UNIVERSITE D'ARTOIS
Laboratoire de Génie Informatique
et Automatique de l'Artois

UNIVERSITE TUNIS EL MANAR
Faculté des sciences de Tunis

RAPPORT DE THESE

Pour l'obtention du Titre

DOCTEUR EN INFORMATIQUE

Présentée par

Ahmed SAMET

Théorie des fonctions de croyance : application des outils de data mining pour le traitement de données imparfaites

Soutenue le 3 décembre 2014
Devant les membres de Jury

JURY

M. MARTIN Arnaud	Professeur à l'Université Rennes 1	Rapporteur
M. ELOUADI Zied	Professeur à l'Institut Supérieur de Gestion de Tunis	Rapporteur
M. DUBOIS Didier	Directeur de recherche CNRS, Université Paul Sabatier	Président de Jury
M. GAMMOUDI Mohamed Mohsen	Professeur à l'ISAM de la Manouba	Examinateur
M. LEFEVRE Eric	Professeur à l'Université d'Artois	Directeur de thèse
M. BEN YAHIA Sadok	Professeur à la Faculté des Sciences de Tunis	Directeur de thèse

Préparée sous convention de cotutelle UTM-FST (Tunisie) - UA-LGI2A (France)

Remerciements

Je présente toute ma reconnaissance à Monsieur Arnaud Martin, professeur à l'Université Rennes 1, et à Monsieur Zied Elouedi, Professeur à l'Institut Supérieur de Gestion de Tunis, pour avoir accepté d'être rapporteurs de ce travail.

Je tiens également à remercier Monsieur Didier Dubois, directeur de recherche CNRS à l'Université Paul Sabatier, d'avoir accepté d'examiner cette thèse.

Merci au Professeur Mohamed Mohsen Gammoudi, professeur à la Institut Supérieur des Arts Multimédia de la Manouba, de m'avoir fait l'honneur de participer à ce jury de thèse.

Il m'est impossible d'exprimer toute ma gratitude à Monsieur Eric Lefèvre, professeur à l'Université d'Artois, pour avoir dirigé mon travail, pour son assistance et sa disponibilité. L'intérêt qu'il a manifesté pour mon travail, ses suggestions et ses remarques ont été d'une importance capitale. Je le remercie pour sa disponibilité, pour ses conseils avisés et pour son investissement constant.

Je tiens à remercier sincèrement Monsieur Sadok Ben Yahia, professeur à la Faculté des Sciences de Tunis qui s'est toujours montré à l'écoute et très disponible tout au long de la réalisation de ce travail, ainsi pour l'inspiration, l'aide et le temps qu'il a bien voulu me consacrer et sans qui ce mémoire n'aurait jamais vu le jour.

Je remercie tous les membres du Laboratoire LGI2A qui m'ont toujours chaleureusement accueilli pendant ces années de thèse.

Je tiens à exprimer ma gratitude à tout le personnel de l'Université d'Artois et Université du littoral côte d'opale pour l'accueil qu'ils m'ont réservé durant mes années d'ATER en France.

Je tiens à exprimer ma reconnaissance à tout le personnel du département des sciences de l'informatique de la FST pour les trois années d'enseignement que j'ai passé chez eux.

Enfin, j'adresse mes plus sincères remerciements à tous mes proches et amis, qui m'ont toujours soutenu et encouragé au cours de la réalisation de cette thèse.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Dès l'antiquité, l'être humain a éprouvé le besoin de représenter des informations imparfaites (imprécises, incertaines,...). On trouve, par exemple, chez Aristote (autour des années -350 av. J.-C.) des assertions du type "si un événement est nécessaire, c'est que son contraire est impossible". Il a fallu attendre l'apparition des théories de l'incertain pour formaliser mathématiquement ce genre d'assertion. Historiquement, l'apparition des théories de l'incertain remonte au XVII$^{\text{ème}}$ siècle avec la théorie des probabilités [51] et ce n'est qu'à partir de la seconde moitié du XX$^{\text{ème}}$ siècle qu'apparaissent des théories qui ne sont plus directement reliées aux probabilités. Ainsi, en 1965, Zadeh [112] introduit la théorie des ensembles flous puis la théorie des possibilités en 1978 [113]. Dempster [25], en 1967, présente la théorie de l'évidence qui sera ensuite reprise et formalisée par Shafer en 1976 [93].
Selon les interprétations données, on peut également trouver le terme théorie des fonctions de croyance pour désigner la théorie de l'évidence. La théorie des fonctions de croyance offre plusieurs avantages. En plus, de constituer un cadre de travail riche et flexible pour la représentation et la manipulation des informations imprécises et incertaines, elle est souvent utilisée pour fusionner des informations issues de plusieurs sources. Ces caractéristiques lui ont valu d'être employée dans de nombreux domaines comme le traitement d'image [84], la décision multi-critères [37],....
Parallèlement à l'avancée des théories de l'incertain, stocker un gros volume d'informations dans des supports de données est devenu une nécessité. Nous assistons ces dernières années à un accroissement considérable de la quantité d'informations stockées dans des grandes bases de données scientifiques, économiques, financières, médicales, etc. Ainsi, l'élaboration d'une nouvelle méthode d'exploitation de ce type de base de données est devenue un réel défi pour la communauté scientifique. La fouille de données s'est illustrée par sa capacité à extraire des informations précieuses et cachées à partir d'un gros volume de données grâce à des outils simples. Ce domaine pluri-disciplinaire se situe au confluent de différents domaines, tels que les statistiques, les bases de données, l'algorithmique, les mathématiques, l'intelligence artificielle,... [83].
Même si des travaux [46, 7] ont déjà été initiés permettant d'établir un lien entre la fouille de données et la théorie des fonctions de croyance, ces travaux n'ont pas exploré l'ensemble des interactions possibles. L'objectif de cette thèse est donc de dégager de nouveaux liens entre ces deux domaines. Ainsi, nous avons dégagé deux contributions possibles démontrant les apports réciproques susceptibles d'exister ente ces thématiques :

- Apport de la fouille de données dans la théorie des fonctions de croyance : nous étudierons l'apport des outils de la fouille de données pour la gestion des contradictions entre fonctions de croyance.
- Apport de la théorie des fonctions de croyance dans la fouille de données : nous nous intéresserons, dans ce cadre, à l'extraction de connaissances dans

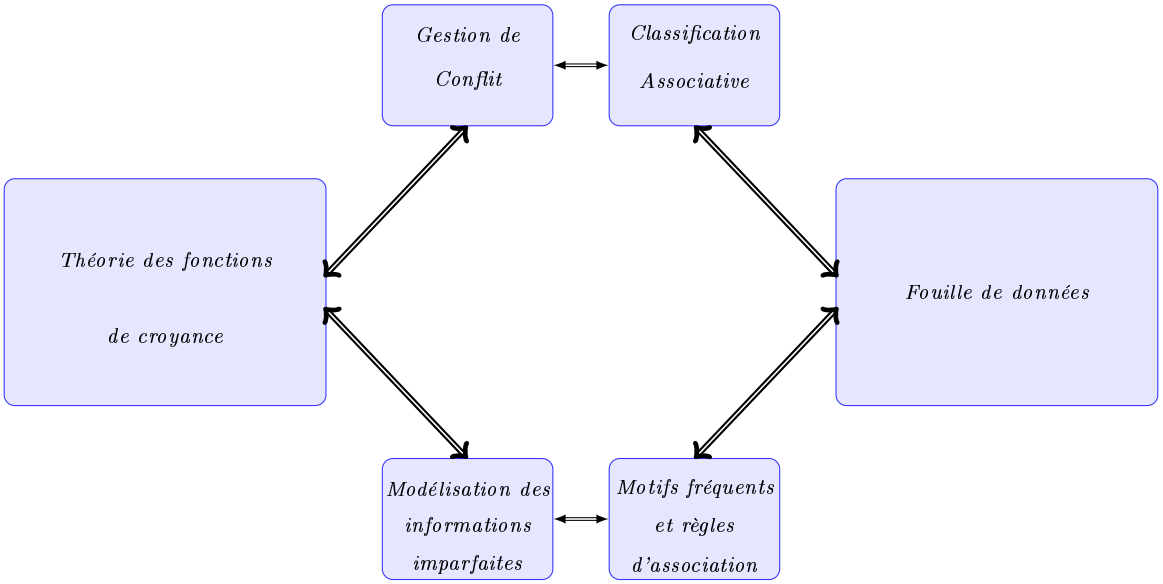

FIG. 1: Schéma des liens pouvant exister entre la fouille de données et la théorie des fonctions de croyance

les bases de données contenant des informations imparfaites.

Apport de la fouille de données pour la gestion du conflit

L'un des aspects de la théorie des fonctions de croyance est la fusion d'information. En effet, si plusieurs informations, issues de sources différentes, sont modélisées sous la forme de fonctions de croyance, il est alors possible de synthétiser ces connaissances en une fonction de croyance unique. L'objectif de cette étape de fusion est d'exploiter la complémentarité et la redondance des différentes sources afin d'obtenir une connaissance globale sous la forme d'une fonction de croyance plus précise et renforcée. Depuis son introduction en 1967 par Dempster [25] et son développement par Shafer [93], de multiples approches de combinaison de sources d'information ont vu le jour [92]. Parmi les approches de combinaison présentent dans la littérature, nous retrouvons la combinaison conjonctive. Elle a été introduite par Smets [105] dans son Modèle des Croyances Transférables (MCT) qui est, contrairement à celui de Shafer, une interprétation non probabiliste de la théorie.
L'une des particularités de cette combinaison est de représenter la contradiction entre les sources d'information combinées par une masse sur l'ensemble vide qu'on appelle communément *conflit*. Smets décrit, dans ses travaux, cette grandeur comme une alarme matérialisant les contradictions existantes entre les connaissances véhiculées par les sources d'information. Elle est aussi un indice de la qualité de la combinaison. Toutefois, lors de la combinaison de nombreuses fonctions de croyance, la masse accordée au conflit devient prépondérante du fait du pouvoir absorbant de

l'ensemble vide. Dans cette situation, la masse du vide conduit parfois à des prises de décisions incohérentes [114]. Ainsi, tout comme il est important d'obtenir des connaissances sur l'état des sources d'information, il est également essentiel de gérer au mieux cette masse avant la prise de décision. Généralement, la gestion du conflit se fait par une redistribution qui consiste à transférer les croyances vers certaines hypothèses solutions. Plusieurs travaux ont cherché à redistribuer cette masse pour une meilleure décision [109, 34]. Dès lors, une question se pose : comment pouvons nous orienter, de façon judicieuse, la redistribution du conflit vers les hypothèses les plus probables et ainsi corriger la fonction de masse résultante de la combinaison conjonctive ? Pour cela, nous supposons que des connaissances supplémentaires, susceptibles de nous aider pour orienter cette redistribution, sont disponibles dans une base de données. L'objectif est alors de déterminer les principes à mettre en œuvre afin d'extraire des informations pertinentes de ces bases.
La fouille de données s'impose comme l'un des meilleurs outils d'extraction de connaissances à partir d'un grand volume de données. Les connaissances extraites, souvent sous forme de motifs ou bien de règles d'association, sont associées à des indicateurs de pertinence qu'on appelle respectivement *support* et *confiance*. Généralement, c'est la règle d'association qui est intéressante quand il s'agit de prendre une décision. C'est une relation causale qui décrit la pertinence d'un choix dans le cas où une condition est vérifiée. Dès lors, une règle d'association peut s'avérer utile pour piloter la gestion du conflit.
Ainsi, dans nos travaux, nous considérons connues une fonction de masse conflictuelle issue de la combinaison conjonctive de plusieurs sources d'information et une base de données décrivant des informations supplémentaires et complémentaires à celles déjà fusionnées. La base de données est étudiée à l'aide des outils de fouille de données afin d'extraire les règles pertinentes. Notre contribution consiste à exploiter ces règles avec leur confiance pour définir la proportion de masse conflictuelle à transférer vers les hypothèses plus probables. Pour cela, nous utilisons les règles d'association ainsi qu'un cadre qui formalisme un ensemble de règles de combinaisons [63]. Au sein de ce cadre, le conflit est redistribué vers toutes les hypothèses possibles en fonction de poids. Dans notre contribution, nous proposons de retrouver ces poids à travers les règles d'association de classification et leur confiance.

L'approche proposée a été mise en œuvre dans le cadre d'une application réelle. Cette application concerne la classification de couronnes d'arbres dans des images hautes-résolutions. Les sources d'information relatives aux couronnes sont utilisées dans le processus de fusion et le conflit apparu est témoin de la contradiction qui subsiste. A ces sources s'ajoute une base de données relative à des informations différentes mais complémentaires de celles déjà fusionnées. A partir de cette base de données, nous extrayons la base générique qui est une base de règles d'association tout aussi informative mais beaucoup plus restreinte en nombre que les bases de règles classiques. Les règles retrouvées sont alors utilisées pour retrouver les facteurs de poids employés pour la redistribution du conflit.

Apport de la théorie des fonctions de croyance pour la modélisation des données

La fouille de données est un domaine en plein essor depuis les années 90. Grâce à sa formalisation et à ses outils simples, elle a su conquérir de nombreuses entreprises qui ont su profiter de ses avantages. Ainsi, les mesures de support et de confiance sont d'une aide précieuse quand il s'agit d'extraire des connaissances depuis un volume de données important. Aux origines de la fouille de données, ces outils ont été développés pour être employés dans le cadre d'une base de données binaires où l'existence d'un attribut est décrit par un "0" ou par un "1". La valeur "1" signifiant que l'attribut existe alors que la valeur "0" désigne le cas contraire.
Depuis le début des années 2000, la communauté de fouille de données a constaté des limites à ce type de représentation. En effet, dans un certain nombre de domaines, l'existence d'un attribut n'est pas toujours connue à l'avance et relève de l'incertain. Pour pallier ce problème, la fouille de données binaires s'est ouverte à d'autres méthodes de représentation de l'information et plus particulièrement à la représentation des imperfections liées à cette information. Ces approches sont par exemple la théorie des probabilités [113], la théorie des ensembles flous [112] ...
De son coté, la théorie des fonctions de croyance permet la représentation des données imparfaites. Elle permet dans un même formalisme d'encoder l'imprécision et l'incertitude contenues dans une information [33]. Ainsi, l'utilisation de la théorie des fonctions de croyance dans la fouille de données trouve tout son sens. Plusieurs travaux ont essayé de lier cette théorie aux bases de données [61] pour représenter des données imparfaites. La base de données générée a été alors appelée base de données évidentielles (au nom de la théorie de l'évidence). Mais ce n'est qu'à la fin des années 2000 que la communauté fouille de données s'est véritablement intéressée à ce type de base de données et dès lors le concept de fouille de données évidentielles a émergé.
Ce domaine est récent et peu de travaux ont été entrepris. Les premières recherches ont essayé d'adapter les outils et les algorithmes originaux à ce nouveau contexte d'extraction. De nouvelles mesures de support et de confiance ont alors été proposées pour ce type de base de données [46]. L'équivalent de l'algorithme d'extraction Apriori [3] a été introduit dans le cadre évidentiel [46]. Cela-dit, plusieurs limites peuvent être dégagées de ce qui a été proposé. Parmi celles-ci, deux limites fondamentales sont à noter :

- la mesure de support est imprécise et coûteuse en temps de calcul,
- il n'existe pas actuellement de classifieur associatif fondé sur une mesure de confiance évidentielle.

Dans un chapitre contribution, après avoir brossé un état de l'art des recherches sur la fouille de données évidentielles et les limites de ces travaux, nous proposons, dans un premier temps, de remédier au problème du temps de calcul exponentiel du support. Pour cela, nous proposons une nouvelle écriture équivalente à ce support mais plus simple. Dans la seconde partie, nous introduisons une nouvelle mesure de

support que nous avons appelée support *précis*. Nous mettons en évidence l'intérêt de cette nouvelle mesure et les avantages qu'elle offre par rapport à sa concurrente. Un comparatif détaillé est illustré sur des bases de données évidentielles générées.
Dans le second chapitre de contributions, nous proposons une nouvelle mesure de confiance qui repose sur la mesure du support déjà énoncée. Cette mesure est cohérente avec ses prédécesseurs issus de la fouille de données binaire, probabiliste,... Un nouveau classifieur, que nous appellerons *EDMA*, est proposé. Celui-ci attribue des classes à des instances évidentielles sur la base de règles évidentielles générées.

Structure du document

Les résultats de nos travaux de recherche sont synthétisés dans ce mémoire, qui est composé de 5 chapitres et d'une conclusion générale.
Dans le premier chapitre, nous présentons les concepts fondamentaux de la théorie des fonctions de croyance à partir de l'interprétation de Smets [105]. Nous présentons également les différentes approches de modélisation des fonctions de croyance.
Le chapitre 2 est dédié à l'état de l'art de la fouille de données. Nous commençons par étudier les concepts fondamentaux de la fouille de données binaires avec les connections de Galois. Par la suite, nous nous attardons sur les différentes variantes imparfaites de la fouille de données. Un comparatif sur ces différentes bases est fourni à la fin du chapitre.
Le chapitre 3 constitue notre premier chapitre contribution. Dans celui-ci, nous développons l'extraction d'informations, à l'aide des outils de la fouille de données, pour la gestion du conflit lors de la combinaison de fonctions de croyance.
Les chapitres 4 et 5 sont relatifs à la deuxième interaction entre les deux domaines. Dans le premier des deux, nous abordons les mesures de support et nous proposons des alternatives aux solutions déjà existantes. Dans le cinquième chapitre, nous introduisons notre classifieur associatif évidentiel qui repose sur la mesure de confiance que nous avons introduite.
Enfin, nous terminons le présent mémoire par une conclusion générale dans laquelle nous résumons l'ensemble de nos travaux et nous présentons quelques perspectives de recherche.

Chapitre 1

Théorie des fonctions de croyance

Sommaire

1.1 Introduction

Connue aussi dans la littérature comme la *théorie de Dempster-Shafer* et surtout *théorie de l'évidence*, la théorie des fonctions de croyance a pour origine les travaux de Dempster [25] sur la généralisation de l'inférence Bayésienne lorsqu'il n'y a pas d'apriori sur les paramètres. Cette généralisation conduit à l'utilisation de probabilités non additives pouvant être combinées par une règle. Néanmoins, l'élaboration du formalisme de la théorie est imputable à Shafer [93]. En 1976, Shafer a montré l'intérêt des fonctions de croyance pour la modélisation de connaissances incertaines. De plus, elle permet de représenter de façon plus naturelle les informations imparfaites. En effet, la représentation des informations imparfaites ou de l'absence d'information, n'est pas bien prise en compte par la théorie des probabilités. Dans ce chapitre, nous décrivons les notions fondamentales de la théorie des fonctions de croyance à travers les différentes fonctions développées. Enfin, nous présenterons les méthodes les plus connues pour l'estimation des fonctions de masse.

