableau 6.1 décrit ces jeux de données, avec pour chaque jeu ; le nombre d'observations (ou instances), le nombre de classes, le nombre de caractéristiques et l'entropie de distribution des classes [12].

Étant donné que l'objectif en l'apprentissage actif est de réduire les coûts de labellisation, une expérience consiste à évaluer les performances du modèle au fur et à mesure que les observations sont progessivement labellisées. Les expériences sont arbitrairement

12. Une entropie de 1 correspond à des classes parfaitement équilibrées et une entropie de 0 indique une totale sur-représentation d'une des classes.

stoppées une fois 60% du jeu de données labellisé (il sera clair en analysant les graphiques qu'il est inutile de poursuivre plus loin l'échantillonnage).

Le modèle est le même pour chaque méthode, il s'agit des K plus proches voisins crédibilistes de Denœux 1995, avec $K = 7$ voisins (voir Hoarau, Martin, J.-C. Dubois et Le Gall 2022 et le chapitre 4 de ce document). Les expériences sont réalisées 100 fois pour obtenir une estimation de l'exactitude moyenne du modèle sur chaque jeu de données. Plusieurs critères sont utilisés pour comparer les résultats, dont l'exactitude, l'aire sous la courbe d'exactitude (ACE) et le rang obtenu par la méthode sur chaque jeu de données. Pour l'évaluation statistique, plusieurs tests sont aussi réalisés, un t-test de Student pour l'évaluation des ACE et un test de Friedman accompagné de la méthode de Wilcoxon-Holm pour la validation des diagrammes de différence critique (voir Demšar 2006).

La figure 6.21 montre 6 des 12 jeux de données où la méthode proposée performe de manière significative. Les performances sur jeu complet (*i.e.* absence d'apprentissage actif) sont représentées par la droite en pointillés et la courbe de tirets représente la méthode proposée avec $\lambda = 0.2$. Chaque graphique montre la prédominance de la méthode proposée par rapport à l'échantillonnage par incertitude, temporairement pour les jeux de données Sonar et Heart (seulement au début de l'apprentissage actif pour Sonar et au milieu pour Heart). En supposant un coût de labellisation identique pour chaque observation, il est possible de tirer quelques conclusions.

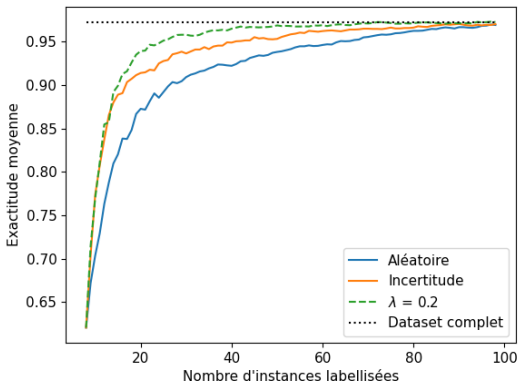
Quelques conclusions à propos de la figure 6.21

— Pour Dog-2 : Quand 99% des performances totales sont atteintes, l'échantillonnage par incertitude parvient à réduire les coûts de labellisation de 62% alors que la méthode proposée parvient à réduire les coûts de 82%.

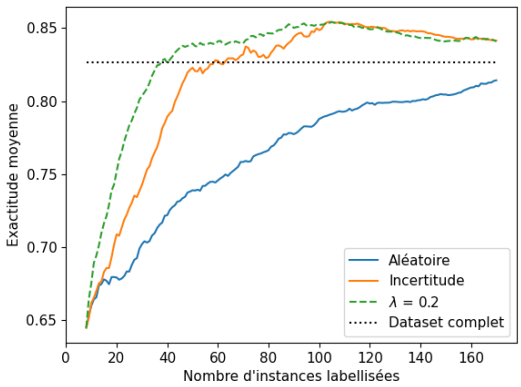
— Pour Ionosphere : En utilisant la méthode proposée, les coûts liés à la labellisation peuvent être réduits d'un facteur 9 avec une perte d'exactitude de 0% par rapport au jeu de données complet. Alors qu'avec l'échantillonnage par incertitude, pour permettre une labellisation 9 fois moins chère, le modèle doit perdre 9% d'exactitude [13].

— Pour Ecoli : 10 requêtes de labellisation sont nécessaires avec l'échantillonnage par incertitude pour atteindre les performances de la méthode proposée avec seulement

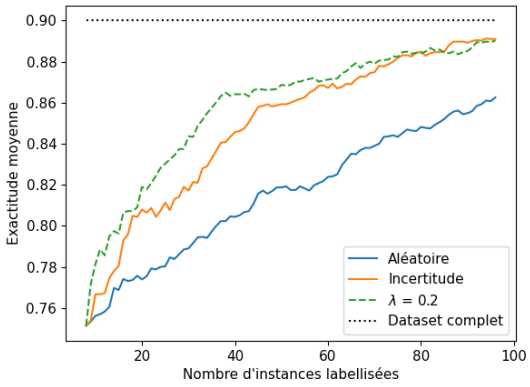
13. Pour ce jeu de données, la réduction du coût de labellisation peut améliorer la performance du modèle, un phénomène qui se produit parfois en apprentissage actif, représenté par la courbe d'apprentissage qui dépasse la ligne horizontale de la performance sur le jeu de données complet.



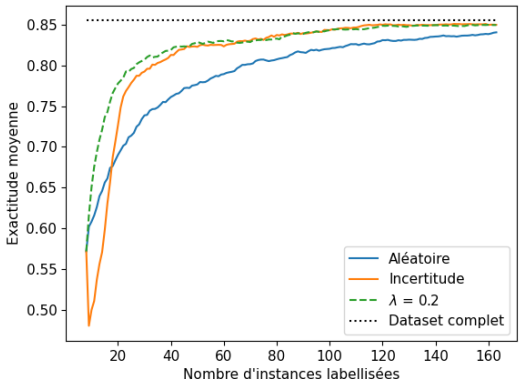
(a) Dog-2



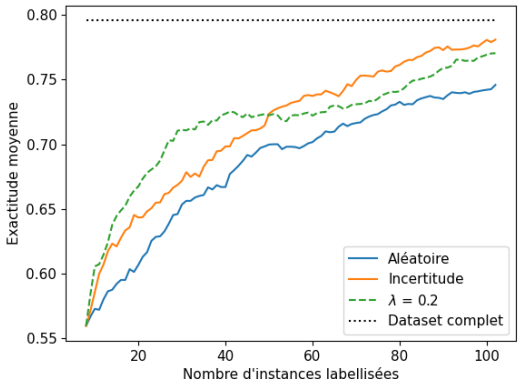
(b) Ionosphere



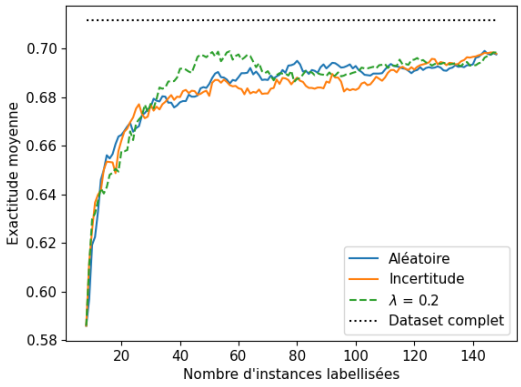
(c) Parkinson



(d) Ecoli

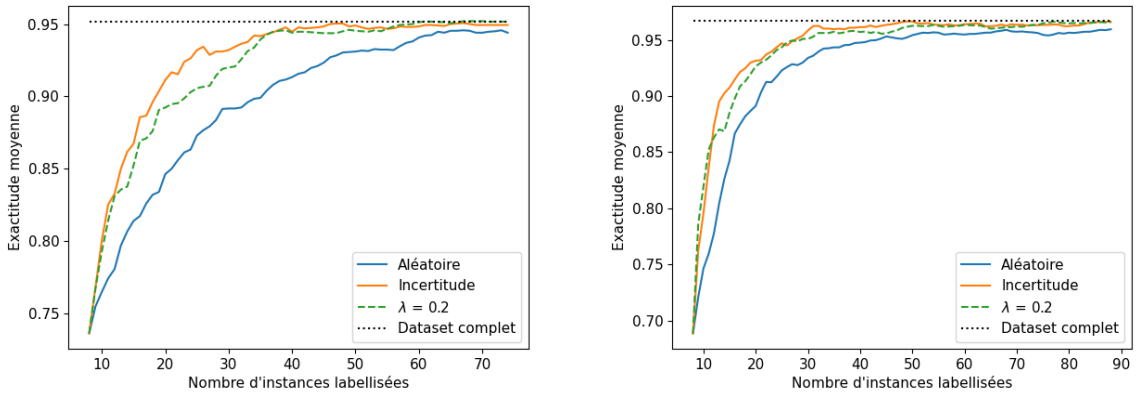


(e) Sonar



(f) Heart

FIGURE 6.21 – Exactitude moyenne *vs.* le nombre d'instances labellisées pour l'échantillonnage aléatoire, l'échantillonnage par incertitude, et la méthode proposée avec $\lambda = 0.2$ et pour 6 jeux de données.



(a) Iris (b) Wine

FIGURE 6.22 – Exactitude moyenne *vs.* le nombre d'instances labellisées pour l'échantillonnage aléatoire, l'échantillonnage par incertitude, et la méthode proposée avec $\lambda = 0.2$ pour les jeux Iris et Wine.

3 requêtes.

Même sur des jeux de données où les performances de la méthode proposée sont plus faibles, l'écart n'est pas toujours conséquent. La figure 6.22 montre deux de ces jeux de données où la méthode proposée performe de manière moins impressionnante. La différence est légèrement plus grande entre la méthode proposée et l'échantillonnage par incertitude sur le jeu de données Iris.

Le tableau 6.2 montre le moyenne des aires sous la courbe d'exactitude pour les trois méthodes étudiées et pour chaque jeu de données. Un t-test statistique est également réalisé entre la meilleure méthode et la seconde pour chaque valeur indiquée en gras. L'échantillonnage aléatoire donne les meilleurs résultats sur le jeu Liver, l'échantillonnage par incertitude donne les meilleurs résultats sur les jeux Seed, Iris et Wine. La méthode proposée est la plus performante sur les 8 autres jeux de données.

Pour affirmer statistiquement quelle méthode est la plus performante, un diagramme de différence critique est tracé. Le premier diagramme est présenté sur la figure 6.23a, il s'agit d'une comparaison de différentes valeurs de λ pour la méthode proposée. Sur la totalité des jeux de données, $\lambda = 0.2$ obtient en moyenne la position 1.83 sur 4 et $\lambda = 0.5$ (*i.e.* autant d'exploration que d'exploitation) obtient en moyenne la position 3.42. Si une ligne relie deux méthodes, cela signifie que malgré la meilleure performance de l'une d'entre elles, les méthodes ne sont pas statistiquement distinctes. Dans l'exemple, $\lambda = 0.2$ est plus performant que $\lambda = 0.3$ et $\lambda = 0.4$, mais sans significativité statistique. Maintenant, la

TABLEAU 6.2 – Moyenne des aires sous la courbe d'exactitude pour l'échantillonnage aléatoire, l'échantillonnage par incertitude et la méthode proposée avec $\lambda = 0.2$ sur chaque jeu de données. Un t-test de Student est également réalisé pour déterminer la significativité de la meilleure méthode.

	Méthode			t-test	
Jeu de données	Aléatoire	Incertitude	$\lambda = 0.2$	t	p-valeur
Breast Cancer	93.87	94.96	**95.31**	1.84	0.0669
Ionosphere	75.77	81.06	**82.40**	2.33	0.0210
Heart	67.89	67.72	**68.08**	0.29	0.7741
Parkinson	80.60	83.93	**84.85**	1.64	0.1034
Sonar	67.79	70.67	**70.94**	0.37	0.7089
Liver	**58.07**	57.37	58.02	0.07	0.9415
Dog-2	90.94	93.06	**94.10**	3.24	0.0014
Seeds	88.70	**89.93**	89.48	0.84	0.4010
Iris	88.22	**91.23**	90.60	1.19	0.2373
Wine	91.66	**93.55**	93.27	0.86	0.3920
Glass	57.33	58.32	**59.05**	0.87	0.3829
Ecoli	78.59	80.89	**81.98**	2.19	0.0300

(a) (b)

FIGURE 6.23 – Diagramme de différence critique pour différentes valeurs λ de la méthode proposée (6.23a) et pour l'échantillonnage aléatoire, l'échantillonnage par incertitude et la méthode proposée (6.23b).

figure 6.23a est obtenue en comparant la méthode proposée avec l'échantillonnage aléatoire et l'échantillonnage par incertitude. En moyenne, la méthode proposée obtient la position 1.33 sur 3 et la significativité statistique de ce résultat est démontrée par l'absence d'une ligne connectant les méthodes. Il peut être intéressant de noter que l'échantillonnage par incertitude ne possède pas non plus de connexion avec l'échantillonnage aléatoire, il s'agit déjà d'une méthode qui permet un gain significatif en performances comparé à l'échantillonnage aléatoire.

6.4 Bilan des méthodes d'échantillonnage

Le calcul de l'incertitude épistémique (non-crédibiliste) possède plusieurs étapes, et n'est pas forcément accessible. Il dépend des observations et il est nécessaire de passer au travers de plusieurs phases de calcul, d'estimation de vraisemblance, de maximum de vraisemblance et d'optimisation. Toute cette complexité comparée aux autres approches est présentée en annexe de chapitre, réservée aux détails expérimentaux. Avec les deux méthodes proposées, l'incertitude crédibiliste et l'incertitude épistémique crédibiliste, un simple calcul en sortie du modèle permet d'obtenir les niveaux d'incertitude.

Il y a évidemment une contrepartie à cette simplicité, et il s'agit du fait que le modèle doit être capable de retourner une fonction de masse pour représenter en sortie une incertitude et une imprécision. De tels modèles ne sont pas nombreux, et ont pour certains déjà été présentés dans ce document au chapitre 4. Parmi eux on compte le très célèbre modèle des K plus proches voisins crédibilistes, de Denœux 1995, les arbres de décision crédibilistes, de Trabelsi, Elouedi et al. 2019, Elouedi, Mellouli et al. 2001, et de Denoeux et Bjanger 2000, les forêts aléatoires crédibilistes, présentées dans ce document, ou encore des réseaux de neurones crédibilistes, de Yuan, Yue et al. 2020. Les méthodes proposées sont bien sûr compatibles avec des modèles probabilistes plus classiques [14] mais toute la profondeur liée à la modélisation de la croyance serait perdue.

Dans les expériences présentées, le paramètre λ se voir attribuer la valeur 0.2, ce qui signifie que le modèle va opter pour plus d'exploitation que d'exploration. C'est une valeur qui permet d'obtenir de bonnes performances. Nos études pour trouver quand il est plus intéressant de faire de l'exploration ou de l'exploitation, résumées sur la figure 6.23a, ne nous permettent pas pour l'instant de conclure davantage. Ces résultats montrent que plusieurs valeur de λ offrent des résultats similaires. La valeur 0.2 est celle qui, d'un point de vue général, offre les meilleures performances : pour la majorité des jeux de données, il est plus intéressant de faire de l'exploitation, sans dépasser un certain seuil, sinon quoi le modèle perdra en performances. Pour de prochains travaux, il serait intéressant d'être capable de modifier le paramètre λ au fil de l'eau, dans le but d'obtenir un modèle encore plus performant capable de choisir entre exploration et exploitation à la volée.

14. Une probabilité est une fonction de masse particulière.

6.5 Conclusion du chapitre

Ce chapitre s'intéresse à un niveau plus bas de modélisation que le chapitre précédent, en se focalisant sur l'échantillonnage par incertitude et sur la représentation des incertitudes du modèle.

Deux méthodes sont alors proposées, l'échantillonnage par incertitude crédibiliste et l'échantillonnage par incertitude épistémique crédibiliste. La première utilise l'incertitude de Klir, combinant la discorde et la non-spécificité et la seconde étend l'incertitude épistémique, et donc réductible, au cadre crédibiliste et à plusieurs classes, en simplifiant la phase calculatoire. Il ne s'agit pas pour ce dernier point de comparer des performances, mais de représenter une nouvelle information en échantillonnage par incertitude et malgré la différence de représentation, ces approches peuvent être appliquées identiquement à l'échantillonnage par incertitude. La contrainte est d'avoir accès à un modèle capable de prédire de manière incertaine et imprécise, en représentant son ignorance.

L'échantillonnage par incertitude crédibiliste est comparé au cours d'expérimentations à l'échantillonnage aléatoire et à l'échantillonnage par incertitude. Sa supériorité en terme de performance est statistiquement significative sur les jeux de données étudiés.

Il est aussi important de noter que l'ajout d'incertitude et d'imprécision dans un label permet, certes de moins altérer sa vraie nature et de gagner en performance, mais rend également cette étape plus sensible. La qualité de la labellisation, qui dépend de l'oracle et du modèle utilisé pour représenter l'imperfection, joue fortement sur les performances finales, pouvant faire varier les résultats de manière plus significative en améliorant la qualité des labels plutôt que la qualité du modèle lui-même.

Annexe du chapitre : expériences d'échantillonnage par incertitude

Cette annexe présente le détail des expériences réalisées en section 6.1 ainsi que les différents paramètres utilisés.

Echantillonage par incertitude

Les trois jeux de données sont générés selon les frontières de décision suivantes :

1. $y = -3x$, pour la figure 6.1a,
2. $y = \sin(0.8x)$, pour la figure 6.1b,
3. $(x-0)^2 + (y-0)^2 = 6$, pour la figure 6.1c.