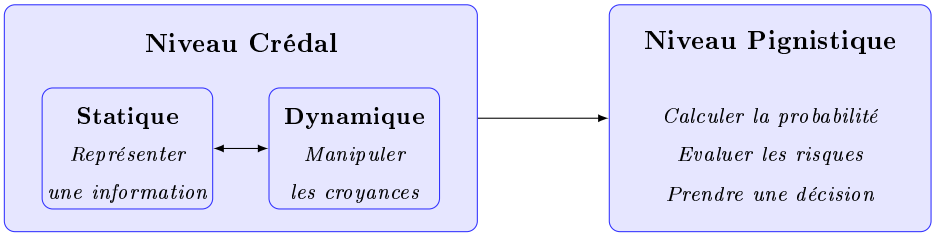

FIG. 1.1: Schéma des deux niveaux du Modèle des Croyances Transférables

1.2 Interprétation de la théorie des fonctions de croyance

Le livre, écrit par Shafer [93] et fondé sur l'article initial de Dempster [25], introduit la théorie. Après cet ouvrage, de nombreuses modifications et généralisations ont été apportées occasionnant ainsi une certaine confusion. Dans cette théorie, on peut identifier 3 points de vue distincts qui sont des extensions ou des variantes du modèle proposé par Shafer. Des détails sur ces différentes interprétations sont proposés par Smets [101].
Le premier modèle repose sur la théorie des probabilités imprécises appelée aussi théorie des probabilités inférieures [107, 116]. Cette théorie suppose l'existence d'une mesure de probabilité P précise, mais non parfaitement connue. On note alors $\mathcal{P}$ l'ensemble des probabilités compatibles avec l'information disponible. Le modèle peut alors être défini soit directement par $\mathcal{P}$, soit par ses enveloppes inférieure et supérieure. Sous certaines conditions, la probabilité inférieure peut alors être considérée comme une fonction de croyance.
Le second modèle proposé par Dempster [25, 26] est un cas particulier des probabilités imprécises. A l'aide de ce modèle, on peut là aussi définir des probabilités inférieure et supérieure. Ce modèle peut être rapproché des hints présentés par Kohlas et Monney [58]. L'interprétation logique de la théorie des fonctions de croyance proposée par Cholvy [18] peut être vue comme un cas particulier du modèle de Kohlas et Monney.
Enfin le dernier modèle, présenté par Smets [99, 97, 105], est une extension des fonctions de croyance, proposée au travers du Modèle des Croyances Transférables (MCT). L'approche proposée par Smets se distingue des approches précédentes de par son caractère fondamentalement non probabiliste. De plus, elle offre une justification axiomatique cohérente avec les principaux concepts de la théorie des fonctions de croyance et a permis de clarifier le lien existant entre la représentation des croyances [100] et la prise de décision. En effet, dans ce Modèle des Croyances Transférables, illustré par la figure 1.1, deux niveaux sont distingués :

- *le niveau crédal* : Ce niveau correspond à la représentation et à la manipulation des informations disponibles. Il peut être divisé en deux parties : statique et dynamique.
- *le niveau pignistique* : C'est à ce niveau qu'a lieu la prise de décision.

1.3 Niveau crédal statique

Dans ce qui suit, nous décrivons les différentes opérations de base réalisées au niveau crédal statique.

1.3.1 Cadre de discernement

La théorie des fonctions de croyance repose sur plusieurs notions parmi lesquelles on trouve le cadre de discernement ou cadre d'intérêt noté Θ. L'ensemble Θ décrit l'ensemble des réponses possibles à un problème tel que :

$$\Theta = \{H_1, H_2, \ldots, H_N\}.$$

Ce cadre peut vérifier la notion du *monde fermé* [96] (*closed-world*), c'est à dire que que l'ensemble est exhaustif et que les hypothèses sont exclusives. Il est aussi possible de s'affranchir de cette condition en considérant Θ comme cadre non exhaustif. Cette approche est alors appelée hypothèse du *monde ouvert* [105] (*open-world*). A partir du cadre de discernement Θ, on déduit l'ensemble noté 2^Θ, comprenant l'ensemble des 2^N sous-ensembles A de Θ :

$$2^\Theta = \{A, A \subseteq \Theta\} = \{\{H_1\}, \{H_2\}, ..., \{H_N\}, \{H_1 \cup H_2\}, ..., \Theta\}.$$

Cet ensemble comprend les hypothèses singletons de Θ mais aussi toutes les disjonctions possibles de ces hypothèses. Cet ensemble sert de référentiel de définition pour l'ensemble des grandeurs utilisées par la théorie des fonctions de croyance pour évaluer la véracité d'une proposition.

1.3.2 Fonction de masse

Une première grandeur appelée *fonction de masse* peut être construite. Une fonction de masse m (appelée aussi *jeu de masse* ou BBA[1]) associée à une source d'information (qui peut être issue d'un capteur, d'un agent, d'un classifieur,...) est définie par :

$$m : 2^\Theta \longrightarrow [0, 1] \tag{1.1}$$

telle que :

$$\begin{cases} \sum_{A \subseteq \Theta} m(A) = 1 \\ m(\emptyset) \geq 0. \end{cases} \tag{1.2}$$

Ce jeu de masse m traduit l'opinion d'un système sur la pertinence des différentes propositions. En effet, la quantité $m(A)$ est interprétée comme la part de croyance placée strictement sur A. Cette représentation est celle qui modélise le

[1] Acronyme du mot anglais Basic Belief Assignment

TAB. 1.1: Exemple de distribution de masse exprimée sur le cadre de discernement $\Theta = \{H_1, H_2, H_3\}$

	Catégorique	Dogmatique	Vide	Simple	Bayésienne	Consonante	Normale
$\emptyset$	0.00	0.02	0.00	0.00	0.00	0.00	0.01
H_1	0.00	0.05	0.00	0.00	0.20	0.00	0.10
H_2	1.00	0.10	0.00	0.00	0.30	0.00	0.09
H_3	0.00	0.13	0.00	0.00	0.50	0.20	0.05
$H_1 \cup H_2$	0.00	0.15	0.00	0.30	0.00	0.00	0.23
$H_1 \cup H_3$	0.00	0.17	0.00	0.00	0.00	0.00	0.12
$H_2 \cup H_3$	0.00	0.38	0.00	0.00	0.00	0.35	0,13
Θ	0.00	0.00	1.00	0.70	0.00	0.45	0.27

mieux l'incertitude car elle attribue un degré de croyance non seulement aux hypothèses singletons (comme les probabilités) mais aussi aux hypothèses composites. La masse $m(\emptyset)$ est appelée *masse conflictuelle* ou *conflit*. Deux interprétations subsistent au sein de la théorie. Une interprétation non probabiliste avancée par les travaux de Smets [105] sur le modèle des croyances transférables qui s'affranchit de la contrainte de normalisation et dans ce cas $m(\emptyset) \geq 0$. D'un autre coté, Dempster défend un raisonnement probabiliste, s'inspirant des probabilités imprécises, contraignant ainsi la normalisation des fonctions de masse $m(\emptyset) = 0$. Tout sous-ensemble $A \subseteq \Theta$ ayant une croyance non nulle est appelé *élément focal*. Nous noterons $f(m)$ l'ensemble des éléments focaux de m. L'union des éléments focaux est appelée *noyau*.

Au sein de la théorie des fonctions de croyance, plusieurs cas particuliers de fonctions de masse subsistent. En effet, soit une fonction de masse m issue d'une source S. Alors m est dite :

- Catégorique : si toute la croyance est attribuée à un élément de Θ c'est-à-dire $m(A) = 1$ et $A \subseteq \Theta$.
- Dogmatique : si Θ n'est pas un élément focal.
- Vide : si Θ est le seul élément focal.
- Simple : si elle a au plus deux éléments focaux, et s'il en existe deux alors Θ est l'un d'entre eux.
- Bayésienne : les éléments focaux sont des singletons.
- Consonante : les éléments focaux sont emboités.

Un exemple de chaque type de fonctions de masse est fourni dans le tableau 1.1.

1.3.3 Autres représentations de fonctions de masse

A partir d'une fonction de masse plusieurs autres fonctions peuvent être déduite.
La fonction de crédibilité :
La crédibilité de A (notée $Bel(A)$) représente la croyance minimale soutenant une proposition A et s'écrit de la manière suivante :

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B). \tag{1.3}$$

Dans le cadre du monde fermé, la fonction $Bel(.)$ vérifie un certain nombre de propriétés telles que :

1. $Bel(\emptyset) = 0$
2. $Bel(\Theta) = 1$
3. $Bel(\bigcup_{i=1,\cdots,n} A_i) \geq \sum_{I \subseteq \{1,\cdots,n\}} (-1)^{|I|+1} Bel(\bigcap_{i \in I} A_i)$

La fonction de plausibilité :
La plausibilité de A (notée $Pl(A)$) est quant à elle calculée comme la somme des masses des éléments ne contredisant pas A et elle est définie comme suit :

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{1.4}$$

Elle s'interprète comme la part de croyance qui pourrait potentiellement être allouée à A, compte tenu des éléments qui ne discréditent pas cette hypothèse. La plausibilité peut être retrouvée directement à partir de la crédibilité :

$$Pl(A) = Bel(\Theta) - Bel(\bar{A}). \tag{1.5}$$

La fonction de plausibilité est aussi la duale de la fonction de crédibilité. Ainsi dans le cadre du monde fermé, quelques propriétés intéressantes peuvent être exprimées :

1. $Pl(\emptyset) = 0$
2. $Pl(\Theta) = 1$
3. $Pl(\bigcap_{i=1,\cdots,n} A_i) \geq \sum_{I \subseteq \{1,\cdots,n\}} (-1)^{|I|+1} Pl(\bigcup_{i \in I} A_i)$

1.4 Niveau crédal dynamique

Dans cette partie, nous présentons un certain nombre de principes de la théorie des fonctions de croyance permettant le transfert de croyance entre les sous-ensembles de Θ.

1.4.1 Affaiblissement

L'affaiblissement est une étape nécessaire lorsqu'on possède une information sur la fiabilité d'une ou des sources d'information existantes. Elle consiste à modifier une fonction de masse, par rapport à un coefficient de fiabilité donné, par transfert de croyance. Les premiers travaux sur l'affaiblissement, dans le cadre des fonctions de croyance, ont été développés par Shafer [93], axiomatisés par Smets [99] et généralisés par Mercier et al. [75, 76]. L'affaiblissement de Shafer consiste à pondérer chaque source par un coefficient α comme suit :

$$\begin{cases} m^{\alpha}(B) = (1-\alpha) \cdot m(B) & \forall B \subseteq \Theta \\ m^{\alpha}(\Theta) = (1-\alpha) \cdot m(\Theta) + \alpha. \end{cases} \tag{1.6}$$

Où $\alpha \in [0,1]$ est appelé taux d'affaiblissement alors que $(1-\alpha)$ est appelé fiabilité. L'affaiblissement de Shafer a pour but le déplacement de la croyance vers l'ignorance Θ. Ce transfert de croyance est dicté par le scalaire α. Deux cas extrêmes peuvent être distingués. Pour $\alpha = 0$, on parle de source fiable et aucune modification sur la fonction de masse n'est apportée. En revanche, pour $\alpha = 1$, la source d'information est dite non fiable et coïncide avec une fonction de masse vide.

Une autre approche d'affaiblissement est utilisée pour corriger l'information fournie par une source à partir d'une méta-connaissance, elle est appelée *affaiblissement contextuel* [75]. L'idée principale de ce dernier consiste à varier la fiabilité d'une source selon le contexte envisagé (objet à reconnaître), c'est-à-dire exploiter la fiabilité de la source pour chaque contexte A_i, $i \in \{1, \cdots, 2^N\}$ sachant que la réalité est A_i. Soit une fonction de masse m fournie par une source S et β_A le degré de fiabilité sachant que la réalité se trouve dans A. L'affaiblissement contextuel de m est donné par :

$$m^{\alpha} = \sum_{B \subseteq A} G(A,B) \cdot m(B), \quad \forall A \subseteq \Theta \tag{1.7}$$

avec

$$G(A,B) = \begin{cases} \prod\limits_{\theta_k \in A \backslash B} \alpha_k \prod\limits_{\theta_l \in \bar{A}} \beta_l & si\ B \subseteq A, \\ 0 \quad sinon \end{cases} \tag{1.8}$$

où $G(A,B)$ représente la fraction de la masse $m(B)$ transférée à A. Cette fraction augmente avec :

- La plausibilité α_k que la source soit non fiable, sachant que la vérité est θ_k, pour tout $\theta_k \in A \backslash B$.

- Le degré de croyance β_l que la source soit fiable, sachant que la vérité est θ_l, pour tout $\theta_l \notin A$.

1.4.2 Conditionnement

Le conditionnement permet de tenir compte, dans une prévision, d'une information complémentaire. C'est l'un des concepts fondamentaux dans la théorie des fonctions de croyance mais également dans les autres théories de l'incertain telles que la théorie des possibilités [113] et la théorie des probabilités [20]. Il a été discuté et axiomatisé au sein de la théorie des probabilités [100]. Il a, ensuite, été étendu dans la théorie des fonctions de croyance avec les travaux de Smets [97].

1.4.2.1 Conditionnement dans la théorie des probabilités

Le conditionnement dans la théorie des probabilités a été initié par Cox [21] dans le courant subjectiviste qu'il représente. Il a proposé cinq postulats intuitifs qui ont donné par la suite les règles de probabilité. Les postulats sont :

- **P1.** cohérence ou non-contradiction,
- **P2.** continuité de la méthode,
- **P3.** universalité ou complétude,
- **P4.** énoncés sans équivoque,
- **P5.** pas de refus d'information et prise en compte de la dépendance du contexte.

Le postulat P5 conduit au conditionnement hypothétique : le degré de confiance dans une proposition A n'est connu que conditionnellement à un état de connaissance e. Un tel degré de confiance est noté $[A|e]$.

Il existe une relation fonctionnelle entre les degrés de confiance s'exprimant de la manière suivante :

$$P(AB|e) = P(A|Be) \cdot P(B|e). \tag{1.9}$$

$$P(A|e) + P(\bar{A}|e) = 1. \tag{1.10}$$

1.4.2.2 Conditionnement dans la théorie des fonctions de croyance

Dans le cadre de son Modèle des Croyances Transférables, Smets a levé le problème que pose le principe d'indifférence dans les probabilités. Ce problème n'est plus d'actualité dans le cadre crédibiliste. En effet, la fonction de masse définie par :

$$m(\Theta) = 1 \; et \; \forall A \subset \Theta, A \neq \Theta, m(A) = 0 \tag{1.11}$$

représente parfaitement l'indifférence (ou l'ignorance totale). La seconde idée de Smets est à l'origine du conditionnement. Le problème se pose de la manière suivante : étant donnée une information nouvelle permettant d'affirmer que la vérité se trouve dans un sous-ensemble B du cadre de discernement Θ, comment modifier un jeu de masses m pour prendre en compte cette nouvelle information ? La formulation

que propose Smets est la suivante :

$$m[B](A) = \begin{cases} \sum\limits_{X \subset \bar{B}} m(A \cup X) & \forall A \subset B \\ 0 \quad sinon \end{cases} \tag{1.12}$$

où $m[B]$ désigne le nouveau jeu de masse obtenu après conditionnement. Le conditionnement a également été défini comme la spécialisation d'une fonction de masse m [97]. La fonction de masse conditionnée doit être la moins informative telle que $pl[B](\bar{B}) = 0$. Ainsi, la matrice de spécialisation [55] C_B associée à un conditionnement sur B est définie par :

$$C_B(A, C) = \begin{cases} 1 & si\ A = B \cap C \\ 0 & sinon. \end{cases} \tag{1.13}$$

On vérifie que :

$$m[B] = C_B \cdot m. \tag{1.14}$$

1.4.2.3 Déconditionnement dans la théorie des fonctions de croyance

Le déconditionnement est l'opération duale du conditionnement. Étant donnée une fonction de masse conditionnelle $m[B]$, retrouver la fonction de masse d'origine m n'est pas toujours possible. Dans la pratique, en se basant sur le principe du minimum d'informations [105], c'est la fonction de masse la moins informative qui est retenue. Celle-ci est notée $m[B]^{\Uparrow\Theta}(A \cap \bar{B}) = m[B](A)$, $\forall A \subseteq B$. La matrice de généralisation [55] D_B associée à un déconditionnement par rapport à B est définie par :

$$D_B(A, C) = \begin{cases} 1 & A = \bar{B} \cup C \\ 0 & sinon. \end{cases} \tag{1.15}$$

1.4.3 Combinaison des croyances

La modélisation d'une information via une fonction de masse est une solution intéressante pour supporter l'imperfection qu'elle contient. La théorie des fonctions de croyance offre un autre atout très avantageux dans le cas d'une disposition de plusieurs informations. En effet, dans un cas multi-source (où chaque information issue d'une source est modélisée par une fonction de masse), il est possible de les combiner afin d'en extraire une information plus fiable.

1.4.3.1 Combinaison dans un cadre unique

La combinaison conjonctive

La combinaison conjonctive (communément appelée somme conjonctive) a été retenue par Smets dans le cadre de son Modèle des Croyances Transférables [98]. Soit deux sources S_1 et S_2, supposées fiables, dont les informations sont représentées

respectivement par les fonctions de masse cognitivement indépendantes m_1 et m_2. On note $m_{\textcircled{\cap}}$ le résultat de la combinaison conjonctive qui représente l'agrégation des informations issues des deux sources considérées :

$$m_{\textcircled{\cap}} = m_1 \textcircled{\cap} m_2.$$

Formellement, la combinaison conjonctive de deux fonctions de masse s'écrit de la manière suivante :

$$m_{\textcircled{\cap}}(A) = \sum_{B \cap C = A} m_1(B) \cdot m_2(C) \qquad A \subseteq \Theta. \tag{1.16}$$

Cette règle vérifie un certain nombre de propriétés intéressantes comme l'*associativité*, la *commutativité* et elle admet un *élément neutre* $m(\Theta) = 1$. Une des particularité de cette règle est qu'elle génère une fonction de masse non normalisée ($m(\emptyset) > 0$). Cette condition, n'étant pas possible dans un monde fermé, il est nécessaire de procéder à une étape de normalisation.