Les incertitudes sont calculées sur une grille de 60*50 points (identique pour chaque expérience). Le modèle utilisé est celui des K plus proches voisins (K-NN), avec une prédiction probabiliste et la version pondérée par distance disponible avec scikit-learn (*cf.* Pedregosa, Varoquaux et al. 2011). Les paramètres non spécifiés sont ceux utilisés par défaut dans scikit-learn. Le modèle est entraîné sur les jeux de données et essaye de prédire la classe de chaque point de la grille en utilisant $K = 10$ voisins. Les mêmes paramètres sont utilisés pour le jeu Iris de Fisher.

Incertitudes épistémique et aléatoire

Le jeu de données présenté sur la figure 6.5a est généré selon la frontière d'équation $y = 0$ et les points sont distribués suivant un logarithme binaire log_2 avec plus de points sur la gauche.

Les incertitudes sont calculées avec la même version des K plus proches voisins. Le processus expérimental suit ce qui est présenté par Nguyen, Shaker et al. 2022. Les formules sont déduites de l'équation (6.4) comme suit :

$$\begin{aligned} \pi(1|x) &= \sup_{\theta \in \Theta} \min[\pi_\Theta(\theta), p_\theta(1|x) - p_\theta(0|x)], \\ \pi(0|x) &= \sup_{\theta \in \Theta} \min[\pi_\Theta(\theta), p_\theta(0|x) - p_\theta(1|x)]. \end{aligned} \tag{6.15}$$

Elles peuvent être réécrites en utilisant $f(a) = 2a - 1$:

$$\begin{aligned}\pi(1|x) &= \sup_{\theta\in\Theta} \min[\pi_\Theta(\theta), f(p_\theta(1|x))],\\ \pi(0|x) &= \sup_{\theta\in\Theta} \min[\pi_\Theta(\theta), f(1 - p_\theta(1|x))],\end{aligned} \tag{6.16}$$

avec :

$$\pi_\Theta(\theta) = \frac{L(\theta)}{L(\hat{\theta})}, \tag{6.17}$$

Pour une estimation par noyau, décrite par Nguyen, Shaker et al. 2022, avec p et n respectivement le nombre d'instances positives et négatives, la vraisemblance $L(\theta)$ est décrite par :

$$L(\theta) = \binom{p+n}{p}\theta^p(1-\theta)^n, \text{ and } \hat{\theta} = \frac{p}{p+n} \tag{6.18}$$

Nous n'utilisons pas d'estimation par noyau, mais une version pondérée par distance des K plus proches voisins, p est donc pris comme étant la somme des distances inverses pour les éléments positifs parmi les K voisins et n réciproquement pour la classe négative. On obtient :

$$\begin{aligned}\pi(1|x) &= \sup_{\theta\in[0,1]} \min\left(\frac{\theta^p(1-\theta)^n}{\left(\frac{p}{p+n}\right)^p\left(\frac{n}{p+n}\right)^n}, 2\theta - 1\right),\\ \pi(0|x) &= \sup_{\theta\in[0,1]} \min\left(\frac{\theta^p(1-\theta)^n}{\left(\frac{p}{p+n}\right)^p\left(\frac{n}{p+n}\right)^n}, 1 - 2\theta\right).\end{aligned} \tag{6.19}$$

La méthode de Brent est appliquée pour trouver un minimum local [15] dans l'intervalle $\theta \in [0, 1]$. Il est important de noter que pour ces expériences, les résultats ne servent pas à comparer un écart de performances, mais à supporter une nouvelle approche d'échantillonnage par incertitude. Dans notre cas, une formule plus appropriée a été déduite, c'est donc celle-là que nous avons utilisée dans les expériences :

$$\begin{aligned}\pi(1|x) &= \sup_{\theta\in[0,1]} \min\left(\frac{\theta^p(1-\theta)^n}{\left(\frac{p}{p+n}\right)^p\left(\frac{n}{p+n}\right)^n}, 2\theta - 1\right),\\ \pi(0|x) &= 1 - \sup_{\theta\in[0,1]} \min\left(\frac{\theta^p(1-\theta)^n}{\left(\frac{p}{p+n}\right)^p\left(\frac{n}{p+n}\right)^n}, 1 - 2\theta\right),\end{aligned} \tag{6.20}$$

15. Ici, il s'agit de trouver un maximum.

et pour les incertitudes épistémique et aléatoire :

$$\begin{aligned} \mathcal{U}_e(x) &= \min[\pi(1|x), \pi(0|x)] - 0.5, \\ \mathcal{U}_a(x) &= 1 - \max[\pi(1|x), \pi(0|x)]. \end{aligned} \tag{6.21}$$

La complexité calculatoire, la dépendance sur les observations, la phase d'optimisation, l'utilisation de la vraisemblance et du maximum de vraisemblance rendent la méthode relativement lourde. Pour les deux méthodes proposées, un simple calcul dépendant de la prédiction du modèle remplace toutes ces étapes.

Incertitude crédibiliste

Le jeu de données présenté sur la figure 6.11a est généré selon la frontière d'équation $y = 0$ et les points sont ajoutés dans l'espace excepté dans le sous-espace vérifiant $(x - 2.8)^2+(y-0)^2 < 5$ et avec imprécision dans le sous-espace vérifiant $(x-4.5)^2+(y-0)^2 < 5$.

Les incertitudes crédibilistes sont calculées selon le modèle des K plus proches voisins crédibilistes de Denœux 1995 déjà introduit au chapitre 4. Les paramètres utilisés sont ceux de la version γ-EKNN, avec $K = 10$ voisins, $\beta = 2$ et $\alpha = 1$ et la formule (6.8) est utilisée avec $\lambda = 0.5$, comme suggéré par Klir et Wierman 1998. Une différence importante est faite par rapport au modèle des K plus proches voisins crédibilistes, au lieu d'utiliser la règle de combinaison conjonctive (2.11), une moyenne des masses est utilisée, ce qui permet de passer α à 1 au lieu de $\alpha < 1$.

Incertitude épistémique crédibiliste

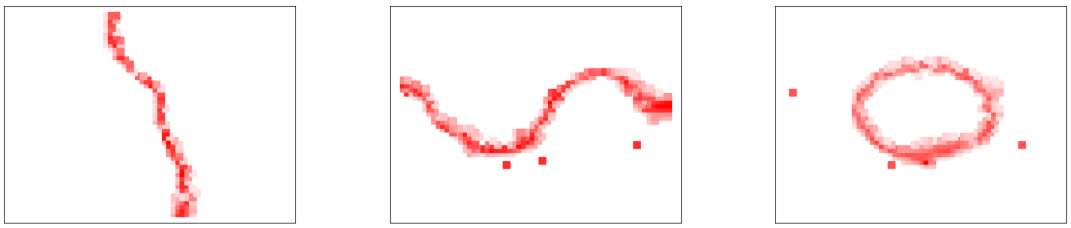
Le jeu de données présenté sur la figure 6.14a est généré selon trois blobs gaussiens avec scikit-learn (*cf.* Pedregosa, Varoquaux et al. 2011) et avec plus d'imprécision (ou d'ignorance) dans le sous espace vérifiant $(2.3 - x)^2 + (3 - y)^2 < 2$.

L'incertitude épistémique crédibiliste est calculée selon le même modèle des K plus proches voisins crédibilistes avec $K = 6$ voisins, $\beta = 2$, $\alpha = 1$ et la formule (6.14) est utilisée.

Non-régression des méthodes introduites

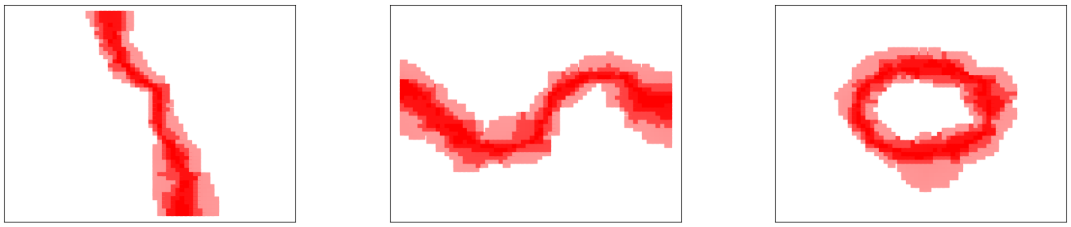
Cette section montre que pour les méthodes présentées, que ce soit l'incertitude épistémique, l'incertitude de Klir ou l'incertitude épistémique crédibiliste, il n'y a pas de

régression par rapport aux mesures classiques d'échantillonnage par incertitude d'équation (2.24) et (6.2). Les zones d'incertitude obtenues sont sensiblement les mêmes que sur la figure (6.2) quand les incertitudes épistémique, aléatoire et crédibiliste sont confondues et indiscernables. Les figures 6.24, 6.25 et 6.26 montrent ces zones d'incertitude pour les jeux de données introduits sur la figure 6.1. Les incertitudes épistémique, crédibiliste et épistémique crédibiliste donnent relativement les mêmes zones d'incertitude que celles utilisées sur la figure 6.2.