Combinaison orthogonale

La normalisation de la combinaison conjonctive est connue sous le nom *règle de combinaison orthogonale* ou sous la désignation de *règle de combinaison de Dempster*. C'est aussi la première règle introduite au sein de la théorie et elle s'écrit de la manière suivante :

$$\begin{cases} m_{\oplus}(A) = \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) \cdot m_2(C) = \frac{1}{1-K} m_{\textcircled{\cap}}(A) & \forall A \subseteq \Theta, A \neq \emptyset \\ m_{\oplus}(\emptyset) = 0 \end{cases} \tag{1.17}$$

où K est défini comme :

$$K = \sum_{B \cap C = \emptyset} m_1(B) \cdot m_2(C) = m_{\textcircled{\cap}}(\emptyset). \tag{1.18}$$

K représente la masse conflictuelle ($m(\emptyset)$) prenant ses valeurs entre $[0, 1]$. Généralement, cette masse traduit le degré de contradiction entre les sources fusionnées. Pour $K = 1$, S_1 et S_2 sont considérées comme complètement conflictuelles et les sources ne peuvent pas être fusionnées. Au contraire, si $K = 0$ les sources sont parfaitement en accord. La règle de combinaison de Dempster peut être généralisée pour J sources distinctes :

$$m_{\oplus}(A) = \oplus_{j=1}^{J} m_j(A) = \frac{1}{1-K} \sum_{A_1 \cap \ldots \cap A_j = A} \left(\prod_{j=1}^{J} m_j(A_j)\right) \qquad \forall A \subseteq \Theta, A \neq \emptyset \tag{1.19}$$

De même que pour la combinaison conjonctive, la règle orthogonale présente des propriétés intéressantes comme l'associativité, la commutativité et un élément neutre qui est la fonction de masse vide.

Autres combinaisons
L'utilisation de la combinaison conjonctive nécessite que toutes les sources fusionnées soient fiables. Cette contrainte est parfois difficilement vérifiable. Plusieurs approches ont été proposées afin de tenir compte de la nature des sources d'information avant, durant et après la fusion. Ces approches s'illustrent dans une nouvelle famille de combinaison autre que celles de l'opérateur conjonctif et orthogonal. C'est dans ce contexte que s'inscrit la combinaison disjonctive qui a été introduite par Dubois et Prade. Cette dernière suppose qu'au moins une des sources fusionnées soit fiable. La combinaison disjonctive s'écrit sous cette forme :

$$m_{\textcircled{\cup}}(A) = \sum_{B \cup C = A} m_1(B) \cdot m_2(C) \qquad A \subseteq \Theta. \tag{1.20}$$

1.4.3.2 Combinaison dans un cadre produit

La combinaison sous ses différentes déclinaisons, comme celles présentées dans les paragraphes précédents, nécessite que les fonctions de masse combinées partagent le même cadre de discernement. Cela-dit, il est également possible de fusionner des sources d'information qui ne partagent pas le même cadre en les translatant vers un cadre produit. Ce principe est connu sous le nom d'*extension* dans la théorie des fonctions de croyance. Soit une fonction de masse m définie sur un cadre de discernement Θ. Il est possible de définir m d'un cadre de discernement produit $\Theta \times \Omega$ noté $m^{\Theta\uparrow\Theta\times\Omega}$ tel que[2] :

$$m^{\Theta\uparrow\Theta\times\Omega}(B) = \begin{cases} m^{\Theta}(A) & si\ B = A \times \Omega,\ A \subseteq \Theta \\ 0 & sinon. \end{cases} \tag{1.21}$$

Cette opération est appelée l'extension vide. Ainsi, la combinaison de m_1 et m_2 ayant respectivement le cadre de discernement Θ et Ω devient :

$$m_{1\times 2}^{\Theta\times\Omega} = m_1^{\Theta\uparrow\Theta\times\Omega} \textcircled{\cap} m_2^{\Omega\uparrow\Theta\times\Omega} \tag{1.22}$$

L'opération duale de l'extension est la *marginalisation*. Elle permet de passer d'un cadre produit $\Theta \times \Omega$ à Θ par un transfert de chaque masse $m^{\Theta\uparrow\Theta\times\Omega}(B)$, $B \subseteq \Theta \times \Omega$ vers sa projection sur Θ :

$$m^{\Theta\times\Omega}(A) = \sum_{\{B \subseteq \Theta\times\Omega, Proj(B\downarrow\Theta)=A\}} m^{\Theta\uparrow\Theta\times\Omega}(B), \quad \forall A \subseteq \Theta \tag{1.23}$$

où la projection $Proj(B \downarrow \Theta)$ est la projection de B sur Θ, définie par : $Proj(B \downarrow \Theta) = \{\omega_1 \in \Theta / \exists \omega_2 \in \Omega; (\omega_1, \omega_2) \in B\}$.

[2] Nous noterons le cadre de discernement en exposant que si nécessaire

1.4.3.3 Gestion du conflit

La fusion de données est une solution intéressante pour l'obtention d'informations plus pertinentes. Historiquement, l'opérateur de Dempster (somme orthogonale) a été le premier opérateur de combinaison défini dans le cadre de la théorie des fonctions de croyance. Cette règle vérifie certaines propriétés intéressantes et son utilisation a été justifiée de manière théorique par plusieurs auteurs [32, 54]. Toutefois, son application n'est pas justifiable à chaque fois et des cas existent où son utilisation risque de mener à des résultats non cohérents. En effet, nous pouvons distinguer les cas où les sources d'information retenues sont dépendantes ou souffrent de non fiabilités. C'est Zadeh [114, 115] par l'intermédiaire d'un exemple qui a montré l'inconsistance du résultat de la combinaison orthogonale dans le cas de sources non fiables. Dans ces cas, l'utilisation de la somme orthogonale est fortement prohibée. Pour la combinaison conjonctive, une masse du vide $m(\emptyset)$ plus ou moins importante peut apparaître selon la contradiction entre les sources d'information fusionnées.

Plusieurs règles de combinaisons ont été proposées pour résoudre le conflit. Deux familles d'approches peuvent être distinguées. La première consiste à agir sur les sources avant de les fusionner en détectant les sources non fiables. La source non fiable est alors affaiblie (voir section 1.4.1). Plusieurs travaux se sont intéressés sur l'estimation des fiabilités des sources d'information moyennant des mesures de distance. Dans ce cadre, nous pouvons citer les travaux [85, 87, 70, 53, 27]. La seconde solution, qui a pour objet de répartir plus finement le conflit, consiste à gérer celui-ci au niveau de la combinaison. La liste des opérateurs de cette famille ne se résume pas juste à la combinaison orthogonale mais aussi à d'autres approches [104].

1.5 Niveau pignistique : Prise de décision

La manipulation de l'information est propre à la partie crédale où un résumé exhaustif est produit sous forme de fonction de croyance m. En ce qui concerne la partie décisionnelle, elle correspond au niveau pignistique. La phase de décision s'appuie sur la distribution pignistique [103] notée $BetP$ obtenue à partir de la fonction de masse m. Elle est aussi appelée probabilité pignistique pour le jeu de probabilité sur les singletons qu'elle génère. La probabilité pignistque, notée $BetP(.)$, consiste à repartir de manière équiprobable la masse d'une proposition A sur les hypothèses contenues dans A. Formellement, la probabilité pignistique $BetP$ est définie par :

$$BetP(H_n) = \frac{1}{1 - m^{\Theta}(\emptyset)} \sum_{A \subseteq \Theta} \frac{|H_n \cap A|}{|A|} \times m(A) \qquad \forall H_n \in \Theta. \tag{1.24}$$

Suite à cette conversion, il est impossible de revenir à la fonction de masse m initiale. En effet, à une fonction de masse on n'associe qu'une seule probabilité pignistique. Par contre, une probabilité pignistique peut être obtenue à partir d'une infinité de

fonctions de masse. Généralement, la décision est prise en choisissant l'élément H_0 possédant la plus grande probabilité pignistique :

$$H_0 = argmax_{H_k \in \Theta} BetP(H_k) \tag{1.25}$$

Le choix d'une hypothèse H s'apparente généralement à la mise en œuvre d'une action $a \in \mathcal{A}$. Néanmoins, cette mise en œuvre présente un risque. Supposons qu'il existe une fonction de coût $\lambda : \mathcal{A} \times \Theta \rightarrow R$ où $\mathcal{A}$ désigne un ensemble fini d'actions telle que $\lambda(a, H_i)$ désigne le coût de choisir l'action quand la vérité est H_i. Pour chaque action $a \in \mathcal{A}$, on peut associer une mesure de risque $BetR(a)$ telle que :

$$BetR(a) = \sum_{H_i \in \Theta} BetP(H_i)\lambda(a, H_i). \tag{1.26}$$

Cette mesure de risque peut être néanmoins bornée par une mesure de risque inférieur [29] qui conduit, par la suite, à une prise de décision pessimiste :

$$InfR(a) = \sum_{A \subseteq \Theta} m(A) min_{H_i \in A} \lambda(a, H_i). \tag{1.27}$$

Son dual existe et est appelé risque supérieur et conduit à une prise de décision optimiste. La fonction risque supérieur $SupR$ s'écrit de la manière suivante :

$$SupR(a) = \sum_{A \subseteq \Theta} m(A) max_{H_i \in A} \lambda(a, H_i). \tag{1.28}$$

1.6 Modélisation des fonctions de croyance

La modélisation des fonctions de croyance est sans doute la partie primordiale pour mettre au point un système de décision performant. En effet, la qualité de la modélisation se reflète dans la partie décisionnelle du système car plus la modélisation est conforme aux informations disponibles plus le résultat de la décision est cohérent. Dans la théorie des fonctions de croyance, on peut identifier deux familles de modélisation de fonctions de masse :

- Modélisation basée sur la vraisemblance [93]
- Modélisation basée sur la distance [119]

Dans ce qui suit, nous détaillons ces deux familles de modélisation existantes dans la littérature en nous fondant sur un problème de classification. Soit un vecteur x que nous cherchons à étiqueter (attribuer une classe). L'ensemble exclusive et exhaustive des hypothèses à qui x peut appartenir constituent notre cadre de discernement $\Theta = \{H_1, \ldots, H_n\}$. Pour cela nous disposons d'une base d'apprentissage $\mathcal{L}$ contenant des vecteurs. Un vecteur d'apprentissage x^i, appartenant à une classe k, possède l'étiquette suivante : $u_k^i = 1$. Un vecteur peut contenir plusieurs attributs (caractéristiques) où x_j^i désigne la caractéristique numéro j du vecteur x^i. Les sources d'information considérées à partir des vecteurs x^i sont vérifiées indépendantes.

1.6.1 Modélisation basée sur la vraisemblance

Proposée sous deux variantes différentes (globale et séparable), la modélisation avec la vraisemblance repose sur la notion de probabilité conditionnelle au classe. En effet, pour les deux familles d'approche, la fonction de vraisemblance d'un objet observé x est une fonction de Θ dans $[0, +\infty[$ définie par $L(H_k|x) = f(x|H_k)$ pour tout $k \in \{1, \cdots, K\}$.

1.6.1.1 Approche globale

L'approche globale est ainsi appelée car elle construit une fonction de masse en se fondant sur toutes les hypothèses H_n du cadre du discernement. Introduite par Shafer [93], cette approche conduit à des fonctions de masse consonantes. Afin d'établir une relation entre la plausibilité et une distribution de probabilité, Shafer a introduit les conditions suivantes :

- La plausibilité doit être proportionnelle à la vraisemblance :

$$Pl(\{H_k\}) = cL(H_k|x) \quad \forall H_k \in \Theta. \tag{1.29}$$

- La fonction de plausibilité doit être consonante et doit vérifier la condition suivante :

$$Pl(A \cap B) = max[Pl(A), Pl(B)] \quad \forall A, B \subseteq \Theta. \tag{1.30}$$

La fonction de plausibilité d'un sous-ensemble A peut être alors estimée de la manière suivante :

$$Pl(A) = c \cdot max_{H_n \in A} L(H_n|x) \tag{1.31}$$

A partir de la fonction de plausibilité, une fonction de masse m peut être construite.

1.6.1.2 Approche séparable

En utilisant la fonction de vraisemblance construite à partir des données d'apprentissage contenues dans $\mathcal{L}$, il est possible de définir une fonction de masse m_k associée à la classe H_k compte-tenu de différents axiomes détaillés dans [4, 5]. Les éléments focaux de cette fonction sont le complémentaire du singleton $\{H_k\}$ et l'ensemble Θ lui-même. Cette fonction de masse est définie de la manière suivante :

$$\begin{cases} m_i(\{H_n\}) = 0 \\ m_i(\overline{H_n}) = \alpha_i \cdot \{1 - R \cdot L(H_n|x)\} \\ m_i(\Theta) = 1 - \alpha_i \cdot \{1 - R \cdot L(H_n|x)\}. \end{cases} \tag{1.32}$$

Un deuxième modèle est défini comme suit :

$$\begin{cases} m_i(\{H_n\}) = \frac{\alpha_i \cdot R \cdot L(H_n|x)}{1+R \cdot L(H_n|x)} \\ m_i(\overline{H_n}) = \frac{\alpha_i}{1+R \cdot L(H_n|x)} \\ m_i(\Theta) = 1 - \alpha_i \end{cases} \tag{1.33}$$

où R est un facteur de normalisation contraint par :

$$R \in [0, (\underset{n \in [1,N]}{max} L(H_n|x))^{-1}] \tag{1.34}$$

Une fonction de masse unique m est retrouvée :

$$m = \oplus_{i=1}^{I} m_i. \tag{1.35}$$

Dans la partie précédente, nous avons considéré l'espace des caractéristiques dans sa globalité. Une autre stratégie consiste à modéliser les informations selon chaque caractéristique x_j (avec $j \in \{1, \cdots, J\}$) du vecteur x à classer. L'équation 1.32 devient :

$$\begin{cases} m_{ij}(\{H_n\}) = 0 \\ m_{ij}(\overline{H_n}) = \alpha_{ij} \cdot \{1 - R_j \cdot L(H_n|x_j)\} \\ m_{ij}(\Theta) = 1 - \alpha_{ij} \cdot \{1 - R_j \cdot L(H_n|x_j)\}. \end{cases} \tag{1.36}$$

L'équation 1.33 devient :

$$\begin{cases} m_{ij}(\{H_n\}) = \frac{\alpha_{ij} \cdot R_j \cdot L(H_n|x_j)}{1+R_j \cdot L(H_n|x_j)} \\ m_{ij}(\overline{H_n}) = \frac{\alpha_{ij}}{1+R_j \cdot L(H_n|x_j)} \\ m_{ij}(\Theta) = 1 - \alpha_{ij} \end{cases} \tag{1.37}$$

Les fonctions de masse construites sont alors fusionnées avec la somme orthogonale comme suit :

$$m_i = \oplus_{j=1}^{J} m_{ij}. \tag{1.38}$$

Une fonction de masse unique m est retrouvée :

$$m = \oplus_{i=1}^{I} m_i. \tag{1.39}$$

1.6.2 Modélisation basée sur la distance

Une méthode de modélisation de fonction de croyance reposant sur les distances a été introduite par Denœux en 1995 [28]. Cette approche s'inspire du principe de la méthode des *k plus proches voisins* (KPPV) introduite par Fix et Hodges en 1951 [39]. La méthode des KPPV est très connue dans le domaine de la reconnaissance de formes. Selon cette règle, un vecteur de classe inconnue se verra appartenir à la classe soutenue par la majorité de ses voisins dans un ensemble d'apprentissage

défini au préalable. Deux types d'approches distance subsistent se distinguant par la granularité des objets modélisés et les niveaux de fusion.

1.6.2.1 Approche multidimensionnelle

Soit un vecteur x à classer possédant un ensemble de caractéristiques. Soit un vecteur x^i de l'ensemble d'apprentissage, ayant l'étiquette $u_n^i = 1$, séparé d'une distance d en terme de caractéristique à x. L'approche d'estimation distance consiste à affecter une part de croyance à l'hypothèse H_n du cadre de discernement selon la distance d retrouvée. Le reste est affecté à l'ignorance totale. Ainsi, chaque vecteur x^i proche de x constituera une source d'information indépendante et sera représenté par une fonction de masse construite de la manière suivante :

$$\begin{cases} m_i(\{H_n\}) = \alpha^i \phi^i(d^i) \\ m_i(\Omega) = 1 - \alpha^i \phi^i(d^i) \end{cases} \tag{1.40}$$

où $0 < \alpha^i < 1$ est constante. $\phi^i(.)$ est une fonction décroissante qui satisfait la condition suivante : $\phi^i(0) = 1$ et $lim_{d\to\infty}\phi^i(d) = 0$. d_i représente la distance Euclidienne entre les vecteurs x et x^i. La fonction ϕ^i peut être écrite sous une forme exponentielle comme suit :

$$\phi^i(d^i) = exp(-\gamma^i(d^i)^2) \tag{1.41}$$

où γ^i est paramètre associé au $i^{ème}$ prototype. Ce paramètre permet de spécifier la vitesse de décroissance de la masse avec la distance selon le prototype. Cette approche a été proposée dans le cadre d'un classifieur appelé *classifieur evidentiel fondé sur la distance*. Afin de retrouver le résultat de la classification, une étape de fusion est nécessaire. Une fonction de masse m unique est obtenue par la combinaison orthogonale des k fonctions de masse retrouvées, de la manière suivante :

$$m = \oplus_{i=1}^{k} m_i. \tag{1.42}$$

1.6.2.2 Approche monodimensionnelle

Une autre variante de l'approche proposée par Denœux consiste à affiner encore plus la granularité des sources d'information par rapport à l'approche multidimensionnelle. Contrairement à l'approche multidimensionnelle qui considère le vecteur x dans sa globalité, l'approche monodimensionnelle consiste à étudier chaque composante de x à part. Le procédé de modélisation des fonctions de croyance de la section 1.6.2.1 est appliqué à la $j^{ème}$ composante de x et la fonction de masse m_{ij} est obtenue de la manière suivante :

$$\begin{cases} m_{ij}(\{H_n\}) = \alpha_j^i \phi_{ij}(d_j^i) \\ m_{ij}(\Theta) = 1 - \alpha_j^i \phi_{ij}(d_j^i) \end{cases} \tag{1.43}$$

où d_j^i est la distance entre x^i et x pour la $j^{ème}$ composante. La fonction ϕ_j^i peut être exprimée de la manière suivante :

$$\phi_{ij}(d_j^i) = exp(-\gamma_j^i (d_j^i)^2). \tag{1.44}$$

Les fonctions de masse construites sont alors fusionnées avec la somme orthogonale comme suit :

$$m_i = \oplus_{j=1}^{J} m_{ij}. \tag{1.45}$$

En utilisant l'équation 1.42, nous obtenons une fonction de croyance m unique.