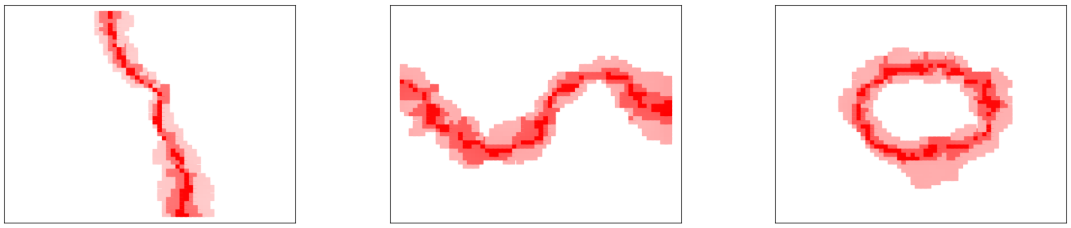
(a) (b) (c)

FIGURE 6.24 – De gauche à droite, les zones d'incertitude correspondant aux jeux de données des figures 6.1a, 6.1b et 6.1c selon l'incertitude épistémique.



(a) (b) (c)

FIGURE 6.25 – De gauche à droite, les zones d'incertitude correspondant aux jeux de données des figures 6.1a, 6.1b et 6.1c selon l'incertitude crédibiliste, résultante de la discorde et de la non-spécificité.



(a) (b) (c)

FIGURE 6.26 – De gauche à droite, les zones d'incertitude correspondant aux jeux de données des figures 6.1a, 6.1b et 6.1c selon l'incertitude épistémique crédibiliste.

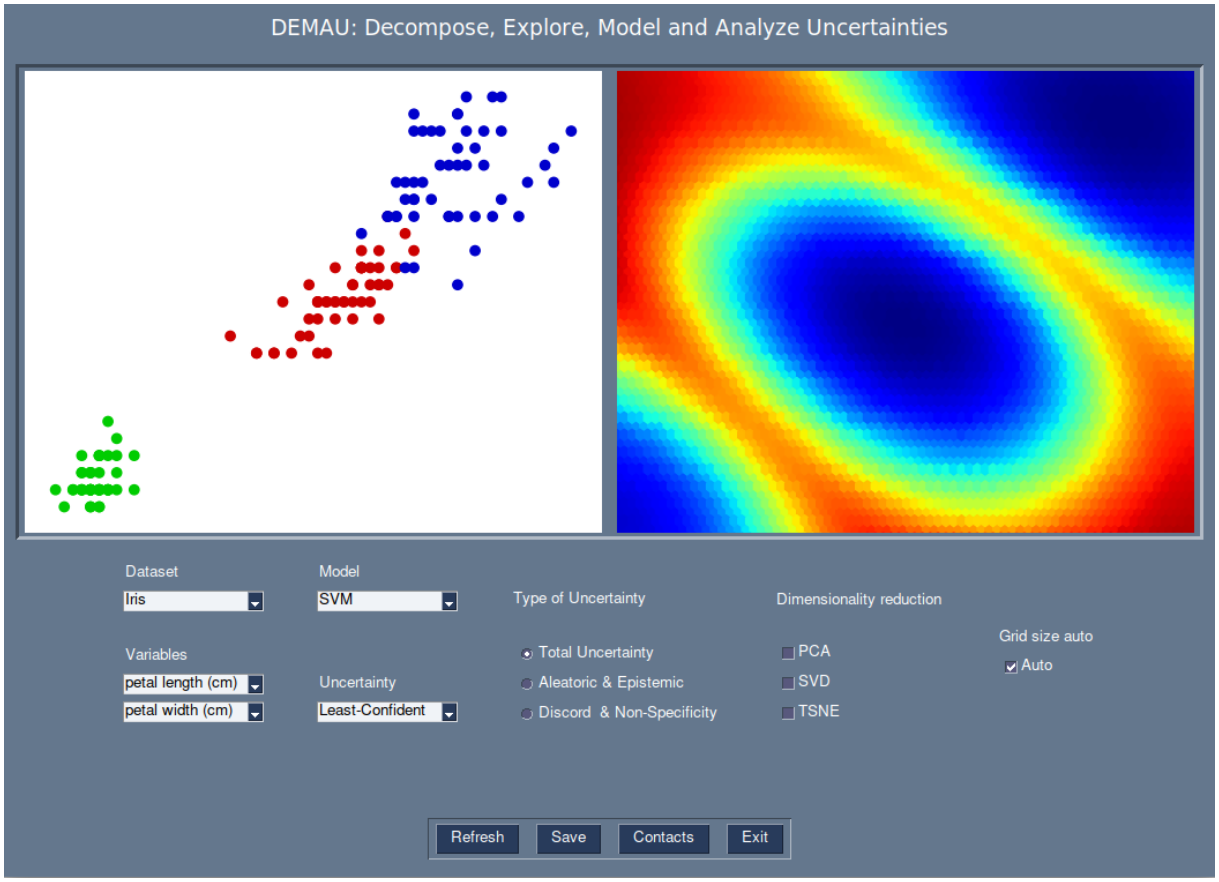


FIGURE 6.27 – DEMAU : Logiciel proposé et utilisé pour générer les zones d'incertitude de modèles d'apprentissage automatique (cf. Hoarau et Lemaire 2024).

Chapitre 7

Conclusion et perspectives de recherche

7.1 Conclusion

Ce document intitulé “Apprentissage actif de données incertaines et imprécises” financé par la Région Bretagne et le département Côtes-d'Armor retrace les activités de recherche développées durant un peu moins de trois années. Deux grandes thématiques ont été introduites, les fonctions de croyance, choisies pour modéliser l'incertitude et l'imprécision dans les données, ainsi que l'apprentissage actif, dont l'intérêt est de travailler sur un nombre réduit d'observations labellisées. La principale problématique étudiée est celle de la qualité et de la quantité des labels en apprentissage automatique. Le terme de “qualité” est choisi pour représenter une meilleure modélisation des labels et le terme de “quantité” fait référence à une réduction des coûts de labellisation. Le choix a été fait de travailler dans un premier temps sur chacun des sujets gravitant autour de la problématique, et d'apporter ensuite une solution globale qui permettrait d'y répondre. L'objectif de cette conclusion est également de présenter les nombreuses perspectives de recherche envisageables afin de poursuivre ces travaux. Les travaux réalisés sont succinctement résumés ci-dessous, et la section suivante présente les perspectives de recherche, d'abord en abordant chaque sujet étudié de manière isolée, puis en se concentrant sur la problématique globale.

Les jeux de données *Credal Dog-7*, *Credal Dog-4*, *Credal Dog-2*, *Credal Bird-10* et *Credal Bird-2* obtenus lors de campagnes de production participative ont été présentés. À notre connaissance, aucun autre jeu de données destiné à l'apprentissage automatique n'offre une représentation aussi exhaustive des imperfections dans les réponses des utilisateurs lors de l'étiquetage. Il s'agit donc d'une contribution visant à fournir à la communauté scientifique de nouveaux jeux de données labellisés de manière incertaine et imprécise pour combler une lacune dans la littérature. Ces jeux de données sont compatibles

non seulement avec les modèles d'apprentissage profond ou traditionnels, mais aussi avec des modèles probabilistes et plusieurs cadres de représentation de l'incertitude. Ils ont été décrits, publiés et mis à disposition en accès libre [1]. Leur utilisation a permis, entre autres dans ce document, de traiter de manière appliquée plusieurs aspects qui étaient jusqu'à présent abordés uniquement de manière théorique ou à l'aide de bruits synthétiques.

Jusqu'à présent, peu de modèles ont réussi à produire des prédictions incertaines et imprécises tout en tenant compte de l'imperfection déjà présente dans les données et en garantissant une interprétabilité aussi facile que celle des arbres de décision. Les propositions existantes dans la littérature concernant les arbres de décision crédibilistes souffrent de surapprentissage lorsque les données sont imparfaitement labellisées. Pour remédier à cela, une nouvelle version du modèle des K plus proches voisins crédibiliste a été développée, offrant un nouveau calcul pour l'un des paramètres et permettant de retrouver une équivalence avec le modèle original dans le cas de données labellisées parfaitement. Deux nouveaux modèles ont également été introduits, à savoir les arbres de décision crédibilistes et les forêts aléatoires crédibilistes, tous deux capables de traiter des labels riches et de produire des prédictions incertaines et imprécises. La robustesse de ces modèles a été démontrée expérimentalement, et les résultats montrent la pertinence d'utiliser ces nouvelles approches en apprentissage automatique.

En ce qui concerne la problématique centrale de la thèse, des expériences ont été menées pour examiner l'impact de l'ajout d'informations lors de la phase de labellisation sur les performances de classification, en utilisant le moins de données labellisées possible. Ces premiers résultats, issus d'une démarche exploratoire, ont montré la pertinence de cette approche. Par la suite, deux méthodes d'échantillonnage ont été développées : l'échantillonnage par incertitude crédibiliste et l'échantillonnage par incertitude épistémique crédibiliste. Ces approches offrent une nouvelle représentation de l'incertitude du modèle en tenant compte de celle déjà présente dans les labels. En plus de permettre d'intégrer l'incertitude des labels, de simplifier le processus de calcul et d'adresser le dilemme d'exploration-exploitation, ces méthodes améliorent les performances de l'apprentissage actif sur des jeux de données classiques. Une analyse expérimentale a confirmé la supériorité statistique de ces méthodes par rapport à l'échantillonnage par incertitude. Elles sont applicables à un large éventail de problèmes d'apprentissage automatique et permettent une réduction significative des coûts sur de nombreux jeux de données. De plus, aucune connaissance préalable de la théorie des fonctions de croyance n'est requise, car

1. Voir le chapitre "Logiciels et Reproductibilité".

les outils développés peuvent s'intégrer de manière presque transparente avec les systèmes existants, offrant ainsi une réduction simple des coûts de labellisation.

D'autres travaux connexes ont également été réalisés, notamment la création de DEMAU (*cf.* Hoarau et Lemaire 2024), un outil éducatif, exploratoire, analytique et libre d'accès permettant de visualiser et d'explorer plusieurs types d'incertitudes pour les modèles de classification en apprentissage automatique. De plus, une étude non biaisée réalisée par Hoarau, Sale et al. 2024 a permis de fournir des résultats tangibles, notamment en mettant en évidence la corrélation entre les mesures d'incertitude épistémique et aléatoire à travers différents cadres théoriques.

Cette thèse se clôture avec les différents éléments qui ont été présentés dans ce document et les multiples contributions mises à disposition de la communauté scientifique mais la pluralité des sujets évoqués et leur possible application dans diverses branches de la recherche fondamentale et applicative permettent d'ouvrir vers de nombreuses perspectives.

7.2 Perspectives de recherche

7.2.1 Perspectives liées à la labellisation imparfaite

Cinq jeux de données ont été proposés et ont permis de combler un manque dans la littérature, leur obtention lors de campagnes de production participative est issue de travaux communs entrepris par Thierry, Hoarau et al. 2022. L'interface elle-même, utilisée pour obtenir les résultats de Hoarau, Thierry, Martin et al. 2023 peut être améliorée, la difficulté étant de laisser la possibilité à un utilisateur de représenter l'imperfection de sa réponse, sans rendre le processus trop contraignant. La qualification des contributeurs (voir Thierry, Martin et al. 2023) constitue aussi une piste de recherche intéressante, qui conditionne la qualité des labels utilisés dans toutes les parties liées à cette thèse. En effet, apporter un poids moindre à un contributeur moins assidu peut permettre d'améliorer la qualité des labels qui s'avèrent, comme nous l'avons vu dans ce document, parfois plus importante que le modèle même.

7.2.2 Perspectives liées à l'apprentissage automatique

Plusieurs modèles de classification crédibilistes ont été présentés et certains ont été proposés au travers de ces travaux (*cf.* Hoarau, Martin, J.-C. Dubois et Le Gall 2022 ; Hoarau,

Martin, J.-C. Dubois et Le Gall 2023). Ces modèles permettent de faire une prédiction incertaine et imprécise, mais dans le cadre de cette thèse, ils permettent également de prendre en compte l'imperfection déjà présente dans les labels. Une des perspectives serait de travailler avec d'autres classifieurs crédibilistes, ou d'autres modèles ne prenant pas en compte des données labellisées imparfaitement, afin de les rendre compatibles avec la théorie des fonctions de croyance et d'observer si, comme pour K-NN, les arbres de décision et les forêts aléatoires, une version crédibiliste peut s'avérer pertinente. Plusieurs travaux ont déjà été menés dans ce sens, avec de nombreux algorithmes qui existent en version crédibiliste, comme des variantes de l'algorithme EM, de Denœux 2011 et de Vannoorenberghe 2007, ou encore une version crédibiliste des réseaux de neurones, de Z. Tong, Xu et al. 2021.

Par ailleurs, le sujet de la classification a été exclusivement traité dans ce document mais l'apprentissage supervisé se compose également de régression. De nombreux parallèles peuvent être faits en ouvrant chacune des problématiques étudiées à la régression crédibiliste (Amini, Schwarting et al. 2020).

7.2.3 Perspectives liées à l'apprentissage actif sur données imparfaites

La principale contribution liée à la problématique constitue la proposition d'échantillonnage par incertitude crédibiliste, adressant le dilemme d'exploration-exploitation et permettant de prendre en compte des labels riches, tout en augmentant les performances en apprentissage actif. Il s'agit d'un large sujet sur lequel une formalisation serait bienvenue, principalement sur la différence fondamentale entre les incertitudes liées à un modèle probabiliste sur labels durs et celles liées à un modèle crédibiliste sur labels riches. Le cœur de la méthode proposée repose sur un paramètre λ qui sert de curseur entre exploration et exploitation. Plusieurs expériences ont été conduites mais aucune n'a encore permis d'atteindre un niveau suffisant d'automatisation concernant l'ajustement de ce paramètre. La valeur 0.2 qui est choisie arbitrairement, offre des résultats très compétitifs. Cette valeur peut se traduire par "le modèle opte pour plus d'exploitation que d'exploration". Cependant, une méthode qui serait capable de modifier ce paramètre au fil de l'eau et en fonction du jeu de données offrirait probablement des performances encore meilleures.

Seul l'échantillonnage par incertitude est étudié en détail dans ce document mais il existe d'autres méthodes d'échantillonnage en apprentissage actif (voir chapitre 2.2.4). Un

échantillonnage par comité de modèles crédibilistes au sein de laquelle plusieurs modèles sont entraînés en parallèle sur les mêmes données semble être une approche qui pourrait être pertinente. Le désaccord est ensuite mesuré pour sélectionner l'instance à labelliser. Il serait alors possible de travailler avec des modèles crédibilistes qui ne donnent plus une prédiction sur Ω mais sur 2^{Ω}. D'autres méthodes peuvent être mises en place, dont l'utilisation ou non des entropies crédibilistes pour calculer le désaccord entre les modèles, ou encore mesurer le conflit avec la théorie des fonctions de croyance.

La comparaison de la méthode proposée avec des modèles récents de décomposition d'incertitudes est sans doute l'une des perspectives les plus cohérentes, et constitue les travaux qui sont conduits en ce moment, prolongeant directement les résultats présentés ici. Les travaux récents de Hüllermeier, Destercke et al. 2022 et de Nguyen, Shaker et al. 2022 sont, à l'heure actuelle et à mon sens, ceux qui méritent d'être explorés, pour ce qui est exclusivement des thématiques liées à la thèse défendue dans ce document.

Certaines de ces perspectives sont déjà en cours d'exploration et ne sont pas détaillées dans ce document car elles dépassent les limites de la problématique. En ce qui concerne la représentation et la quantification de l'incertitude, des travaux ont été entrepris en collaboration avec l'université Ludwig-Maximilian afin de proposer une étude quantitative non biaisée de plusieurs outils de quantification d'incertitude à travers différents cadres théoriques de représentation, comme décrit par Hoarau, Sale et al. 2024. Pour ce qui est de l'apprentissage actif (voir Hoarau, Shaker et al. 2024), une critique des outils actuels utilisés pour séparer l'incertitude épistémique et aléatoire est également proposée, et introduit un nouveau cadre déterministe permettant de capturer une estimation plus fine des différents types d'incertitudes. De plus, un projet en collaboration avec le laboratoire IRIT est envisagé pour explorer la partie non supervisée de l'apprentissage automatique, ce qui complète les travaux menés ici qui se concentrent sur l'apprentissage supervisé.

En conclusion, ce travail ouvre de nombreuses perspectives de recherche dans les domaines de la labellisation imparfaite, de l'apprentissage automatique et de l'apprentissage actif. Il souligne l'importance de prendre en compte l'imperfection dans les données pour développer des modèles plus efficaces et robustes en intelligence artificielle. Par ailleurs, il est essentiel de veiller à réduire les coûts économiques, écologiques et sociétaux associés aux méthodes d'apprentissage automatique. À titre d'exemple, le recours à des travailleurs kenyans payés moins de 2 dollars de l'heure par OpenAI pour l'étiquetage de contenu toxique, tel que rapporté par Perrigo 2023 dans le magazine TIME, souligne la nécessité de développer des approches plus soucieuses de ces problématiques.