1.7 Conclusion

Dans ce chapitre, nous avons exploré un certain nombre de principes fondamentaux de la théorie des fonctions de croyance. Le problème de représentation des fonctions de croyance a donné naissance à plusieurs interprétations [25, 101]. Parmi ces modèles, nous y trouvons le Modèle de Croyances Transférables qui est une interprétation non probabiliste. C'est aussi une lecture multi-niveau de la théorie où deux niveaux peuvent être dissociés. Dans le premier niveau, les croyances sont manipulées et deux sous-niveaux peuvent être distingués. Dans le niveau crédale statique, l'information est modélisée et dans le niveau crédale dynamique les croyances sont manipulées. Finalement au sein du second niveau, le niveau pignistique, la décision est prise par le biais de la probabilité pignistique. La théorie des fonctions de croyance représente une large variété d'imperfection [33], lui permettant d'être au cœur de nombreux autres domaines. Elle est souvent utilisée pour représenter les données imparfaites. Parmi ces domaines, nous y trouvons les bases de données [61, 60]. Vers le début des années 90, un nouveau domaine est apparu visant à exploiter les bases de données afin d'extraire des connaissances. Ce domaine, appelé fouille de données, a connu un essor important. De nombreux travaux ont été élaboré afin de pousser plus loin l'exploitation des connaissances pour assimiler les données imparfaites. Plusieurs formalismes de modélisation d'information ont été utilisés : la théorie des fonctions de croyance en fait partie.

Chapitre 2

Fondement de la fouille de données et ses variantes

Sommaire

2.1 Introduction

Le *data mining*, que l'on peut traduire par *fouille de données* en français, est apparu au milieu des années 1990 aux États-Unis. La fouille de données désigne la phase d'extraction des connaissances dans les bases de données (KDD ou Knowledge Discovery in Databases). Ses premières applications furent menées sur l'analyse du panier de la ménagère (en anglais Market Basket Analysis). Au départ, la fouille de données s'est donc intéressée aux bases de données des supermarchés afin d'identifier les améliorations possibles des ventes d'articles grâce à des décisions stratégiques.

Son essor est sans doute dû à la simplicité des outils statistiques apportant efficacité et performance.
La naissance du data mining est essentiellement due à la conjonction des deux facteurs suivants :

- l'accroissement exponentiel, dans les entreprises, de données liées à leur activité (données sur la clientèle, les stocks, la fabrication, la comptabilité ...) qu'il serait dommage de ne pas exploiter car elles contiennent des informations-clé pour la prise de décisions stratégiques.
- les progrès très rapides des matériels et des logiciels.

L'objectif poursuivi par le data mining, comme illustré dans la figure 2.1, est donc celui de la valorisation des données contenues dans les importantes bases de données. En effet, pour exploiter un volume important de données brutes, une étape de traitement est nécessaire afin de les mettre sous un format adéquat. Elles sont ensuite étudiées pour retrouver des éléments fréquents dans la base de données ou des règles. Ces derniers constituent des connaissances de valeur pour une prise de décision par la suite.
La fouille de données s'apparente généralement à deux notions fondamentales qui sont : les *motifs fréquents* et les *règles d'association*. Ces deux notions sont fondamentales et font la réussite et l'extension de la fouille de données dans divers domaines. Ainsi, grâce à son analyse, la fouille de données est utilisée souvent à des fins de classification, de prédiction et d'apprentissage. Cette simplicité d'utilisation et ses performances de calcul séduisantes ont réussi à placer cette discipline au cœur de nombreux autres domaines tels que : le traitement d'images [77], le web [69], la sécurité des réseaux [62] ....
Dans ce chapitre, nous détaillerons les principes fondamentaux de la fouille de données. Les définitions seront introduites en s'inspirant des connections de Galois (issue de l'Analyse des Concepts Formels). Les différentes notions de bases, ainsi que les algorithmes de fouille de données seront détaillés. Un état de l'art sur les différentes variantes possibles du data mining sera présenté ainsi que les mesures d'exploration dédiées.

2.2 Analyse des Concepts Formels

Dans ce qui suit, nous utilisons le cadre théorique présenté dans [24].

2.2.1 Notion d'ordre partiel

Soit E un ensemble. *Un ordre partiel* sur l'ensemble E est une relation binaire $\leq$ sur les éléments de E, tel que pour $x, y, z \in E$, nous avons les propriétés suivantes [24] :

1. *Réflexivité* : $x \leq x$
2. *Anti-symétrie* : $x \leq y$ et $y \leq x \Rightarrow x = y$

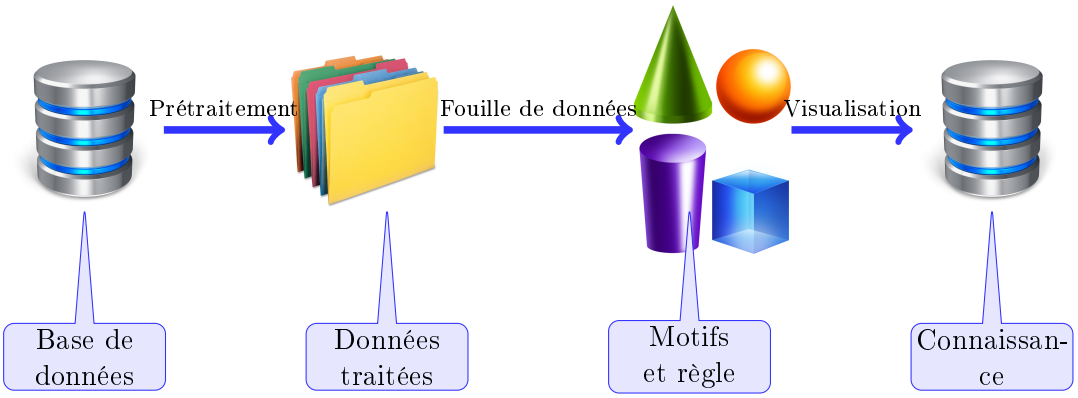

FIG. 2.1: Processus d'extraction de connaissances

3. *Transitivité* : $x \leq y$ et $y \leq z \Rightarrow x \leq z$

Un ensemble E doté d'une relation d'ordre $\leq$, noté $(E,\leq)$, est appelé *ensemble partiellement ordonné* [24].

2.2.1.1 Relation de couverture

Soit E un ensemble partiellement ordonné $(E,\leq)$ et x, y deux éléments de E. La *relation de couverture* entre les éléments de E, notée $\prec$, est la réduction transitive de la relation d'ordre et elle est définie comme suit :
$x \prec y$ si et seulement si $x \leq y$ et tel qu'il n'existe pas d'élément $z \in E$, tel que $x \leq z \leq y$, pour $z \neq x$ et $z \neq y$. Si $x \prec y$, alors y *couvre* x ou bien y est un *successeur* immédiat de x [24].

2.2.1.2 Les ensembles minorants et majorants

Soit un sous-ensemble $S \subseteq E$ de l'ensemble partiellement ordonné (E,$\leq$). Un élément $u \in E$ est un *majorant*, ou *borne-sup*, de S si pour tout élément $s \in S$, nous avons $s \leq u$. L'ensemble des majorants de S est noté *UB(S)*. D'une manière duale, un élément $v \in E$ est un *minorant*, ou *borne-inf*, de S si pour tout élément $s \in S$, nous avons $v \leq s$. L'ensemble des minorants de S est noté *LB(S)* [24] :

$$UB(S) = \{u \in E \mid \forall s \in S,\ s \leq u\}$$
$$LB(S) = \{v \in E \mid \forall s \in S,\ v \leq s\}.$$

Le *plus petit majorant* d'un ensemble S, s'il existe, est le plus petit élément de l'ensemble $UB(S)$ des majorants de S. Le *plus grand des minorants* d'un ensemble S, s'il existe, est le plus grand élément de l'ensemble $LB(S)$ des minorants de S [24].

2.2.1.3 Treillis d'inclusion

Un ensemble partiellement ordonné $(E,\leq)$ non vide est un *treillis* si pour tout couple d'éléments $(x, y) \in E$, l'ensemble $\{x, y\}$ possède un plus petit majorant, noté $x \vee y$, et un plus grand minorant, noté $x \wedge y$.

L'ensemble partiellement ordonné $(E,\leq)$ est un *treillis d'inclusion* si pour tout sous-ensemble $S \subseteq E$, les éléments $UB(S)$ et $LB(S)$ existent [24].

2.2.2 Connexion de Galois : Fermeture et générateur minimal

2.2.2.1 Opérateur de fermeture

Soit un ensemble partiellement ordonné $(E,\leq)$. Une application f de $(E,\leq)$ dans $(E,\leq)$ est un *opérateur de fermeture*, si et seulement si pour tout $S \subseteq E$ et $S' \subseteq E$, les trois conditions suivantes sont vérifiées [24] :

1. *Isotonie* : $S \leq S' \Rightarrow f(S) \leq f(S')$
2. *Extensivité* : $S \leq f(S)$
3. *Idempotence* : $f(f(S)) = f(S)$

2.2.2.2 Contexte formel

Un *contexte formel* ou *contexte d'extraction* est un triplet $\mathcal{K} = (\mathcal{O},\mathcal{I},\mathcal{R})$, où $\mathcal{O}$ et $\mathcal{I}$ sont deux ensembles finis et $\mathcal{R}$ une relation binaire entre $\mathcal{O}$ et $\mathcal{I}$, $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$. L'ensemble $\mathcal{O}$ est appelé ensemble de *transactions* (objets) et $\mathcal{I}$ est appelé ensemble d'*items* (ou attributs). L'ensemble d'item est appelé *itemset*. $\mathcal{R}$ une relation binaire définie sur $\mathcal{O} \times \mathcal{I}$. Chaque couple *(o, i)* $\in \mathcal{R}$ signifie que l'objet $o \in \mathcal{O}$ contient l'item $i \in \mathcal{I}$.

Exemple 1 *Dans le tableau 2.1 où* $\mathcal{I} = \{$*Eau, Fromage, Lait, Chocolat, Pain* $\}$ *est l'ensemble des items, alors que* $\{T1, T2, T3, T4, T5\}$ *est l'ensemble des transactions dans le contexte d'extraction D. Eau est un item alors que {Eau, Fromage} est un itemset.*

On définit le *support* d'un *itemset* X noté, *support(X)*, comme étant le nombre de lignes dans le contexte formel $\mathcal{K}$ contenant X[1]. Le support peut être exprimé formellement de la manière suivante :

$$Support(I) = \frac{|\{t | t \in \mathcal{O}, \forall i \in I, (t,i) \in \mathcal{R}\}|}{|\mathcal{O}|}. \tag{2.1}$$

Le support exprime la probabilité d'apparition d'un itemset sur l'ensemble des objets constituant le contexte formel. Le tableau 2.1 illustre un exemple d'un contexte formel. L'ensemble des itemsets fréquents forment l'ensemble $\mathcal{IF}$.

[1] Le support peut être défini aussi comme une fréquence d'apparition relative.

TAB. 2.1: Exemple de contexte d'extraction D

	Eau	Fromage	Lait	Chocolat	Pain
T1	X		X	X	
T2		X	X		X
T3	X	X	X		X
T4		X			X
T5	X	X	X		X

Exemple 2 *Dans le tableau 2.1, pour un support minimal minsup* $= \frac{4}{5}$*, les items Fromage, Lait et Pain sont considérés comme fréquents avec un support égal à* $\frac{4}{5}$.

Soit l'application ϕ de l'ensemble des parties de $\mathcal{O}$ (*i.e.* l'ensemble de tous les sous-ensembles de $\mathcal{O}$), noté par $\mathcal{P}(O)$, dans l'ensemble des parties de $\mathcal{I}$, noté par $\mathcal{P}(I)$. L'application ϕ associe à un ensemble d'objets $O \subseteq \mathcal{O}$, l'ensemble des items $i \in \mathcal{I}$ communs à tous les objets $o \in O$ [41], ainsi :

$$\phi : \mathcal{P}(\mathcal{O}) \rightarrow \mathcal{P}(\mathcal{I})$$
$$\phi(O) = \{i \in \mathcal{I} | \forall o \in O, (o, i) \in \mathcal{R}\}.$$

Dualement, considérons l'application ψ de l'ensemble des parties de $\mathcal{I}$ dans l'ensemble des parties de $\mathcal{O}$. Cette application associe à un ensemble d'items $I \subseteq \mathcal{I}$, l'ensemble d'objets $o \in \mathcal{O}$ communs à tous les items $i \in I$, telle que :

$$\psi : \mathcal{P}(\mathcal{I}) \rightarrow \mathcal{P}(\mathcal{O})$$
$$\psi(I) = \{o \in \mathcal{O} | \forall i \in I, (o, i) \in \mathcal{R}\}.$$

Le couple d'applications (ϕ, ψ) définit une connexion de Galois entre l'ensemble des parties de $\mathcal{O}$ et l'ensemble des parties de $\mathcal{I}$ tel que pour $O \subset \mathcal{O}$ et $I \subset \mathcal{I}$, les applications $\omega = \phi \circ \psi$ et $\gamma = \psi \circ \phi$ sont appelés *les opérateurs de fermeture de la connexion de Galois* [41].

Exemple 3 *Dans le Tableau 2.1, on retrouve :*

$$\phi(T1) = \{Eau, Lait, Chocolat\}$$

où $\phi(T1)$ *représente l'ensemble des produits présents dans la transaction* $T1$*. De manière équivalente :*

$$\psi(Eau) = \{T1, T3, T5\}$$

où $\psi(Eau)$ *regroupe l'ensemble des transactions contenant le produit Eau.*

2.2.2.3 Concept formel

Soient deux ensembles O et I tels que $O \subset \mathcal{O}$ et $I \subset \mathcal{I}$, un concept formel *(O, I)* est une paire, telle que si $O \times I \subset O' \times I' \subset \mathcal{R}$ alors $O = O'$ et $I = I'$. Les

ensembles O et I sont appelés, respectivement, l'*extension* (*domaine*) et l'*intension* (*co-domaine*) du concept *(O, I)*.
L'ensemble de tous les concepts formels extraits à partir d'un contexte formel ($\mathcal{O}, \mathcal{I}$, $\mathcal{R}$) est noté par $\mathcal{ECF}$. Une relation d'ordre, notée par $\leq$, est définie sur cet ensemble comme suit : $(O_1,\ I_1) \subseteq (O_2,\ I_2) \Leftrightarrow O_1 \subseteq O_2\ ,\ I_2 \subseteq I_1$.

2.2.2.4 Itemset fermé fréquent

Étant donné un opérateur de fermeture de la connexion de Galois ω sur un ensemble partiellement ordonné ($\mathcal{I}$,$\leq$), un élément $I \in \mathcal{I}$ est un *itemset fermé* si l'image de I par l'opérateur de fermeture est égale à lui-même [78] : $\omega(I) = I$. Un itemset fermé est donc un ensemble maximal d'items communs à un ensemble d'objets.
Étant donné un seuil minimal de support appelé *minsup*, un itemset fermé I est fréquent si et seulement si $support(I) \geq minsup$ [78].
Dans la suite, nous présentons trois propriétés relatives aux itemsets fermés [78]. L'ensemble des itemsets fréquents forment l'ensemble $\mathcal{IFF}$.

Propriété 1 *Tous les sous-ensembles d'un itemset fermé fréquent sont fréquents.*

Propriété 2 *Tous les super-ensembles d'un itemset fermé non fréquent sont non fréquents.*

Propriété 3 *Le support d'un itemset I est égal au support de sa fermeture $\omega(I)$, qui est le plus petit itemset fermé contenant I, i.e., $support(I) = support(\omega(I))$.*

Exemple 4 *Dans le Tableau 2.1 et $minsup = \frac{3}{5}$, l'itemset $\{Eau, Lait\}$ est considéré comme fermé fréquent car :*

$$\omega(\{Eau, Lait\}) = \phi \circ \psi(\{Eau, Lait\}) = \{Eau, Lait\}$$

Puisque le support de {Eau, Lait} est égal à $\frac{3}{5}$ alors {Eau, Lait} est un itemset fermé fréquent.

2.2.2.5 Générateur minimal

Un itemset $g \subseteq \mathcal{I}$ est un *générateur minimal* d'un itemset fermé I si et seulement si $\omega(g) = I$ et $\nexists\ g' \subset g$ tel que $\omega(g') = I$ [78].

Exemple 5 *L'item Eau est un générateur minimal de fermé $\{Eau, Lait\}$ car $\omega(\{Eau\}) = \{Eau, Lait\}$ et il n'existe pas d'autres items inclus dans Eau ayant la même fermeture.*

2.2.2.6 Classe d'équivalence

L'opérateur de fermeture ω induit une relation d'équivalence sur l'ensemble de parties de $\mathcal{I}$, *i.e.*, l'ensemble de parties est partitionné en des sous-ensembles dis-

joints, appelés aussi *classes d'équivalence*. Dans chaque classe, tous les éléments possèdent la même valeur de support et la même fermeture. Les générateurs minimaux d'une classe sont les plus petits éléments incomparables, tandis que l'itemset fermé constitue l'élément le plus large de cette classe [12].

2.2.3 Règle d'association

La formalisation du problème d'extraction des règles associatives a été introduite par Agrawal et al. en 1993 [2]. Une règle associative r est une relation causale entre itemsets de la forme $r : X \rightarrow (Y - X)$, dans laquelle X et Y sont des itemsets fréquents, tel que $X \subset Y$. Les itemsets X et $(Y - X)$ sont appelés, respectivement, *prémisse* et *conclusion* de la règle r. La génération des règles associatives est réalisée à partir d'un ensemble $\mathcal{F}$ d'itemsets fréquents dans un contexte d'extraction $\mathcal{K}$, pour un seuil minimal de support *minsup*. Les règles associatives valides (de confiance) sont celles dont la mesure de confiance, $Conf(r) = \frac{Supp(Y)}{Supp(X)}$, est supérieure ou égale à un seuil minimal de confiance, défini par l'utilisateur, qui sera noté dans la suite $minconf$. Si $Conf(r) = 1$ alors r est appelée *règle associative exacte*, sinon elle est appelée *règle associative approximative* [78].
Ainsi, chaque règle associative, $X \rightarrow (Y - X)$, est caractérisée par :

1. **La mesure de support** : elle correspond au nombre de fois où l'association est présente, rapporté au nombre de transactions comportant l'ensemble des items de Y. Le niveau de support permet de mesurer la fréquence de l'association [3].
2. **La mesure de confiance** : elle correspond au nombre de fois où l'association est présente, rapportée au nombre de présences de X. Le niveau de confiance permet de mesurer la force de l'association [3].