Tenez-vous sur vos gardes, philosophes et amis de la connaissance ! Évitez le martyre, évitez de souffrir "pour l'amour de la vérité" ! Évitez même de vous défendre ! Vous ne ferez que ruiner toute l'innocence et l'impartialité supérieure de votre conscience, vous raidir contre les objections et les provocations, vous transformer en sots, en bêtes, en bœufs, si, alors que vous êtes déjà aux prises avec le danger, la calomnie, le soupçon, l'ostracisme et aux autres séquelles, plus brutales encore, de la haine, vous avez au surplus à vous poser en défenseurs de la vérité sur la terre. Comme si la "vérité" était désarmée et faible au point d'avoir besoin de défenseurs ! Et justement de vous, lugubres paladins de l'esprit, qui vous rencognez dans votre trou pour filer vos toiles d'araignées !

Friedrich Nietzsche, *Par-delà bien et mal.*

PUBLICATIONS

Articles publiés

— Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2022), « Imperfect Labels with Belief Functions for Active Learning », in : *Belief Functions : Theory and Applications*, p. 44-53

— Constance Thierry, Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2022), « Real bird dataset with imprecise and uncertain values », in : *Belief Functions : Theory and Applications*, p. 275-285

— Arthur Hoarau, Constance Thierry, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2023), « Datasets with Rich Labels for Machine Learning », in : *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, p. 1-6

— Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2023), « Evidential Random Forests », in : *Expert Systems with Applications* 230

— Arthur Hoarau, Vincent Lemaire, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2024b), « Méthode crédibiliste pour l'extraction d'incertitudes sans dépendance aux observations », in : *Extraction et Gestion des Connaissances (EGC)*

— Arthur Hoarau, Vincent Lemaire, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2024a), « Evidential uncertainties and rich labels for active learning », in : *Machine Learning*

— Arthur Hoarau, Constance Thierry, Jean-Christophe Dubois et Yolande Le Gall (2024), « A mean distance between elements of same class for rich labels », in : *Belief Functions : Theory and Applications*, Belfast, United Kingdom

En cours de publication et manuscrits-auteur

— Arthur Hoarau, Yusuf Sale, Paul Hofman et Eyke Hüllermeier (2024), « Quantitative analysis of uncertainty measures », in : *On demand*

— Arthur Hoarau, Mohammad Hossein Shaker et Eyke Hüllermeier (2024), « Efficient reduction of uncertainty by focusing on real epistemic uncertainty », in : *On demand*

Logiciels et Reproductibilité

Accès aux codes sources

Jeux de données *Credal*

`https://data.mendeley.com/datasets/4hz3wx6wm5`

Arthur Hoarau, Constance Thierry, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2023), « Datasets with Rich Labels for Machine Learning », in : *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, p. 1-6

Modèle des K plus proches voisins crédibiliste

`https://github.com/ArthurHoa/imperfect-eknn`

Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2022), « Imperfect Labels with Belief Functions for Active Learning », in : *Belief Functions : Theory and Applications*, p. 44-53

Arbre de décision crédibiliste et forêt aléatoire crédibiliste

`https://github.com/ArthurHoa/conflict-edt`

`https://github.com/ArthurHoa/evidential-random-forest`

Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2023), « Evidential Random Forests », in : *Expert Systems with Applications* 230

Echantillonnage par incertitudes crédibilistes

`https://github.com/ArthurHoa/evidential-uncertainty-sampling`

Arthur Hoarau, Vincent Lemaire, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2024b), « Méthode crédibiliste pour l'extraction d'incertitudes sans dépendance aux observations », in : *Extraction et Gestion des Connaissances (EGC)*

Apprentissage actif crédibiliste

`https://github.com/ArthurHoa/evidential-active-learning`

Arthur Hoarau, Vincent Lemaire, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2024a), « Evidential uncertainties and rich labels for active learning », in : *Machine Learning*

Logiciel - DEMAU : Decompose, Explore, Model and Analyze Uncertainties

`https://github.com/ArthurHoa/DEMAU`

Arthur Hoarau et Vincent Lemaire (2024), « DEMAU : Decompose, Explore, Model and Analyse Uncertainties », in : *Soumis à ECMLPKDD*

BIBLIOGRAPHIE

Abe, Naoki, Bianca Zadrozny et John Langford (août 2006), « Outlier detection by active learning », in : *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2006, p. 504-509.

Abiodun, Oludare Isaac, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed et Humaira Arshad (2018), « State-of-the-art in artificial neural network applications : A survey », in : *Heliyon* 4.*11*.

Aggarwal, Charu, Xiangnan Kong, Quanquan Gu, Jiawei Han et Philip Yu (2014), *Active Learning : A Survey, Data Classification : Algorithms and Applications*, CRC Press.

Amini, Alexander, Wilko Schwarting, Ava Soleimany et Daniela Rus (2020), « Deep Evidential Regression », in : *Advances in Neural Information Processing Systems*, sous la dir. de H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan et H. Lin, t. 33, Curran Associates, Inc., p. 14927-14937.

Amit, Yali et Donald Geman (oct. 1997), « Shape Quantization And Recognition With Randomized Trees. », in : *Neural Computation* 9, p. 1545-1588.

Bondu, Alexis et Vincent Lemaire (oct. 2008), « Etat de l'art sur les méthodes statistiques d'apprentissage actif », in : *Revue des Nouvelles Techniques de l'Information*.

Bondu, Alexis, Vincent Lemaire et Marc Boullé (2010), « Exploration vs. exploitation in active learning : A Bayesian approach », in : *The 2010 International Joint Conference on Neural Networks (IJCNN)*, p. 1-7.

Bramer, Max (jan. 2013), « Avoiding Overfitting of Decision Trees », in : *Principles of Data Mining*, Springer London, p. 121-136.

Breiman, L. (1996), « Bagging predictors », in : *Machine Learning* 24, p. 123-140.

Breiman, Leo (oct. 2001), « Random Forests », in : *Machine Learning* 45, p. 5-32.

Breiman, Leo, Jerome Friedman, Charles J. Stone et R.A. Olshen (1984), *Classification and Regression Trees*, Taylor & Francis.

Chapelle, Olivier et Alexander Zien (2005), « Semi-Supervised Classification by Low Density Separation », in : *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, sous la dir. de Robert G. Cowell et Zoubin Ghahramani,

Proceedings of Machine Learning Research, Reissued by PMLR on 30 March 2021., PMLR, p. 57-64.

Charpentier, Bertrand, Daniel Zügner et Stephan Günnemann (2020), « Posterior Network : Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts », in : *Advances in Neural Information Processing Systems*, sous la dir. de H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan et H. Lin, t. 33, Curran Associates, Inc., p. 1356-1367.

Cohn, David, Rich Caruana et Andrew McCallum (jan. 2008), « Semi-supervised clustering with user feedback », in : *Constrained Clustering : Advances in Algorithms, Theory, and Applications*, p. 17-31.

Cohn, David A., Zoubin Ghahramani et Michael I. Jordan (1996), « Active Learning with Statistical Models », in : *Journal of Artificial Intelligence Research* 4.*1*, p. 129-145.

Côme, Etienne, Latifa Oukhellou, Thierry Denoeux et Patrice Aknin (2009), « Learning from partially supervised data using mixture models and belief functions », in : *Pattern Recognition* 42.*3*, p. 334-348.

Dempster, Arthur P. (1967), « Upper and Lower Probabilities Induced by a Multivalued Mapping », in : *The Annals of Mathematical Statistics* 38.*2*, p. 325-339.

Demšar, Janez (2006), « Statistical comparisons of classifiers over multiple data sets », in : *The Journal of Machine learning research* 7, p. 1-30.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li et Li Fei-Fei (2009), « ImageNet : A large-scale hierarchical image database », in : *2009 IEEE Conference on Computer Vision and Pattern Recognition*, p. 248-255.

Denœux, Thierry (1995), « A k-Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory », in : *IEEE Transactions on Systems, Man, and Cybernetics* 219, p. 804-813.

— (août 2006), « The cautious rule of combination for belief functions and some extensions », in : *2006 9th International Conference on Information Fusion, FUSION*, p. 1-8.

— (jan. 2011), « Maximum Likelihood Estimation from Uncertain Data in the Belief Function Framework », in : *Knowledge and Data Engineering, IEEE Transactions on* 25.

— (2019), « Decision-making with belief functions : A review », in : *International Journal of Approximate Reasoning* 109, p. 87-110.

— (2021), « Belief functions induced by random fuzzy sets : A general framework for representing uncertain and fuzzy evidence », in : *Fuzzy Sets and Systems* 424, Uncertainty, p. 63-91.

Denœux, Thierry et Lalla Zouhal (sept. 2001), « Handling possibilistic labels in pattern classification using evidential reasoning », in : *Fuzzy Sets and Systems* 122, p. 409-424.

Denoeux et Bjanger (2000), « Induction of decision trees from partially classified data using belief functions », in : *Systems, Man, and Cybernetics, 2000 IEEE International Conference* 4, p. 2923-2928.