Exemple 6 *Dans le Tableau 2.1, la règle d'association* $Fromage \rightarrow Lait$ *possède un support égal à* 3 *et une confiance de* $\frac{3}{4}$.

Ainsi, étant donné un contexte d'extraction $\mathcal{K}$, le problème de l'extraction des règles associatives dans $\mathcal{K}$ consiste à déterminer l'ensemble des règles associatives dont le support et la confiance sont au moins égaux respectivement aux valeurs de *minsup* et $minconf$. Ce problème peut être décomposé en deux sous-problèmes comme suit :

1. Déterminer l'ensemble des itemsets fréquents dans $\mathcal{K}$, i.e., les itemsets dont le support est supérieur ou égal à *minsup*.
2. Pour chaque itemset fréquent I_1, générer toutes les règles associatives de la forme $r : I_2 \rightarrow I_1$ et dont la confiance est supérieure ou égale à $minconf$.

Cette décomposition du problème est à la base du premier algorithme de fouille de données à savoir *Apriori* [3]. Cet algorithme est détaillé dans la section suivante.

2.3 Algorithmes d'extraction

Dans cette partie, nous détaillons quelques algorithmes pionniers de la fouille de données. Un algorithme a été proposé par Agrawal et Srikant [3] et a servi pendant des décennies comme référentiel de comparaison. L'algorithme est appelé *Apriori*. Celui-ci permet l'extraction des motifs fréquents et des règles d'association.

2.3.1 Algorithme Apriori

Le premier algorithme pour l'extraction des motifs fréquents est Apriori [3]. Apriori est un algorithme de génération de fréquents par niveau. Son principe consiste à générer tous les itemsets niveau par niveau (débute à partir du premier niveau) et ne retient que ceux qui ont un support supérieur ou égal au *minsup* (i.e., fréquents). Les itemsets retenus sont sauvegardés et utilisés pour générer les candidats du niveau suivant. La même opération de calcul de support est effectuée pour l'élagage des non fréquents. Les détails d'Apriori sont illustrés dans l'algorithme 1. Ce principe est dicté par la propriété d'anti-monotonie du support :

Propriété 4 *Soit deux itemsets X et Y tel que $X \subset Y$, la mesure de support supp est dite anti-monotone si et seulement si $supp(X) \leq supp(Y)$.*

Algorithm 1 Algorithme Apriori

Require: $\mathcal{D}$

Ensure: $Frequent$

1: $k \leftarrow 1$
2: $F_k \leftarrow \{i | i \in I \text{ et } supp(i) \geq minsup\}$
3: **while** $C_k \neq \emptyset$ **do**
4: $\quad F_k \leftarrow$candidats de C_k dont le support $\geq minsup$
5: $\quad C_{k+1} \leftarrow$ candidats sont générés à partir de F_k
6: $\quad k \leftarrow k + 1$
7: **end while**
8: $Frequent \leftarrow \bigcup_k F_k$

2.3.2 Algorithme d'extraction des règles d'association

Agrawal *et al.* [2] ont proposé un premier algorithme pour la génération de règles associatives à partir des itemsets fréquents. Le principe de la génération de règles associatives valides est le suivant : pour chaque k-itemset fréquent I_k de taille $k \geq 2$, chaque sous-ensemble I_x de I_k est déterminé et la valeur du rapport $\frac{support(I_k)}{support(I_x)}$ est calculée. Si cette valeur est supérieure ou égale au seuil de confiance minimale *minconf*, alors la règle associative $I_x \Rightarrow (I_k\text{-}I_x)$ est générée.

Une extension de cet algorithme a été proposée dans [3]. Afin de réduire le nombre d'opérations réalisées pour la génération de règles associatives. Elle est fondée sur la propriété suivante :

Propriété 5 *Étant donné un itemset I_i, le support d'un sous-ensemble I_j de I_i est supérieur ou égal au support de I_i [3].*

2.3.3 Autres Algorithmes

Jusqu'à ce jour, plusieurs algorithmes d'extraction des motifs fréquents et de génération des règles d'association de confiance ont été proposés. Dans cette partie, nous détaillons brièvement quelques méthodes qui ont marquées la communauté de fouille de données. Cette sous section dresse un bref état de l'art, non exhaustif, des travaux de recherche dans le domaine de l'extraction des motifs fréquents. Trois familles d'approches peuvent subsister pour l'extraction des motifs fréquents et des règles d'association. La première famille d'approche repose sur les méthodes d'exploration de l'espace de recherche et sur la représentation des motifs sous forme de treillis. Dans cette famille d'approche, on retrouve Apriori déjà décrit dans la sous section 2.3.1 mais aussi d'autres algorithmes comme *Partition* [91], *Pascal* [9] et *FP-Growth* [45]. L'algorithme FP-Growth [45] utilise une structure de données appelée Fréquent Pattern Tree. Il permet de trouver les itemsets fréquents dans une base de transactions. Grâce à la structure FP-tree on conserve l'ensemble des éléments fréquents de la base des transactions dans une structure compacte. Ainsi, il n'est plus nécessaire de devoir parcourir la base de transactions. De plus, ces éléments se retrouvent triés ce qui accélère la recherche des règles d'association.
Une des limites des approches de la première famille est la complexité de recherche dans une base de données pouvant être énorme. D'autres alternatives qui constituent la deuxième famille d'approches, s'appuient sur l'extraction des itemsets fermés fréquents. Ces approches reposent sur la fermeture de la connexion de Galois [41] pour résoudre le problème d'extraction de règles d'association [80]. Elles sont fondées sur un élagage du treillis des itemsets fermés, en utilisant les opérateurs de fermeture de la connexion de Galois [41]. Plusieurs algorithmes ont été proposés dans la littérature tel que *CLOSE* [79], *CLOSET* [82] et *Charm* [118], dont le but est de découvrir les itemsets fermés fréquents. L'algorithme CLOSE proposé par Pasquier et al. [80] est le premier algorithme permettant l'extraction des itemsets fermés fréquents. Cet algorithme s'appuie sur les générateurs minimaux pour calculer les itemsets fermés. Ces générateurs minimaux sont utilisés pour construire un ensemble d'itemsets fermés candidats (itemsets fermés potentiellement fréquents), qui sont les fermetures des générateurs minimaux. Ainsi, étant donné un contexte d'extraction $\mathcal{K}$, CLOSE génère toutes les règles d'association en trois étapes successives [11] :

1. Découverte des itemsets fermés fréquents ;
2. Dérivation de tous les itemsets fréquents à partir des itemsets fermés fréquents obtenus durant la première étape ;

3. Pour chaque itemset fréquent, génération de toutes les règles ayant une confiance au moins égale à $minconf$.

Les algorithmes d'exploration comme Apriori ont été testés sur des bases plutôt éparses (les transactions de supermarchés) et donnent des résultats probants dans leur traitement. En revanche, les résultats d'Apriori deviennent moins attractifs dans le cas des bases denses où l'extraction des motifs fréquents longs devient une nécessité. La troisième famille se spécialise sur la recherche des motifs maximaux tels que :

$$\mathcal{FM} = \{I \in \mathcal{IF} | \forall I' \subseteq I, Support(I') \geq minsup\}. \tag{2.2}$$

Plusieurs algorithmes ont été proposées. Parmi eux nous pouvons citer MaxMiner [10], Pincer Search [67] et MaxEclat [117].

2.4 Fouille de données et ses variantes

2.4.1 Notion d'imperfection

La majorité des informations manipulées dans la réalité sont souvent imparfaites. Plusieurs auteurs ont cherché à les répertorier et à analyser ces imperfections comme [14, 102]. Parmi les types d'imperfections qui peuvent toucher une information, on peut distinguer entre-autres :

- Imprécision
- Incertitude
- Inconsistance

L'imprécision est relative à l'information en elle-même. Par exemple, nous savons que le client C1 a acheté soit du chocolat noir soit du chocolat au lait, ou bien il a acheté au moins 3 morceaux. La première est dite une information imprécise catégorique puisque elle prend ses valeurs dans un ensemble fini de catégories. En revanche, la seconde est dite information imprécise quantitative car elle gère une quantité prenant ses valeurs dans un intervalle. L'incertitude est quand à elle liée à la véracité de l'information en elle-même. Comme exemple : on peut demander à la caissière si le dernier client C1 a vraiment acheté des chocolats. Même si la réponse fournie est précise, elle peut être fausse. Enfin, l'inconsistance est relative à la redondance de données qui souffrent de conflit. Un exemple serait deux tickets de caisse d'un même client C1 pour une même date avec la présence de chocolat dans l'un et pas dans l'autre.

2.4.2 Fouille de données binaires

Dans les sections précédentes, nous avons défini la fouille de données à travers ses différentes notions. Ce type de fouille de données repose sur des bases de données binaires. L'appartenance d'un item à une transaction ne peut prendre que deux valeurs possibles : 0 ou 1. La présence de l'item dans la transaction est représentée

	Eau	Fromage	Lait
T1	5	0	2
T2	2	1	2
T3	0	2	1

TAB. 2.2: Exemple d'une base de données quantitative

	Eau	Fromage	Lait
T1	gazeuse	maroilles	poudre
T2	minérale	emmental	liquide
T3	minérale		liquide

TAB. 2.3: Exemple d'une base de données qualitative

par la valeur 1 et 0 indique son absence. Or, dans la réalité, le manque ou bien même l'imperfection de l'information font que de l'utilisation de ces bases de données binaires est difficile. C'est dans ce cadre que la communauté fouille de données s'est intéressée à d'autres types de bases de données tels que :

- Les bases de données quantitatives (tableau 2.2) : des bases de données intégrant des valeurs quantitatives.
- Les bases de données qualitatives (tableau 2.3) : des bases de données comportant des attributs qualités (ou catégories).
- Les bases de données temporelles (tableau 2.4) : les données sont associées à un attribut temporel.

D'autres types de bases de données subsistent dans la littérature. Elles dépendent du type de données en entrée. En effet, des travaux se sont intéressés à des bases de données mixtes qui regroupent des attributs de type différents [17, 108]. L'hétérogénéité et la complexité des attributs dans ces bases de données actuelles ont poussé à l'exploration de nouvelles méthodologies de modélisation. En effet, les formalismes de modélisation de l'incertain ont connu, une avancée remarquable ces dernières décennies et leur utilisation est devenue une nécessité. Plusieurs formalismes de modélisation comme les ensembles flous [112], la théorie des possibilités [113] et même la théorie des fonctions de croyance se sont illustrés comme solutions de représentation et de manipulation de ces types de données [72]. De l'application de ces formalismes, différents types de base de données sont apparus. Dans la section suivante, les différentes variantes de bases de données dans un contexte incertain sont présentées. Nous nous intéresserons aussi aux différentes mesures d'extraction de connaissances dans le cadre de ces bases.

	Date1	Date2	Date3
T1	lait chocolat	lait beurre	pain
T2		pain beurre	eau

TAB. 2.4: Exemple d'une base de données temporelle

	Lait	Chocolat	Pain
T1	0.3	0.8	1.0
T2	0	0.5	0.5
T3	0.7	1	0.2
T4	0.5	0.5	0

TAB. 2.5: exemple d'une base de données probabiliste

2.4.3 La fouille de données probabilistes

Les limites des bases binaires à représenter les informations imparfaites ont conduit à l'apparition de bases de données probabilistes [1]. Soit une base de données probabilistes $\mathcal{PDB}$ de n transactions et d'un ensemble $\mathcal{I} = \{i_1, \ldots, i_k\}$. L'appartenance d'un item à une transaction est représentée par une probabilité $pr_{Ti}(i)$ tel que $i \in \mathcal{I}$. Le tableau 2.5 illustre un exemple d'une base de données probabilistes. Le degré de présence d'un itemset X dans une transaction Ti a été défini de la manière suivante [19] :

$$pr_{Ti}(X) = \prod_{x \in X} pr_{Ti}(x). \tag{2.3}$$

Ainsi, le support de X sur $\mathcal{PDB}$ peut être déduit de la façon suivante :

$$Support(X) = \frac{\sum_{i=1}^{n} pr_{Ti}(X)}{|\mathcal{PDB}|}. \tag{2.4}$$

Exemple 7 *Dans la base de données du tableau 2.5, le support de l'itemset {Lait, Chocolat} est retrouvé de la manière suivante :*

$$Support(\{Lait, Chocolat\}) = \frac{(0.3 \times 0.8) + (0 \times 0.5) + (0.7 \times 1) + (0.5 \times 0.5)}{4} = 0.29 \tag{2.5}$$

Une version Apriori a été introduite par Chui et al. [19] pour la génération des itemsets fréquents. Cette variante probabiliste d'Apriori a été désignée par *U-*

Apriori. Tout comme dans la version originelle, U-Apriori repose aussi sur l'anti-monotonie de son support pour l'élagage des candidats. FP-Growth aussi a été traduit dans une version probabiliste avec le support décrit dans l'équation 2.4. Introduit par Leung et al., UF-growth [65] permet aussi la représentation des motifs fréquents sous forme d'arbre. Cette approche offre une certaine rapidité par rapport à Apriori en s'affranchissant de l'étape de la génération des candidats.

2.4.4 La fouille de données floues et possibilistes

Dans la littérature, plusieurs types de motifs ont été définis selon le type de corrélation à extraire et selon la nature des données à partir desquelles l'extraction est faite. En effet, dans les bases de données quantitatives, les motifs ainsi que la technique d'extraction diffèrent de ceux employés dans les bases de données binaires. De même, les motifs décrivant une corrélation entre les attributs diffèrent de ceux décrivant une corrélation entre les variations des attributs et de ceux qui intègrent des contraintes temporelles (i.e., les motifs séquentiels) [71]. Cela va de même pour les bases de données avec des quantificateurs linguistiques. La théorie des ensembles flous s'apparente aux traitements de ce type de données. D'ailleurs, la mesure du support a été redéfinie dans le cadre flou et dans ce qui suit, nous présentons ses fondements mathématiques[2].
Soit un contexte d'extraction flou $\mathcal{FDB} = (\mathcal{O}, \mathcal{I}, \tilde{R})$ tel que $\mathcal{O}$ décrit un ensemble fini d'objets (transaction), $\mathcal{I}$ un ensemble fini d'attribut (item), $\tilde{\mathcal{R}}$ une relation floue (i.e., $\tilde{\mathcal{R}} \subseteq \mathcal{O} \times \mathcal{I}$). Le couple (o, i) appartient à $\mathcal{R}$, signifie que l'attribut (item) i appartenant à $\mathcal{I}$ est vérifié par l'objet (transaction) o appartenant à $\mathcal{O}$ avec un degré α (i.e., $\mu_o(i) = \alpha$).

La mesure de support flou repose aussi sur la notion de dénombrement. En effet, comme dans le cadre binaire, le support est calculé par comptage des éléments présents dans les transactions. En fouille de données floues, le support repose sur une variable de comptage *count* qui est calculée de la manière suivante :

$$count(i) = \sum_{o=1}^{|\mathcal{O}|} \mu_o(i). \tag{2.6}$$

Le degré de présence (support) de l'item i dans la base de données est retrouvé de la manière suivante :

$$Support(i) = \frac{count(i)}{|\mathcal{O}|}. \tag{2.7}$$

Pour un itemset X de taille q tel que $x_i \in X$ et $i \in [1, q]$, le comptage devient alors :

$$support(X) = \frac{\sum_{o=1}^{|\mathcal{O}|} min\{\mu_o(x_i), i = 1..q\}}{|\mathcal{O}|}. \tag{2.8}$$

[2]Les concepts fondamentaux de la théorie des ensembles flous sont décrits dans l'annexe A.

	Lait	Chocolat	Pain
T1	5(0.2)+6(0.8)+7(0.5)	2	2
T2	2	2(0.3)+3(1.0)+4(0.4)	1
T3	4	0	1

TAB. 2.6: Exemple d'une base de données possibilistes

A partir de ces mesures d'estimation du support, une version d'Apriori a été proposée par Hong et al. [48] reposant sur la génération des motifs fréquents niveau par niveau. Dans une seconde étape, l'algorithme *fuzzy AprioriTid* génère les règles d'association de confiance.
L'apport de la théorie des possibilités a été étudié dans le cadre de la fouille de données. Elle a été souvent associée à la fouille de données floues et surtout employée dans le cadre des données qualitatives et quantitatives imprécises. A notre connaissance, les premiers travaux de l'intégration des possibilités dans la fouille de données sont attribués à Djouadi et al. [31]. Généralement, les valeurs quantitatives précises sont regroupées sous forme de classe afin de travailler sur un espace réduit d'ensemble $E = \{E1, \ldots, En\}$. En revanche, parfois les valeurs quantitatives sont imprécises et la seule information que l'on possède est le degré de possibilité d'appartenance de l'objet aux différents éléments de Ei comme le démontre le tableau 2.6. Cette notion a été étendue dans les travaux de Weng et Chen [108, 17] où les informations imprécises peuvent être assimilées à des valeurs quantitatives précises, des intervalles ou bien même à des catégories.

La mesure de support pour ces derniers est retrouvée de la manière suivante :

$$sup_{DB}(B) = \sum_{Ti \subseteq DB} \frac{sup(A_{Ti}, B)}{|DB|} \tag{2.9}$$

où A_{Ti} est un d-itemset [17] (i.e., $A_{Ti} = \{ai = (it_i, v_i)\}$ où (it_i, v_i) est un d-item[3] de la transaction Ti de DB et B est un r-itemset [17] (i.e., $B = \{ai = (ic_i, f_i)\}$ où (ic_i, f_i) est un r-item[4]). Le $sup(A_{Ti}, B)$ est calculé de la façon suivante :

$$sup(A, B) = Min_{j=1}^{n} sup(ai_j, b_j) \tag{2.10}$$

avec ai_j et b_j sont les $j^{ème}$ composantes des itemsets A et B et $sup(A, B)$ dénote le degré d'appartenance de B à A. Le $sup(ai_j, b_j)$ peut être écrit d'une façon généralisée de la manière suivante :

$$sup(a_i, b_j) = p_{ir} \times sim_r(ai, b_j) \tag{2.11}$$

[3] d-item est un item de la forme $(it_i, \frac{p_{i1}}{x_{i1}} + \frac{p_{i2}}{x_{i2}} + ...)$ ou it_i est un item et x_{ij} les possiblités de valeurs prises par it_i chacune avec un degré de possibilité p_{ij}.
[4] r-item est un item linguistique ou catégorique.

où sim_r est une mesure de similarité.

2.4.5 La fouille de données évidentielles

L'apport de la théorie des fonctions de croyance sur la représentation des données imparfaites dans les bases de données a été étudié [61, 46, 7]. Il est possible de représenter les informations sous forme de fonctions de masse. Ainsi, elle peut être assimilée à une base de connaissances d'expert (les transactions) où chacun d'entre eux exprime une opinion par rapport à une question (item). Le premier à s'être intéressé à ce type de représentation est Lee [61] avec le concept de base de données évidentielles[5]. Dans ce qui suit, nous présenterons la définition et le concept de base évidentielle. L'appartenance d'un item à une transaction est exprimée grâce à une fonction de masse. Chaque cellule de la ligne j et de colonne i de la base de données exprime l'opinion de l'expert i par rapport à la question j et elle peut s'écrire de la manière suivante :

$$m_{ij} : 2^{\Theta_i} \rightarrow [0, 1]$$

où

$$\begin{cases} m_{ij}(\emptyset) = 0 \\ \sum\limits_{A \subseteq \Theta_i} m_{ij}(A) = 1. \end{cases} \tag{2.12}$$

L'intérêt majeur d'une telle modélisation des connaissances est l'assimilation de plusieurs types d'imperfections. En effet, dans [33], Dubois et Prade ont mis en évidence la complémentarité de l'imprécision et de l'incertitude dans les informations imparfaites. Le tableau 2.7 schématise un exemple d'une base de données évidentielles.

Transaction	Attribute A	Attribute B
T1	$m_{11}(A_1) = 0.7$ $m_{11}(\Theta_A) = 0.3$	$m_{21}(B_1) = 0.4$ $m_{21}(B_2) = 0.2$ $m_{21}(\Theta_B) = 0.4$
T2	$m_{12}(A_2) = 0.3$ $m_{12}(\Theta_A) = 0.7$	$m_{22}(B_1) = 1$

TAB. 2.7: Exemple d'une base évidentielle

Une mesure de support crédibiliste a été proposée par [46]. Cette mesure sera détaillée dans le chapitre 4.

[5] La désignation est inspirée d'un des autres noms de la théorie : théorie de l'évidence.

2.4.6 Discussion

Dans cette partie, nous récapitulons les différentes variantes de fouille de données déjà détaillées dans les sections précédentes. Le choix de la théorie utilisée peut dépendre de la nature de l'imperfection gérée comme il est indiqué dans le tableau 2.8. Ce dernier détaille les cinq variantes de fouille de données introduites au préalable avec leurs caractéristiques. Mise à part la fouille de données binaires qui ne formalise pas l'imperfection avec sa représentation en 0 et 1, les autres variantes supportent l'imperfection en représentant les données catégoriques et quantitatives. L'extraction des motifs fréquents et des règles d'association ont été étudiés sous plusieurs types d'imperfection et avec des algorithmes différents. Il est important de noter que même dans leurs différences, la majorité des algorithmes de fouille de données imparfaites se rejoignent sur l'utilisation du principe d'extraction Apriori. En effet, la génération des fréquents et par la suite les règles d'association se fait par niveau (i.e., génération des candidats niveau par niveau et élagage des infréquents) en satisfaisant la propriété d'anti-monotonie.

Malgré la jeunesse de cet intérêt (début des années 2000 jusqu'à présent), chaque théorie apporte un plus par rapport à la représentation de connaissance. La fouille de données probabiliste, floue, possibiliste et même évidentielle ont toutes apporté un gain en terme de représentation des informations incertaines et imprécises. Toutefois, la théorie des fonctions de croyance permet de représenter de façon plus souple les imperfections de différentes natures. Par ailleurs, des fonctions de masses particulières correspondent à des mesures d'autres théories de l'incertain 1.3.2. Par conséquent, les bases évidentielles peuvent être considérées comme une généralisation des bases de données binaires, probabilistes et floues [89].

Par ailleurs, l'interaction entre la théorie des fonctions de croyance et la fouille de données peut ne pas se résumer uniquement à la modélisation de l'imperfection dans les bases de données. La figure 2.2, illustre les interactions possibles entre la fouille de données et la théorie des fonctions de croyance. En effet, il est possible d'envisager un apport de la fouille de données pour résoudre des problèmes propres à la théorie des fonctions de croyance. Les motifs fréquents ainsi que les règles d'association sont des informations précieuses qui peuvent être d'un grand secours dans le cas où la prise de décision n'est pas possible. C'est dans ce cadre que nous envisageons d'étudier les apports et les interactions possibles entre ces deux domaines différents pour la résolution de leurs problèmes respectifs.

2.5 Conclusion

Dans ce chapitre, nous avons présenté les principes fondamentaux de la fouille de données. La fouille de données se base sur deux notions qui sont la mesure du support des motifs et de la confiance de règles d'association. Dans un but de représentation plus condensé des motifs, nous avons détaillé le rapport existant entre les connections de Galois du domaine de l'Analyse des Concepts Formels (ACF) et la fouille de données. Parmi, les algorithmes cités, nous trouvons l'algorithme

Type de base	fonction de représentation	Type de données représentées	type d'imperfections supportées	Algorithmes
Binaire	0 ou 1	binaire	aucun	Apriori[3], CLOSE[80]
Probabiliste	fonction de probabilité $p_{Ti}(x)$	catégorique quantitative	incertaine	U-Apriori[19] UF-Growth[65]
Floue	fonction d'appartenance $\mu_A(x)$	catégorique quantitative	imprécise	fuzzy AprioriTid[48]
Floue/Possibiliste	fonction d'appartenance $\mu_A(x)$+ possibilité $\pi(x)$	catégorique quantitative(valeur,interval)	imprécise incertaine	UDM[108]
Evidentielle	fonction de masse mij	quantitative catégorique	imprécise incertaine	BIT[46] FIMED[7]

TAB. 2.8: Récapitulatif des propriétés des différentes variantes de fouille de données

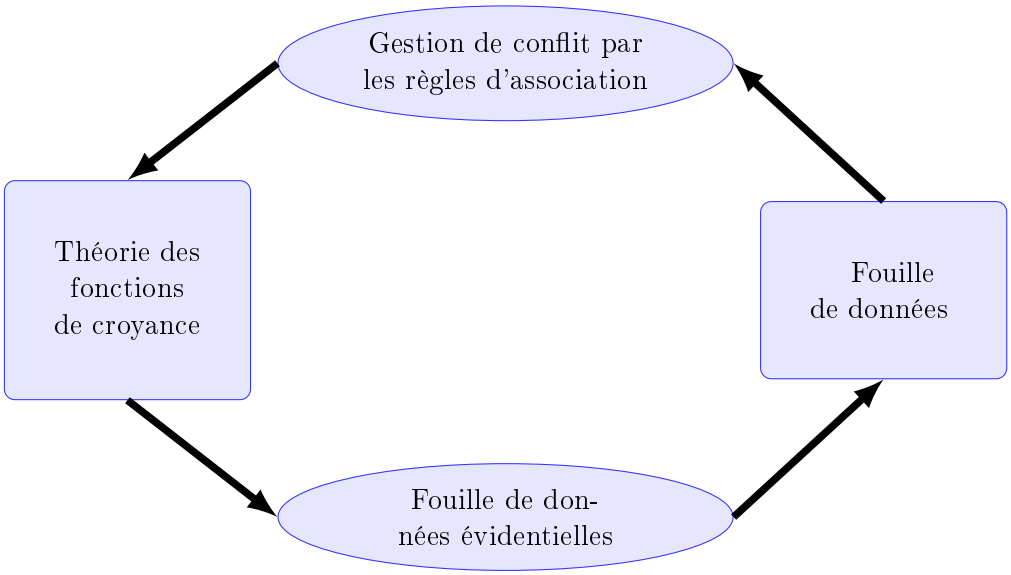

FIG. 2.2: Les relations entre la fouille de données et la théorie des fonctions de croyance

CLOSE [79] qui repose sur ces notions. Depuis son introduction vers le début des années 90, la fouille de données binaires a apporté un gain en extraction de connaissance avec des temps de calcul toujours aussi attractifs. Cela-dit, l'évolution des données et les imperfections qu'elles contiennent ont donné vie à de nouvelles disciplines de fouille de données. Ainsi, dans la deuxième partie de ce chapitre, nous avons exploré les variantes existantes dans la littérature de la fouille de données. La nature des données traités a poussé la communauté à explorer d'autres formes de représentation de données se fondant sur les théories de l'incertain. Des variantes de fouille de données sont apparues et dans ce chapitre, nous avons vu les plus importantes. Parmi ces types entrevus, nous nous sommes intéressés à la fouille de données évidentielles. Cette dernière matérialise l'apport que peut avoir la théorie des fonctions de croyance dans la représentation des données. D'un autre coté, la fouille de données peut contribuer pour résoudre des problèmes au sein de la théorie des fonctions de croyance. Un des aspects de contribution serait l'utilisation des connaissances extraites d'une base de données par une méthode de fouille de données. Les connaissances extraites peuvent être d'une aide appréciable dans le cas ou un système de fusion affiche des limites et aucune décision ne peut être prise. Ce type d'interaction est étudiée dans le chapitre suivant.

Chapitre 3

Règles d'association pour la gestion du conflit

Sommaire

3.1 Introduction

La théorie des fonctions des fonctions de croyance est souvent utilisée pour la classification multi-sources [106]. Cela-dit, elle présente une limite quand il s'agit de fusionner des sources non fiables car dans ce cas un conflit peut apparaître. La résolution du conflit peut dépendre de la gestion des connaissances [36, 64]. En effet, une information exogène, de nature différente (bases de données, information simple,...), peut être utilisée pour gérer le conflit. Tout comme la théorie des fonctions de croyance peut contribuer à résoudre des limites de la fouille de données

classique, et notamment en terme de modélisation de connaissances la fouille de données peut résoudre des problèmes propres à la théorie des fonctions de croyance. Nous nous intéressons plus particulièrement au problème de la gestion du conflit en présence d'une information supplémentaire représentée par une base de données. Nous étudions, l'impact des règles d'association et notamment les règles génériques afin de résoudre ce problème. Ces règles d'association de classification (d'où le nom classification associative) sont employées par la suite pour redistribuer le conflit entre les hypothèses les plus pertinentes.
Dans ce chapitre, nous dévoilons les fondements de la classification associative qui se base sur les règles issues de la fouille de données. Nous présentons les règles d'association issues des bases génériques qui offrent concisions et informativités. Cette base de règles est utilisée dans le cadre de la classification multi-sources d'une image haute-résolution forestière. En effet, nous mettons en évidence l'importance de ces règles d'association pour la gestion du conflit issu de la fusion de sources d'information.

3.2 Base générique IGB

Les algorithmes d'extraction de motifs fréquents et de règles d'association se butent sur un problème de taille à savoir le volume du résultat. En effet, le dénominateur commun entre ces approches réside dans le nombre important de motifs et de règles d'association générées. Cette contrainte amplifie la difficulté de la tâche de traitement lorsque plusieurs règles intéressantes subsistent sans que nous sachions laquelle utiliser. Ainsi, dans cette partie, nous nous intéressons à la réduction des règles d'association.

3.2.1 Base générique de règles exactes et base générique de règles approximatives

3.2.1.1 Base générique de règles exactes

Soit $\mathcal{RA}$ l'ensemble des règles d'association extraites à partir d'un contexte d'extraction $\mathcal{K}$. Une règle $R : R_a \Rightarrow R_c \in \mathcal{RA}$ est dite règle redondante (ou dérivable) par rapport à $R_1 : R_{a1} \Rightarrow R_{c1}$ si R satisfait les conditions suivantes :

1. $Support(R) = Support(R_1) \wedge Confiance(R) = Confiance(R_1)$;
2. $R_{a1} \subseteq R_a \wedge R_{c1} \subseteq R_c$.

La base générique exacte $\mathcal{GBE}$ [8] est définie comme l'ensemble réduit des règles ayant une confiance totale ($\forall R \in \mathcal{GBE}, Confiance(R) = 1$). $\mathcal{GBE}$ est définie comme suit :

$$\mathcal{GBE} = \{R : g \rightarrow (c - g) | c \in \mathcal{IFF} \wedge g \in \mathcal{G}_c \wedge g \neq c\} \tag{3.1}$$

où $\mathcal{IFF}$ est l'ensemble des fermés fréquents, $\mathcal{G}_c$ est l'ensemble des générateurs minimaux de c. D'après la formule, on peut constater que les règles exactes sont

construites à partir des itemsets fermés fréquents I. Le générateur minimal de I constitue notre prémisse et le reste représente la partie conclusion. Ainsi, avec une telle construction, la confiance est maximale.

3.2.1.2 Base générique de règles approximatives

Soient $\mathcal{IFF}$ l'ensemble des itemsets fermés fréquents et $\mathcal{G}$ l'ensemble des générateurs minimaux. La base générique de règles approximatives $\mathcal{GBA}$ est définie comme suit :

$$\mathcal{GBA} = \{R : g \rightarrow (c - g) | c \in \mathcal{IFF} \wedge g \in \mathcal{G} \wedge \omega(g) \subset c \wedge confiance(R) \geq minconf\} \tag{3.2}$$

où $\omega()$ est l'opérateur de fermeture. Contrairement à la base générique exacte, la base approximative relie plusieurs itemsets différents. Cela a pour effet une confiance inférieure à 1.

3.2.2 Base générique informative

L'étude de la littérature [8] a montré que seul le couple ($\mathcal{GBE}$, $\mathcal{GBA}$) est extrait sans perte d'information. Toutefois, le nombre de règles d'association générées est très important.

Soient $\mathcal{IFF}$ l'ensemble des itemsets fermés fréquents et $\mathcal{G}_c$ l'ensemble des générateurs minimaux de tous les itemsets fermés fréquents inclus ou égaux à un itemset fermé fréquent I. La base générique informative $\mathcal{IGB}$ est définie comme suit :

$$\mathcal{IGB} = \{R : g_s \Rightarrow (I - g_s) | I \in \mathcal{IFF} \wedge I - g_s \neq \emptyset \wedge g_s \in \mathcal{G}_c, c \in \mathcal{IFF} \wedge c \subseteq I \wedge confiance(R) \geq minconf \wedge \nexists g' | g' \subset g_s \wedge confiance(g' \Rightarrow I - g') \geq minconf\}. \tag{3.3}$$

Ainsi, les règles génériques de $\mathcal{IGB}$ représentent des implications entre des prémisses minimales et des conclusions maximales (en terme du nombre d'items). En effet, il a été prouvé, dans la littérature, que ces règles sont les plus générales (i.e., convoyant le maximum d'information) [8, 59].

Exemple 8 *Pour $minconf = 0.5$, le tableau 3.2 illustre la base de règles $\mathcal{IGB}$ retrouvée à partir du contexte d'extraction du tableau 3.1.*

3.2.3 Classification par les règles d'association génériques

La base de règles $\mathcal{IGB}$ présente un bon rapport compacité/informativité. En effet, elle garantit trois propriétés intéressantes. La première est la *couverture* large des règles qu'elle contient. Puisque les règles présentent une prémisse minimale et une conclusion maximale, elle couvre une variété large de cas de classification. Le

Trans ID	Attribut1	Attribut2	Classe
T1	P1	P2	*C1*
T2	P1	P4	*C1*
T3	P1	P2	*C1*
T4	P3	P4	*C2*
T5	P3	P2	*C2*
T6	P3	P2	*C2*
T7	P1	P2	*C1*
T8	P3	P2	*C2*
T9	P5	P2	*C3*

TAB. 3.1: Exemple d'un contexte d'extraction

Règle d'association	Confiance
$\emptyset \rightarrow P2$	0.78
$P1 \rightarrow \{P2, C1\}$	0.75
$C1 \rightarrow \{P1, P2\}$	0.75
$\{P2, P1\} \rightarrow C1$	1.00
$\{P2, C1\} \rightarrow P1$	1.00
$P1 \rightarrow C1$	1.00
$P3 \rightarrow \{P2, C2\}$	0.75
$C2 \rightarrow \{P2, P3\}$	0.75
$P3 \rightarrow C2$	1.00
$\{P2, P3\} \rightarrow C2$	1.00
$\{P2, C2\} \rightarrow P3$	1.00

TAB. 3.2: La base de règle $\mathcal{IGB}$ extraite à partir du contexte d'extraction du tableau 3.1

deuxième avantage de la classification par la base $\mathcal{IGB}$ est l'informativité [42] puisque le support et la confiance des règles dérivées sont déduites. Finalement, cette base offre l'avantage de la *compacité* par rapport aux autres bases [42]. Le classifieur *GARC*[1] repose sur la base $\mathcal{IGB}$. Il ne retient que les règles contenant un libellé de

[1] Generic Association Rules based Classifier

classe dans leur partie conclusion.

Exemple 9 *Le tableau 3.3 illustre les règles de classification extraites à partir de la base $\mathcal{IGB}$ du tableau 3.2.*

Règle d'association	Confiance
$P1 \rightarrow C1$	1
$P3 \rightarrow C2$	1

TAB. 3.3: Les règles de classification extraites à partir de la base $\mathcal{IGB}$

3.3 Gestion de conflit par les règles d'association

3.3.1 Motivation

La fusion de plusieurs sources d'information par la combinaison conjonctive génère du conflit. Les origines de son apparition ont été étudiées dans plusieurs travaux [85, 56, 70, 44, 52, 23, 95][2]. Une fois apparues, certaines approches privilégient la résolution du problème par un retour en arrière vers la partie crédale statique [70, 85]. Ceci a pour but de déterminer les sources à l'origine de la contradiction. Plusieurs travaux comme [85, 70] se sont intéressés à l'estimation des fiabilités des sources d'information à partir de leurs fonctions de masse respectives. Ces dernières utilisent les mesures de conflit intrinsèque [43, 110] et extrinsèque [104, 22] afin de déterminer la fiabilité de la source. Toutefois, ces approches peuvent présenter des limites dans certains cas de figure comme les sources singulières [57]. Une autre stratégie consiste à choisir un opérateur de combinaison différent dans le cas d'absence de certitude de la fiabilité des sources. L'opérateur de combinaison disjonctive peut être une solution. Cela-dit, une décision après combinaison peut s'avérer parfois difficile en raison de l'élément absorbant Θ. Pour cela, des informations supplémentaires sont parfois nécessaires afin d'affiner le résultat de la combinaison. Dans cette partie, nous supposons l'existence d'une information supplémentaire sous forme de base de données concernant la scène étudiée. Nous nous proposons de profiter de cette information afin de mieux affiner le résultat de la fusion conjonctive suite à la contradiction constatée. On peut envisager d'intégrer cette information supplémentaire à trois niveaux :

- **Approches d'affaiblissement** : L'information est utilisée pour déterminer les sources non fiables et elles sont affaiblies avant fusion [74].
- **Approches de fusion** : L'information supplémentaire est représentée sous forme de fonction de masse puis intégrée avec les autres sources existantes dans le processus de fusion [36].