Denoeux, Thierry, Orakanya Kanjanatarakul et Songsak Sriboonchitta (2019), « A New Evidential K-Nearest Neighbor Rule based on Contextual Discounting with Partially Supervised learning », in : *International Journal of Approximate Reasoning* 113, p. 287-302.

Dua, Dheeru et Casey Graff (2017), *UCI Machine Learning Repository*, URL : `http://archive.ics.uci.edu/ml`.

Dubois, Didier et Henri Prade (1988), « Representation and combination of uncertainty with belief functions and possibility measures », in : *Computational Intelligence* 4.*3*, p. 244-264.

— (août 2001), « Possibility Theory, Probability Theory and Multiple-Valued Logics : A Clarification », in : *Annals of Mathematics and Artificial Intelligence* 32, p. 35-66.

Dudani, Sahibsingh A. (1976), « The Distance-Weighted k-Nearest-Neighbor Rule », in : *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6.*4*, p. 325-327.

Elouedi, Zied, Khaled Mellouli et Philippe Smets (2001), « Belief decision trees : theoretical foundations », in : *International Journal of Approximate Reasoning* 28.*2*, p. 91-124.

Essaid, Amira, Arnaud Martin, Grégory Smits et Boutheina Ben Yaghlane (2014), « Uncertainty in Ontology Matching : A Decision Rule-Based Approach », in : *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, sous la dir. d'Anne Laurent, Olivier Strauss, Bernadette Bouchon-Meunier et Ronald R. Yager, Springer International Publishing, p. 46-55.

Fiche, Anthony, Arnaud Martin, Jean-Christophe Cexus et Ali Khenchaf (juill. 2010), « Continuous belief functions and α-stable distributions », in : *Information Fusion (FUSION), 2010 13th Conference on*, Information Fusion (FUSION), Edinburgh, United Kingdom, 7 pages.

Fix, Evelyn et J. L. Hodges (1951), « Discriminatory Analysis. Nonparametric Discrimination : Consistency Properties », in : *International Statistical Review / Revue Internationale de Statistique* 57.*3*, p. 238-247.

Fletcher, R. (2000), « Newton-Like Methods », in : *Practical Methods of Optimization*, John Wiley & Sons, Ltd, chap. 3, p. 44-79.

Fredriksson, Teodor, David Issa Mattos, Jan Bosch et Helena Olsson (nov. 2020), *Data Labeling : An Empirical Investigation into Industrial Challenges and Mitigation Strategies*, p. 202-216.

Hacohen, Guy, Avihu Dekel et Daphna Weinshall (2022), « Active Learning on a Budget : Opposite Strategies Suit High and Low Budgets », in : *International Conference on Machine Learning, 2022, Baltimore, Maryland, USA*, sous la dir. de Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu et Sivan Sabato, t. 162, Proceedings of Machine Learning Research, PMLR, p. 8175-8195.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren et Jian Sun (juin 2016), « Deep Residual Learning for Image Recognition », in : *IEEE Conference on Computer Vision and Pattern Recognition*, p. 770-778.

Hemmer, Patrick, Niklas Kühl et Jakob Schöffer (2020), « DEAL : Deep Evidential Active Learning for Image Classification », in : *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, p. 865-870.

Hoarau, Arthur et Vincent Lemaire (2024), « DEMAU : Decompose, Explore, Model and Analyse Uncertainties », in : *Soumis à ECMLPKDD*.

Hoarau, Arthur, Vincent Lemaire, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2024a), « Evidential uncertainties and rich labels for active learning », in : *Machine Learning*.

— (2024b), « Méthode crédibiliste pour l'extraction d'incertitudes sans dépendance aux observations », in : *Extraction et Gestion des Connaissances (EGC)*.

Hoarau, Arthur, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2022), « Imperfect Labels with Belief Functions for Active Learning », in : *Belief Functions : Theory and Applications*, p. 44-53.

— (2023), « Evidential Random Forests », in : *Expert Systems with Applications* 230.

Hoarau, Arthur, Yusuf Sale, Paul Hofman et Eyke Hüllermeier (2024), « Quantitative analysis of uncertainty measures », in : *On demand*.

Hoarau, Arthur, Mohammad Hossein Shaker et Eyke Hüllermeier (2024), « Efficient reduction of uncertainty by focusing on real epistemic uncertainty », in : *On demand*.

Hoarau, Arthur, Constance Thierry, Jean-Christophe Dubois et Yolande Le Gall (2024), « A mean distance between elements of same class for rich labels », in : *Belief Functions : Theory and Applications*, Belfast, United Kingdom.

Hoarau, Arthur, Constance Thierry, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2023), « Datasets with Rich Labels for Machine Learning », in : *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, p. 1-6.

Hora, Stephen C. (1996), « Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management », in : *Reliability Engineering & System Safety* 54.*2*, Treatment of Aleatory and Epistemic Uncertainty, p. 217-223.

Hüllermeier, Eyke, Sébastien Destercke et Mohammad Hossein Shaker (2022), « Quantification of Credal Uncertainty in Machine Learning : A Critical Analysis and Empirical Comparison », in : *Conference on Uncertainty in Artificial Intelligence*.

Hüllermeier, Eyke et Willem Waegeman (2021), « Aleatoric and Epistemic Uncertainty in Machine Learning : An introduction to concepts and methods », in : *Machine Learning* 110, p. 457-506.

Jousselme, Anne-Laure, Dominic Grenier et Éloi Bossé (2001), « A new distance between two bodies of evidence », in : *Information Fusion* 2.*2*, p. 91-101.

Kendall, Alex et Yarin Gal (2017), « What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision ? », in : *31st International Conference on Neural Information Processing Systems (NIPS)*.

Khosla, Aditya, Nityananda Jayadevaprakash, Bangpeng Yao et Li Fei-Fei (2011), « Novel Dataset for Fine-Grained Image Categorization », in : *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO.

Klir, George J. et Mark J. Wierman (1998), « Uncertainty-based information : Elements of generalized information theory. », in : *Springer-Verlag*.

Lee, S.K. (1992), « Imprecise and uncertain information in databases : an evidential approach », in : *Eighth International Conference on Data Engineering*, p. 614-621.

Lefevre, Eric, Olivier Colot et Patrick Vannoorenberghe (juin 2002), « Belief function combination and conflict management », in : *Information Fusion* 3.*2*, p. 149-162.

Lewis, David D. et William A. Gale (1994), « A Sequential Algorithm for Training Text Classifiers », in : *SIGIR 94*, sous la dir. de Bruce W. Croft et C. J. van Rijsbergen, London : Springer London, p. 3-12.

Martens, Timo, Lorenzo Perini et Jesse Davis (2023), « Semi-Supervised Learning From Active Noisy Soft Labels For Anomaly Detection », in : *Machine Learning and Knowledge Discovery in Databases : Research Track : European Conference, ECML PKDD 2023, Turin, Italy*, Turin, Italy : Springer-Verlag, p. 219-236.

Martin, Arnaud (2019), « Conflict management in information fusion with belief functions », in : *Information quality in information fusion and decision making*, sous la dir. d'Eloi Bossé et Galina L. Rogova, Information Fusion and Data Science, p. 79-97.

Nguyen, Vu-Linh, Mohammad Hossein Shaker et Eyke Hüllermeier (2022), « How to Measure Uncertainty in Uncertainty Sampling for Active Learning », in : *Machine Learning* 111, p. 89-122.

Pan, Qian, Deyun Zhou, Yongchuan Tang, Xiaoyang Li et Jichuan Huang (2019), « A Novel Belief Entropy for Measuring Uncertainty in Dempster-Shafer Evidence Theory Framework Based on Plausibility Transformation and Weighted Hartley Entropy », in : *Entropy* 21.*2*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot et E. Duchesnay (2011), « Scikit-learn : Machine Learning in Python », in : *Journal of Machine Learning Research* 12, p. 2825-2830.

Perrigo, Billy (jan. 2023), *Exclusive : The $2 Per Hour Workers Who Made ChatGPT Safer*, en, URL : `https://time.com/6247678/openai-chatgpt-kenya-workers/` (visité le 23/02/2023).

Quinlan, J.R. (1987), « Simplifying decision trees », in : *International Journal of Man-Machine Studies* 27.*3*, p. 221-234.

— (1993), *C4.5 : Programs for Machine Learning*, Morgan Kaufmann series in machine learning, Elsevier Science.

Ramel, Sébastien, Frédéric Pichon et François Delmotte (2018), « Active Evidential Calibration of Binary SVM Classifiers », in : *Belief Functions : Theory and Applications*, sous la dir. de Sébastien Destercke, Thierry Denoeux, Fabio Cuzzolin et Arnaud Martin, Springer International Publishing, p. 208-216.

Roh, Yuji, Geon Heo et Steven Euijong Whang (2021), « A Survey on Data Collection for Machine Learning : A Big Data - AI Integration Perspective », in : *IEEE Transactions on Knowledge and Data Engineering* 33.*4*, p. 1328-1347.