[2]Les travaux d'affaiblissement que nous avons proposées dans [85, 87] sont récapitulés dans l'annexe B

– **Approches de redistribution du conflit** : Le conflit est conservé et l'information supplémentaire est utilisée pour redistribuer le conflit sur les hypothèses viables [64].

Dans notre problème, nous étudions le cas où l'information supplémentaire n'apporte aucune indication sur la fiabilité des sources déjà existantes. Dans ce cas, une approche de gestion de conflit par affaiblissement n'est pas envisageable. Nous supposons que cette information ne peut être transformée en une fonction de masse et intégrée dans le processus de fusion car elle n'est pas relative à la même scène et l'information qu'elle véhicule est de nature différente. Sachant ces hypothèses, comment extraire ces informations ? comment modéliser cette information ? et comment faire pour redistribuer le conflit ?
Dans ce qui suit, nous présentons notre solution qui consiste à modéliser l'information supplémentaire sous forme de règles d'association générées à partir d'un contexte d'extraction $\mathcal{K}$. Ces règles, qui se distinguent par des propriétés de généricité et compacité, sont utilisées pour redistribuer le conflit via l'opérateur de gestion de conflit présenté dans [63]. Cette méthode de gestion de conflit est dite contextuelle car elle dépend du contexte dans lequel nous nous situons et des règles d'association valides.

3.3.2 Cadre générique pour la gestion de conflit

Introduit par Lefèvre en 2001 [63], le cadre générique est une approche de gestion de conflit. Elle a été introduite dans le but d'unifier les opérateurs déjà existants et nombreux. Le but de cette famille d'opérateurs de combinaison est de redistribuer la masse conflictuelle K de manière locale sur un ensemble de proposition. La masse conflictuelle est redistribuée vers l'ensemble des propositions $A \subseteq \Omega$ selon un poids ω à déterminer. Ainsi, la fonction de masse, après gestion de conflit, s'écrit de la manière suivante :

$$\begin{cases} m(A) = m_{\textcircled{\cap}}(A) + m^c(A) & \forall A \subseteq \Omega \\ m(\emptyset) = 0 \end{cases} \tag{3.4}$$

où $m_{\textcircled{\cap}}(A)$ est le résultat de la combinaison conjonctive pour la proposition A. $m^c(A)$ est une partie de la masse conflictuelle K (i.e., $m_{\textcircled{\cap}}(\emptyset)$) attribuée à A et elle s'écrit de la manière suivante :

$$m^c(A) = \omega(A, m_1, \cdots, m_j, \cdots, m_J).K \quad \forall A \subseteq \Omega \tag{3.5}$$

tel que :

$$\sum_{A \subseteq \Omega} \omega(A, m_1, \cdots, m_j, \cdots, m_J) = 1. \tag{3.6}$$

La dernière somme garantit que toute la masse conflictuelle est redistribuée avec une répartition relative aux poids ω. Ainsi, le cadre générique permet de généraliser un certain nombre d'opérateurs de gestion de conflit comme ceux de Inagaki [49],

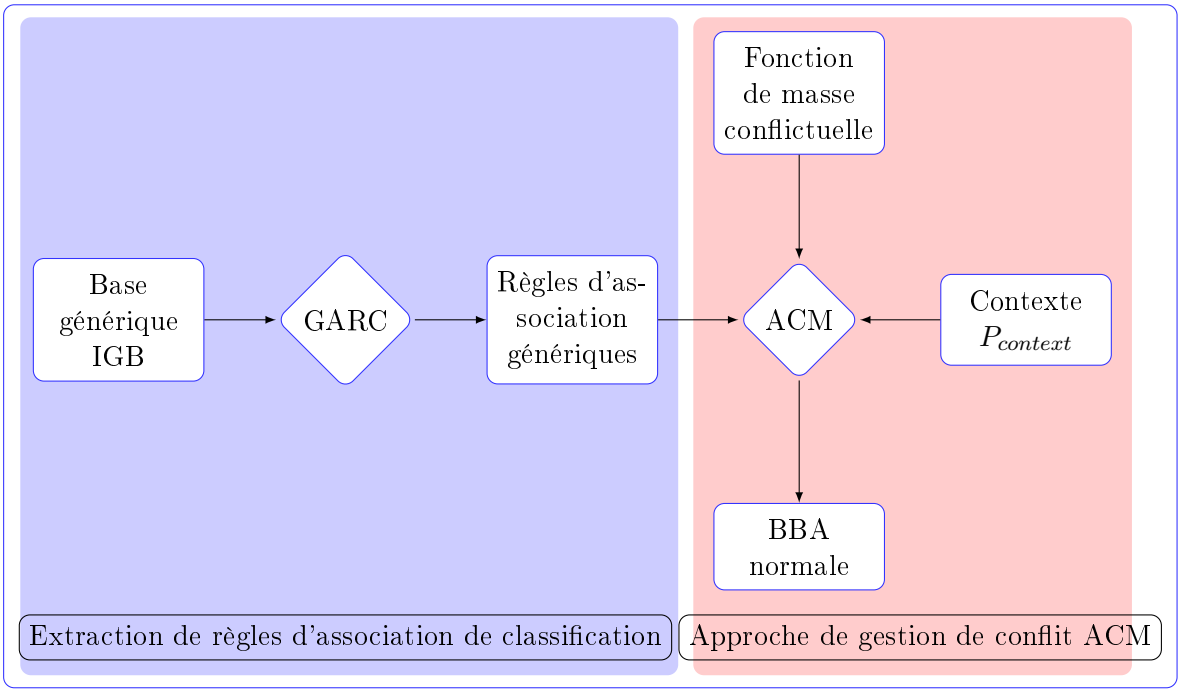

FIG. 3.1: Architecture de l'approche ACM

Smets [98], Yager [109] et Dubois et Prade [34]. Dans [63], il a été mis en évidence les valeurs particulières permettant de retrouver ces autres opérateurs de combinaison.

3.3.3 Approche de gestion de conflit associative

Dans cette partie, nous présentons l'approche ACM[3] [90], dont l'architecture est décrite dans la figure 3.1. L'approche consiste à utiliser une information supplémentaire sous forme de base de données afin de répartir le conflit entre hypothèses. La base de données est étudiée afin d'extraire des connaissances sous forme de règles d'association de classification. Nous avons fait le choix d'utiliser les règles d'association génériques pour les avantages déjà mentionnés dans la section 3.2.3. Dans un second niveau, nous appliquons une gestion de conflit contextuelle (selon le contexte étudié) sur la fonction de masse m en utilisant la base des règles construite. Comme il est indiqué dans la partie gestion de conflit par ACM dans figure 3.1, l'approche proposée consiste à sélectionner les règles intéressantes à partir du contexte étudié. Ces règles permettent de définir les valeurs de croyance transférées entre les hypothèses grâce aux facteurs de poids ω du cadre générique. Ainsi, dans la partie suivante, nous détaillons l'approche suivie pour le calcul des facteurs de poids du cadre générique à partir des règles d'association génériques. Un exemple applicatif concret est donné par la suite.

[3] en anglais Associative Conflict Management approach.

Dans ce qui suit, nous présentons notre méthode pour déterminer les valeurs de poids de la gestion de conflit par le cadre générique. Soit deux sources S_1 et S_2 pour lesquelles nous n'avons aucune information concernant leurs fiabilités. Le résultat de la combinaison conjonctive a donné lieu à une masse conflictuelle $m_{\cap}(\emptyset) > 0$.

Définition 1 *Soit une fonction de masse conflictuelle $m_{\cap}$ définie sur le cadre de discernement Ω et relative à un contexte d'extraction $\mathcal{D}$. Soit l'information supplémentaire $P_{context}$ relative à la scène étudiée et nécessaire pour la gestion de conflit. En réalité, $P_{context}$ est un itemset de taille fixe tel que $\forall i, i \subseteq P_{context}, i \subseteq \mathcal{I}$. Ayant les informations suivantes, la gestion de conflit par la méthode ACM s'écrit ainsi :*

$$\begin{cases} W_{ass}(C, m_{\cap}) = max\{Confiance(R) |\ \nexists R' \left|\frac{Pr(R')}{P_{context}}\right| > \left|\frac{Pr(R)}{P_{context}}\right| \wedge Cl(R) = C \wedge Cl(R') = C\} \\ W_{ass}(\Theta, m_{\cap}) = 1 - W_{ass}(C, m_{\cap}). \end{cases} \tag{3.7}$$

Il est important de noter que la redistribution du conflit ainsi décrite par l'équation 3.7 ne requiert que deux éléments focaux. Le premier élément focal est la classe représentée par la partie conclusion de la règle d'association utilisée. Le deuxième élément est l'ignorance totale sur lequel une partie des croyances est attribuée en cas d'absence d'informations. La confiance de la règle utilisée détermine la valeur du facteur de poids et ainsi la nature de la redistribution du conflit. Trois cas de figure peuvent se présenter. Le premier correspond au cas où la règle utilisée est une règle d'association exacte ($Confiance = 1$). Alors, toute la masse conflictuelle est transférée vers l'hypothèse constituant la conclusion de la règle. Si la règle d'association utilisée est une règle approximative ($Confiance < 1$), une partie de $m_{\cap}(\emptyset)$ est attribuée à l'élément focal représentée par la conclusion de la règle. Quand au reste, il est assigné à l'ignorance totale Θ. Le troisième cas de figure correspond au cas où aucune règle n'a d'intersection avec $P_{context}$. Dans ce cas, tout le conflit est attribué à l'ignorance et la gestion de conflit est équivaut à celle de Yager [109].

La propriété de l'unicité de facteur de poids est garantie telle que :

$$\sum_{A \subseteq \Theta} W_{ass}(A, m_{\cap}) = 1. \tag{3.8}$$

Ainsi le cadre générique peut s'écrire de la manière suivante :

$$m(A) = m_{\cap}(A) + W_{ass}(A, m_{\cap}) \cdot m_{\cap}(\emptyset) \qquad \forall A \subseteq \Theta$$

où P est l'ensemble des classes constituant les conclusions des règles associatives génériques (i.e. $Cl(R)$).

Exemple 10 *Dans cet exemple, nous étudions le problème de classification relative au contexte d'extraction du tableau 3.1. Soit la fonction de masse $m_{\cap}$ conflictuelle*

relative au cadre de discernement $\Omega = \{C1, C2, C3\}$ *:*

$$\begin{cases} m_{\textcircled{\cap}}(\{C3\}) &= 0.0001 \\ m_{\textcircled{\cap}}(\emptyset) &= 0.9999. \end{cases}$$

Après analyse, une information supplémentaire nous parvient nous informant que nous avons $P1$ *et* $P2$ *comme paramètre (i.e.,* $P_{context} = \{P1, P2\}$*). A partir de la base de règles de classification construite, nous avons la règle* $R1 : P1 \rightarrow C1$ *qui est plus générique que* $R2 : \{P1, P2\} \rightarrow C1$*. Les deux règles ont la même valeur de confiance 100%. Ainsi les valeurs réelles des facteurs de poids sont retrouvées de la manière suivante :*

$$W_{ass}(\{C1\}, m_{\textcircled{\cap}}) = 1$$

Ayant cette indication, l'application de ACM à la fonction de masse $m_{\textcircled{\cap}}$ *donne :*

$$\begin{cases} m(\{C1\}) = W_{ass}(\{C1\}, m_{\textcircled{\cap}}) \cdot m_{\textcircled{\cap}}(\emptyset) \\ m(\{C3\}) = m_{\textcircled{\cap}}(\{C3\}) + W_{ass}(\{C3\}, m_{\textcircled{\cap}}) \cdot m_{\textcircled{\cap}}(\emptyset) \\ m(\emptyset) = 0. \end{cases}$$

ce qui produit finalement :

$$\begin{cases} m(\{C1\}) = 0.9999 \\ m(\{C3\}) = 0.0001 \\ m(\emptyset) = 0. \end{cases}$$

3.4 Classification associative d'images forestières

Dans cette partie, nous nous intéressons à l'évaluation de l'approche de gestion de conflit associative dans le cadre applicatif de la classification d'images hautes-résolutions.

3.4.1 Modélisation du problème : Détermination de couronnes d'arbre

Nous considérons le problème de classification de couronnes d'arbre dans une image haute-résolution forestière. L'image traitée, comme l'illustre l'échantillon de la figure 3.2, représente la forêt de Ain-Drahim. Dans cette image, quatre types d'arbre sont majoritairement présents i.e., $\{Zen\ Oak, Cork\ Oak, Arboretum, Coniferous\ tree\}$. Cet ensemble constituera notre cadre de discernement. Plusieurs sources d'information peuvent être extraites à partir de l'image. Nous pouvons les classer en trois catégories :

- **Information spectrale** : Les couronnes d'arbre sont étudiées selon leurs niveaux de gris. Dans ce problème, chaque couronne d'arbre est représentée par

la valeur moyenne du niveau du gris des pixels qu'elle contient.

- **Information de texture** : Ceux sont les informations relevant de l'organisation des niveaux de gris dans l'image. Dans notre cas, nous nous sommes intéressés aux informations suivantes : moyenne, variance, énergie, contraste et entropie.
- **Information de structure** : Ceux sont les informations de structure de la couronne d'arbre comme la surface, le diamètre, le périmètre et l'ellipticité.

Vu la nature des sources dégagées et le facteur d'incertitude présent dans le processus de classification, nous avons utilisé la théorie des fonctions de croyance pour la classification. Les trois familles de sources d'information sont utilisées pour notre problème de fusion. Il existe, une quatrième source, fournissant une information tout aussi précieuse mais se distinguant des trois autres. Cette source est relative à :

- **Information Spatiale** : Elle étudie la disposition spatiale des couronnes et elle est retrouvée par l'analyse spectrale d'une région dans l'image.

La dernière source d'information n'a donc pas de relation avec les trois autres sources car elle n'entrevoit pas la même scène que les trois autres. L'intégrer au processus de fusion directement serait une erreur. De la même manière, l'information spatiale n'apporte aucunement une indication sur la fiabilité des autres sources. La solution la plus adéquate afin de profiter de cette information supplémentaire consiste à redistribuer le conflit.

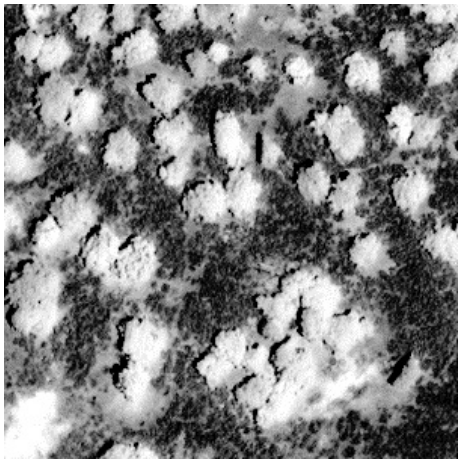

FIG. 3.2: Exemple d'une image forestière

3.4.2 Classification distance des couronnes d'arbre

Dans le chapitre 1, nous avons décrit les modélisations possibles et notamment celles fondées sur les vraisemblances et les distances. Le classifieur distance se distingue par sa simplicité et génère des fonctions de masse simple à deux éléments

Valeur du Conflit	**[0, 0.2)**	**[0.2, 0.4)**	**[0.4, 1)**
Pourcentage de nombre de fonctions de masse	12%	14%	74%

TAB. 3.4: Valeurs du conflit enregistrées pour les couronnes d'arbre

focaux. Cette particularité facilite, par la suite, l'opération de fusion et s'impose comme la solution adéquate vu le nombre de sources d'information que nous devons gérer. Dans un autre coté, élaborer une approche de vraisemblance, nécessite la connaissance apriori d'une fonction de vraisemblance. La solution que nous proposons repose donc sur le classifieur distance. Par ailleurs, ayant des sources de nature composite, la variante mono-dimensionnelle s'avère être le choix le plus judicieux.

Soit Θ le cadre de discernement tel que $\Theta = \{Zen\ Oak, Cork\ Oak, Arboretum, Coniferous\ tree\}$. Comme indiqué précédemment, quatre familles de sources d'information sont utilisées. Les familles de source *Spectrale*, de *Texture* et *Structurelle* sont utilisées dans l'étape de fusion d'information alors que l'information *Spatiale* est utilisée pour la redistribution du conflit par l'approche ACM.

Au sein de chaque famille, plusieurs caractéristiques peuvent être considérées. Celles-ci sont alors vues comme autant de sources d'information. Pour chaque couronne donnée, nous étudions ses K Plus Proches Voisins au vu de la caractéristique considérée (Equation (1.43)). Pour chaque caractéristique d'une famille de source, nous déterminons les 4 fonctions de masse ($K = 4$) issues du voisinage. Elles sont ensuite combinées par la règle de combinaison de Dempster (Equation (1.45)). Le résultat, une fonction de masse unique, indique l'appartenance de la couronne aux classes du cadre du discernement selon la caractéristique étudiée. La même opération est conduite sur les 10 caractéristiques que nous avons étudiées. Les fonctions de masse des caractéristiques sont fusionnées via l'équation 1.42. Cette classification est appelée DMC[4][84].

Afin d'étudier la nature conflictuelle des sources d'information choisies, nous appliquons une combinaison conjonctive (Equation (1.16)) au deuxième niveau de combinaison. Cette modification aura pour effet, d'une part d'afficher le degré de contradiction des sources et d'autre part de profiter de l'information spatiale en appliquant l'approche ACM. Il est à remarquer que ACM devient plus efficace lorsque le conflit devient important. Le tableau 3.4 montre le degré de conflit enregistré après combinaison. En effet, 74% des couronnes classées présentent une masse conflictuelle supérieur à 0.4. Un conflit avec un tel degré est susceptible de changer la classification initiale d'une couronne avec une redistribution du conflit vers une autre classe.

Dans ce travail, l'information spatiale est considérée comme une information supplémentaire. Une base de données relative à l'information spatiale des régions est construite. En effet, la base de données contient des informations de texture des régions d'apprentissage de chaque classe représentée dans le cadre de discernement. Cette base de données est étudiée avec les outils de fouille de données précédem-

[4]Distance Model Classifier

ment décrits pour retrouver les règles de classification génériques essentielles à la redistribution du conflit.

Dans l'approche de classification, chaque couronne est classée par fusion des sources selon la méthode décrite précédemment. L'approche ACM est appliquée par la suite sur toutes les fonctions de masse conflictuelles. Pour cela, la région entourant la couronne est étudiée et les informations de texture sont extraites. Ces informations sont celles que nous avons considérées dans la sous section 3.3.3 comme le $P_{context}$. Le $P_{context}$ sert à retrouver la règle à appliquer pour la redistribution de conflit. Une fois la redistribution effectuée, une décision est prise selon la probabilité pignistique pour déterminer la classe de la couronne étudiée.