Russell, Stuart et Peter Norvig (2010), *Artificial Intelligence : A Modern Approach*, 3e éd., Prentice Hall.

Samet, Ahmed, Éric Lefèvre et Sadok Ben Yahia (2014), « Evidential Database : A New Generalization of Databases ? », in : *Belief Functions : Theory and Applications*, sous la dir. de Fabio Cuzzolin, Springer International Publishing, p. 105-114.

Schmarje, Lars, Johannes Brünger, Monty Santarossa, Simon-Martin Schröder, Rainer Kiko et Reinhard Koch (2021), « Fuzzy Overclustering : Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy », in : *Sensors* 21.*19*.

Schmarje, Lars, Claudius Zelenka, Ulf Geisen, Claus-C Glüer et Reinhard Koch (oct. 2019), « 2D and 3D Segmentation of Uncertain Local Collagen Fiber Orientations in SHG Microscopy », in : *Pattern Recognition, 41st DAGM German Conference*, p. 374-386.

Senge, Robin, Stefan Bösner, Krzysztof Dembczynski, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff et Eyke Hüllermeier (2014), « Reliable classification : Learning classifiers that distinguish aleatoric and epistemic uncertainty », in : *Information Sciences* 255, p. 16-29.

Settles, Burr (2009), *Active Learning Literature Survey*, Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Seung, H. S., M. Opper et H. Sompolinsky (1992), « Query by Committee », in : *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, Pittsburgh, Pennsylvania, USA : Association for Computing Machinery, p. 287-294.

Shafer, Glenn (1976), *A Mathematical Theory of Evidence*, Princeton : Princeton University Press.

Shannon, Claude. E. (1948), « A Mathematical Theory of Communication », in : *Bell System Technical Journal* 27.*3*, p. 379-423.

Siciliano, Roberta (1998), « Exploratory Versus Decision Trees », in : *COMPSTAT*, sous la dir. de Roger Payne et Peter Green, Heidelberg : Physica-Verlag HD, p. 113-124.

Singh, Aarti, Robert Nowak et Parmesh Ramanathan (2006), « Active learning for adaptive mobile sensing networks », in : *2006 5th International Conference on Information Processing in Sensor Networks*, p. 60-68.

Smets, P. (1990), « The combination of evidence in the transferable belief model », in : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.*5*, p. 447-458.

Smets, Philippe (1997), « Imperfect Information : Imprecision and Uncertainty », in : *Uncertainty Management in Information Systems : From Needs to Solutions*, sous la dir. d'Amihai Motro et Philippe Smets, Boston, MA : Springer US, p. 225-254.

Smets, Philippe (2005), « Belief functions on real numbers », in : *International Journal of Approximate Reasoning* 40.*3*, p. 181-223.

Smets, Philippe et Robert Kennes (1994), « The transferable belief model », in : *Artificial Intelligence* 66.*2*, p. 191-234.

Steinhaus, Hugo (1957), « Sur la division des corps matériels en parties », in : *Bulletin L'Académie Polonaise des Science*.

Strat, Thomas M. (1984), « Continuous Belief Functions for Evidential Reasoning », in : *AAAI*.

Sugeno, M. (1993), « Fuzzy Measures and Fuzzy Integrals—A Survey », in : *Readings in Fuzzy Sets for Intelligent Systems*, sous la dir. de Didier Dubois, Henri Prade et Ronald R. Yager, Morgan Kaufmann, p. 251-257.

Thierry, Constance, Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (2022), « Real bird dataset with imprecise and uncertain values », in : *Belief Functions : Theory and Applications*, p. 275-285.

Thierry, Constance, Arnaud Martin, Jean-Christophe Dubois et Yolande Le Gall (oct. 2021), « Validation of Smets' hypothesis in the crowdsourcing environment », in : *6th International Conference on Belief Functions*, Shanghai, China.

— (2023), « Estimation of the qualification and behavior of a contributor and aggregation of his answers in a crowdsourcing context », in : *Expert Systems with Applications* 216.

ThinkML-Team (mars 2022), *Top AI Achievements of 2021*, en, URL : `https://thinkml.ai/top-ai-achievements-of-2021/` (visité le 23/02/2023).

Tong, Simon et Daphne Koller (2002), « Support Vector Machine Active Learning with Applications to Text Classification », in : *J. Mach. Learn. Res.* 2, p. 45-66.

Tong, Zheng, Philippe Xu et Thierry Denœux (2021), « An evidential classifier based on Dempster-Shafer theory and deep learning », in : *Neurocomputing* 450, p. 275-293.

Trabelsi, Asma, Zied Elouedi et Eric Lefevre (juill. 2019), « Decision tree classifiers for evidential attribute values and class labels », in : *Fuzzy Sets and Systems* 366, p. 46-62.

Vannoorenberghe, Patrick (2007), « Estimation de modèles de mélanges finis par un algorithme EM crédibiliste », in : *Traitement Du Signal* 24, p. 103-113.

Willett, Rebecca, Robert Nowak et Rui Castro (2005), « Faster Rates in Regression via Active Learning », in : *Advances in Neural Information Processing Systems*, sous la dir. d'Y. Weiss, B. Schölkopf et J. Platt, t. 18, MIT Press.

Yager, Ronald R. (1987), « On the dempster-shafer framework and new combination rules », in : *Information Sciences* 41.*2*, p. 93-137.

Yuan, Bin, Xiaodong Yue, Ying Lv et Thierry Denoeux (août 2020), « Evidential Deep Neural Networks for Uncertain Data Classification », in : *Knowledge Science, Engineering and Management (Proceedings of KSEM 2020)*, Lecture Notes in Computer Science, Springer Verlag.

Zadeh, L.A. (1965), « Fuzzy sets », in : *Information and Control* 8.*3*, p. 338-353.

— (1978), « Fuzzy sets as a basis for a theory of possibility », in : *Fuzzy Sets and Systems* 1.*1*, p. 3-28.

Zhang, Haifei, Benjamin Quost et Marie-Hélène Masson (2023), « Cautious weighted random forests », in : *Expert Systems with Applications* 213.

Zhu, Daniel, Arnaud Martin, Jean-Christophe Dubois, Yolande Le Gall et Vincent Lemaire (oct. 2021), « Modèle crédibiliste pour l'échantillonnage en apprentissage actif », in : *Rencontres francophones sur la logique floue et ses applications*, Paris, France.

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong et Qing He (2021), « A Comprehensive Survey on Transfer Learning », in : *Proceedings of the Institute of Radio Engineers* 109, p. 43-76.

COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

Titre : Apprentissage actif de données incertaines et imprécises

Mot clés : Apprentissage actif, Incertitudes, Labellisation, Fonctions de croyance

Résumé : Ce document expose les recherches effectuées dans le cadre d'une thèse sur l'apprentissage actif de données incertaines et imprécises, soutenue grâce au financement de la Région Bretagne et du département Côtes-d'Armor. Deux principaux axes de recherche ont été explorés : les fonctions de croyance pour modéliser l'incertitude dans les données, et l'apprentissage actif pour travailler avec un nombre limité d'observations labellisées. La thèse s'est penchée sur la qualité et la quantité des labels en apprentissage automatique, visant à améliorer la modélisation des labels (qualité) tout en réduisant les coûts de labellisation (quantité). Des jeux de données à labels riches ont été proposés et mis à la disposition de la communauté scientifique. De nouveaux modèles ont été développés, des arbres de décision et des forêts aléatoires crédibilistes, tous capables de produire des prédictions incertaines et imprécises. Deux méthodes d'échantillonnage, fondées sur l'incertitude crédibiliste, ont été proposées et ont montré une augmentation des performances en apprentissage actif sur des jeux de données classiques. Enfin, des perspectives de recherche future ont été envisagées, notamment l'amélioration des méthodes d'échantillonnage par incertitude crédibiliste. Les travaux en cours comprennent la comparaison de la méthode proposée avec d'autres modèles de décomposition d'incertitudes, en se basant sur des recherches récentes liées à la thèse.

Title: Active learning of uncertain and imprecise data

Keywords: Active Learning, Uncertainties, Labelling, Belief functions

Abstract: This document outlines the research conducted within the scope of a thesis on active learning of uncertain and imprecise data, supported by funding from the Brittany Region and the Côtes-d'Armor Department. Two main research areas were explored: belief functions for modeling uncertainty in data and active learning to work with a limited number of labeled observations. The thesis focused on the quality and quantity of labels in machine learning, aiming to enhance label modeling (quality) while reducing labeling costs (quantity). Datasets with rich labels were proposed and made available to the scientific community. Novel models were developed, including evidential decision trees and evidential random forests, all capable of producing uncertain and imprecise predictions. Two sampling methods, based on evidential uncertainty, were proposed and demonstrated improved performance in active learning on conventional datasets. Finally, future research perspectives were considered, particularly improving methods for evidential uncertainty-based sampling. Ongoing work involves comparing the proposed method with other uncertainty decomposition models, drawing from recent research related to the thesis.