3.5 Expérimentation et résultat

Dans cette partie, nous suivons une validation de l'approche proposée en deux phases. Dans un premier lieu, nous montrons l'intérêt d'utiliser les règles de la base $\mathcal{IGB}$ plutôt que d'autres existantes dans la littérature. Ensuite, nous nous intéressons aux résultats de classification de ACM en la comparant avec d'autres approches de classification par fusion de sources d'information.

3.5.1 Apport des règles d'association générique et la classification associative

Dans ce qui suit nous démontrons expérimentalement l'apport des bases génériques. Pour cela, nous comparons les performances de la base $\mathcal{IGB}$ dans la classification associative par rapport à l'approche référence CBA [68]. CBA est un classifieur qui filtre les règles d'associations issues de l'algorithme Apriori. Les tests sont conduits sur la base de données construite à partir des informations spatiales collectées de l'image. La base de données contient 546 instances.

En terme de nombre de règles générées, *GARC* présente un avantage conséquent par rapport à l'algorithme CBA comme l'illustre la figure 3.3

La base de règles génériques présente non seulement un nombre moindre de règles mais aussi retient celles de meilleures qualités. En effet, grâce à l'élagage des règles redondantes, $\mathcal{IGB}$ ne retient que celles ayant la plus petite prémisse.
Le classifieur *GARC* présente des résultats compétitifs par rapport à d'autres approches connues en classification. Les expérimentations conduites sur un benchmark démontrent l'efficacité de *GARC* [15]. En effet, il présente de meilleurs résultats que, par exemple, les arbres de décision [15]. Dans notre cas d'étude, *GARC* nous assure non seulement de bons résultats de classification mais aussi un degré d'appartenance à la classe sous forme d'une mesure de confiance.

3.5.2 Apport dans la classification et la gestion de conflit

Dans cette partie, nous comparons notre approche de classification par redistribution de conflit à d'autres approches classiques. Les approches de classification ont

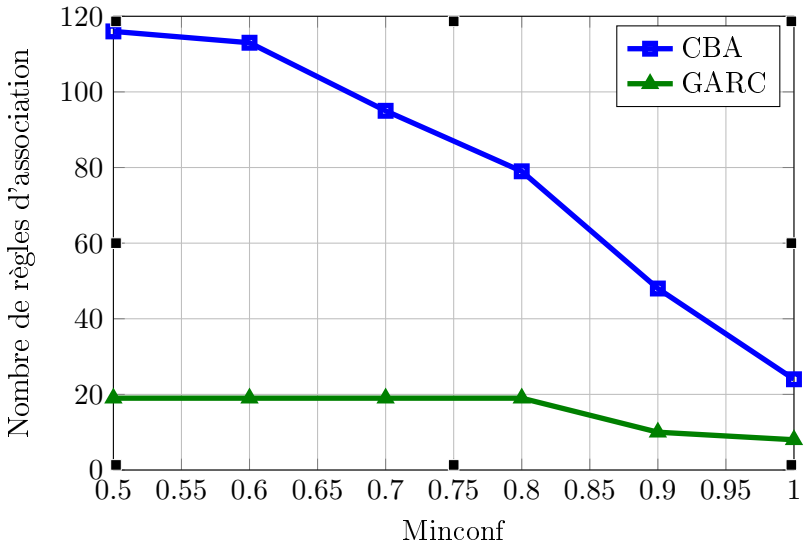

FIG. 3.3: Nombre de règles d'association générées pour $minsup = 0.1$

été testées sur une base de 399 couronnes d'arbre. L'approche ACM de classification a été comparée au classifieur distance DMC décrit dans la sous-section 1.6.2. Ce classifieur évidentiel gère le conflit grâce la combinaison de Dempster qui répartit le conflit proportionnellement aux masses des hypothèses. Cette comparaison aura pour but de montrer l'apport de la gestion de conflit contextuelle dans l'amélioration des résultats. Nous avons également comparé ACM à un classifieur évidentiel, appelé DMCE, qui intègre l'information supplémentaire. DMCE repose sur les mêmes bases de DMC mais se distingue en intégrant l'information supplémentaire dans son système de fusion de sources. Pour cela, une fonction de masse est construite à partir de l'information spatiale de l'image où nous avons adopté la modélisation distance fondée sur les prototypes. Cela dit, le cadre discernement de cette dernière, noté Θ_r, est différent de celui des trois autres sources et contient $\Theta_r = \{Zen\ Oak_r, Cork\ Oak_r, Arboretum_r, Coniferous\ tree_r\}$. Afin de l'intégrer au processus de fusion, nous devons unifier leurs cadres de discernement et pour cela une extension vide est opérée. L'extension vide d'une fonction de masse m de Θ_r à Θ peut s'effectuer de la manière suivante :

$$m^{\Theta_r \downarrow \Theta}(\rho(B)) = m^{\Theta_r}(B) \quad \forall B \subseteq \Theta_r. \tag{3.9}$$

où ρ est la transformation de Θ_r à Θ qui est définie dans notre cas par :

$$\begin{cases} \rho(\{Zen\ Oak_r\}) = \{Zen\ Oak\} \\ \rho(\{Cork\ Oak_r\}) = \{Cork\ Oak\} \\ \rho(\{Arboretum_r\}) = \{Arboretum\} \\ \rho(\{Coniferous\ tree_r\}) = \{Coniferous\ tree\} \end{cases}$$

Une fois que les fonctions de masse sont définies sur le même cadre, la combinaison devient possible. L'opération d'extension n'est possible que lorsque le groupement d'arbres (région) étudié ne contient qu'un seul type d'arbre.

	Zen Oak			Cork Oak			Arboretum			Coniferous tree		
Classifieur	ACM	DMC	DMCE	ACM	DMC	DMCE	ACM	DMC	DMCE	ACM	DMC	DMCE
Zen Oak	**95.19%**	80.76%	86.20%	4.80%	15.38%	11.03%	0.00%	0.00%	0.00%	0.01%	3.86%	2.77%
Cork oak	12.65%	29.11%	30.17%	**78.48%**	50.63%	53.07%	6.32%	12.65%	10.61%	2.55%	7.61%	6.15%
Arboretum	2.08%	4.13%	7.14%	11.03%	28.27%	14.28%	**71.03%**	35.86%	41.83%	15.86%	31.74%	36.75%
Coniferous tree	0.00%	3.82%	17.39%	7.08%	29.89%	18.47%	19.56%	32.60%	28.81%	**73.36%**	33.69%	35.33%

TAB. 3.5: Comparatif de performance : ACM vs DMC

Comme démontré dans le tableau 3.5, l'approche ACM présente de meilleurs résultats que l'approche de fusion classique DMC pour toutes les classes. Cette amélioration s'explique par la gestion de conflit contextuelle de chaque couronne relativement à la région dans laquelle elle se trouve. Ceci démontre également l'intérêt de la source information spatiale dans la classification. L'amélioration est aussi conséquente en comparant ACM à DMCE, démontrant l'intérêt de la gestion de conflit par redistribution. Ce résultat est obtenu grâce à l'apport des règles d'association dans le choix de l'hypothèse sur laquelle le conflit va être transféré. Quand aux résultats de l'approche DMCE, ils sont expliqués par le même poids des sources avant fusion. En effet, les quatre sources fusionnées sont toutes considérées fiables. Ainsi, le résultat peut être erroné même dans le cas où la source d'information spatiale apporte des informations correctes.

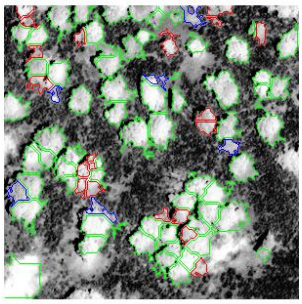

FIG. 3.4: Région *ZenOak* classée avec l'approche DMC.

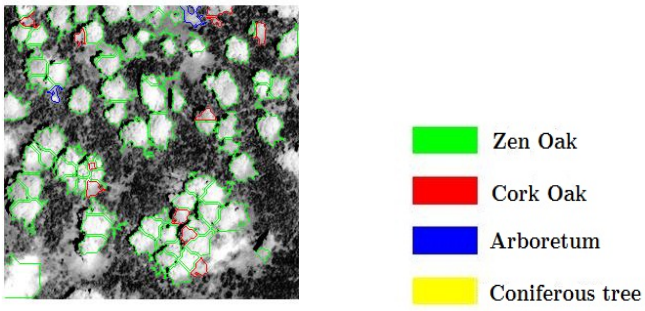

FIG. 3.5: Région *ZenOak* classée avec l'approche ACM.

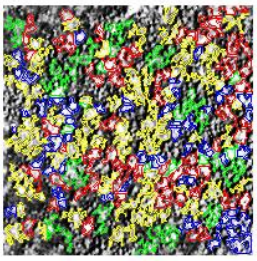

FIG. 3.6: Région *Coniferoustree* classée avec l'approche DMC.

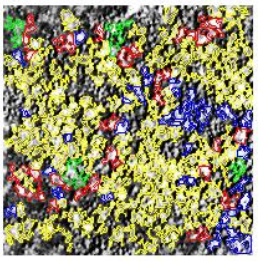

FIG. 3.7: Région *Coniferoustree* classée avec l'approche ACM.

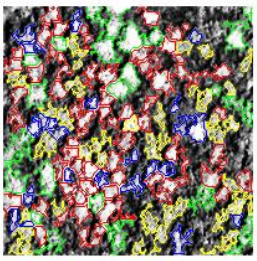

FIG. 3.8: Région *CorkOak* classée avec l'approche DMC.

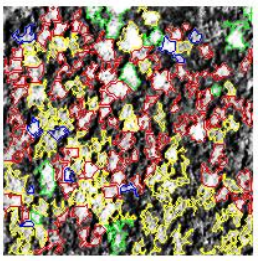

FIG. 3.9: Région *CorkOak* classée avec l'approche AMC.

La figure 3.4 montre une région ne contenant que des arbres appartenant la classe *ZenOak* classée par l'approche DMC. Dans cette image, les classes *Zenoak*, *CorkOak*, *Arboretum* et *Coniferoustree* sont colorées en *vert*, *rouge*, *bleu* et *jaune*. En la comparant à la figure 3.5, nous pouvons voir que ACM a réellement amélioré la classification où certaines couronnes initialement classées *CorkOak* et *Arboretum* ont changé vers la classe *ZenOak* ce qui est le cas en vérité. La comparaison a été faite sur des images contenant les autres types d'arbres comme l'illustre les figures 3.6, 3.7, 3.8 et 3.9. Même dans ces images les améliorations de ACM sont clairement visibles.

3.6 Conclusion

Dans les chapitres précédents, nous avons vu les limites des algorithmes d'extraction des motifs et de règles d'association de fouille de données. La limite majeure est le nombre important des règles retrouvées. La majorité de ces règles sont redondantes et n'apportent aucune information supplémentaire. Dans ce chapitre, nous avons étudié les règles d'association génériques qui présentent l'avantage d'être

de taille très inférieure aux règles générées par les approches classiques. En terme d'informativité, les règles génériques sont équivalentes à celles générées par Apriori (sans aucun élagage). Ces règles ont été utilisées pour étudier l'apport de la fouille de données dans la gestion de conflit dans le cadre de la théorie des fonctions de croyance.

Nous nous sommes, plus particulièrement, intéressés au problème de classification de couronnes d'arbre dans une image haute-résolution. Dans ce cas d'étude, nous avons enregistré la présence d'une source d'information supplémentaire. Cette information, présente sous forme de base de données a été étudiée par les outils de fouille de données et nous avons extrait les connaissances sous forme de règles génériques. Ces règles sont par la suite employées dans une approche de gestion de conflit que nous avons appelé ACM. Cette approche associe les confiances de règles d'association retenues à un cadre générique de gestion de conflit. Les résultats de ACM améliorent ceux des approches classiques. La contribution présentée dans ce chapitre matérialise l'apport de la fouille de données dans la résolution d'un problème propre à la théorie des fonctions de croyance. Dans le chapitre suivant, nous étudierons la relation inverse c'est à dire l'apport de la théorie des fonctions de croyance dans la fouille de données.

Chapitre 4

Estimation des supports des motifs fréquents dans les bases évidentielles

Sommaire

4.1 Introduction

La majorité des algorithmes de fouille de données a été appliqué sur des bases de données précises et certaines. Dans ces bases de données, seule la présence d'un item est indiquée. En effet, dans les bases binaires un attribut peut avoir deux valeurs possibles : 0 si l'attribut n'existe pas et 1 dans le cas contraire. Malheureusement, dans la réalité, les informations rassemblées souffrent d'imperfection pour de multiples raisons comme la fiabilité des capteurs, les erreurs de mesure humaine, l'absence d'informations, etc.
Dans le chapitre 2, nous avons fait l'inventaire des bases de données permettant de gérer les données imparfaites. Parmi ces bases, nous retrouvons les bases de données évidentielles [61, 60]. En effet, grâce à la théorie des fonctions de croyance, les bases de données évidentielles représentent parfaitement les informations *imprécises* et *non fiables*.
Dans ce chapitre, nous nous intéressons à la fouille de données dans les bases de données évidentielles. Nous présentons, dans un premier temps, une simplification

de la méthode de calcul de support existante. Cette modification permet d'améliorer les performances, en terme de temps de calcul, des algorithmes déjà présents dans la littérature. Malgré cette amélioration, la mesure du support actuelle a des limites qui seront énoncées dans ce chapitre. Ces limites donneront lieu à la deuxième partie de ce chapitre qui sera consacrée à la présentation d'une nouvelle mesure d'estimation du support que nous noterons : *mesure de support précis*. L'apport de cette mesure sera justifié par une étude expérimentale sur des bases de données évidentielles construites à partir de benchmark.

4.2 Estimation du support dans les bases évidentielles

Les bases de données évidentielles permettent de représenter les données souffrant d'imprécisions et d'imperfections. Les apports de ces bases de données ont été détaillées dans le tableau 2.8. Les données sont représentées sous forme de fonctions de masse comme dans l'équation 2.12. Dans cette partie, nous détaillons les définitions de base de la fouille de données évidentielles et les différentes mesures introduites dans la littérature. Pour cela, nous considérons l'exemple de base de données évidentielles du tableau 4.1.

Transaction	Attribut A	Attribut B
T1	$m_{11}(A_1) = 0.7$	$m_{21}(B_1) = 0.4$
	$m_{11}(\Theta_A) = 0.3$	$m_{21}(B_2) = 0.2$
		$m_{21}(\Theta_B) = 0.4$
T2	$m_{12}(A_2) = 0.3$	$m_{22}(B_1) = 1$
	$m_{12}(\Theta_A) = 0.7$	

TAB. 4.1: Exemple de base de données évidentielle $\mathcal{EDB}$

Contrairement aux bases de données binaires les colonnes ne représentent plus les items mais les attributs (les questions). L'item, en revanche, est un élément focal correspondant à un attribut donné. Un itemset correspond à la conjonction d'éléments focaux appartenant à des attributs différents.

Exemple 11 *Dans le tableau 4.1,* A_1 *est un item évidentiel et* $\{\Theta_A\ B_1\}$ *un itemset tel que* $A_1 \subset \{\Theta_A\ B_1\}$ *et* $A_1 \cap \{\Theta_A\ B_1\} = A_1$.

Dans le cadre de la fouille de données évidentielles, deux types d'opérateur peuvent être définis : l'inclusion et l'intersection. Soient deux itemsets évidentiels X et Y, l'opérateur d'inclusion est défini comme suit :

$$X \subseteq Y \iff \forall x_i \in X, x_i \subseteq y_i.$$

où x_i et y_i sont les i^{imes} éléments de respectivement X et Y. Pour les mêmes itemsets X et Y, l'intersection est défini de la manière suivante :

$$X \cap Y = Z \iff \forall z_i \in Z, z_i \subseteq x_i \text{ et } z_i \subseteq y_i.$$

Une règle d'association évidentielle R est une relation causale entre deux itemsets qui peut s'écrire de la manière suivante : $R : X \rightarrow Y$ où $X \cap Y = \emptyset$.

Exemple 12 *$A_1 \rightarrow B_1$ est une règle d'association évidentielle.*

Dans cette partie, nous présentons les différents travaux qui se sont intéressés à associer la théorie des fonctions de croyance pour la représentation des informations dans un problème de fouille de données. Le concept de fouille de données evidentielles n'a été étudié que vers la fin des années 2000. L'une des premières approches associant les deux domaines a été proposée par Shy et al. [94]. Ces auteurs ont présenté une approche de génération de règles d'association. La théorie des fonctions de croyance a été alors utilisée que pour exprimer la relation de corrélation entre les attributs.
C'est en 2005 que la première approche d'estimation du support dans les bases évidentielles a vu le jour. Hewawasam et al. [47] ont non seulement proposé une mesure de support crédibiliste mais aussi une représentation des fréquents sous forme d'arbre. L'arbre des items evidentiels (BIT)[1] apporte rapidité et simplicité dans la détermination du support des itemsets. Bach-Tobji et al. [7] ont introduit une amélioration de l'approche BIT dans le cadre d'un problème de maintenance des itemsets fréquents. L'algorithme FIMED [7] est fondé sur la même mesure de support que nous allons détaillé dans ce qui suit.

Soit l'itemset évidentiel $X = \prod_{i \in [1...n]} x_i$ pour lequel nous calculons le support à partir de différentes méthodes [7, 47, 46]. x_i est un item évidentiel appartenant au cadre de discernement Θ_i. Puisque aucun des items x_i de X ne partage le même cadre de discernement que les autres, l'emploi d'une règle de combinaison directe n'est pas possible.

Une première fonction de masse m_j relative à la transaction numéro j est calculée de la manière suivante :

$$m_j(X) = \prod_{x_i \in X} m_{ij}(x_i) \tag{4.1}$$

où $m_j(X)$ est la fonction de masse issue du produit Cartésien de toutes les BBA de la transaction T_j. Ainsi, la fonction de masse de l'itemset X dans tout $\mathcal{EDB}$ devient :

$$m_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} m_j(X). \tag{4.2}$$

Le support de X dans $\mathcal{EDB}$ est déduit de la manière suivante :

$$Support_{\mathcal{EDB}}(X) = Bel_{\mathcal{EDB}}(X). \tag{4.3}$$

La mesure de support basée sur le produit Cartésien satisfait quelques propriétés

[1] en anglais Belief Itemset Tree

intéressantes comme l'*anti-monotonie*. Cette propriété soutient que chaque itemset construit à partir d'un autre non fréquent, l'est aussi. L'opposé est tout aussi valable, tous les itemsets inclus dans un autre fréquent le sont aussi. Ainsi, grâce à cette propriété, plusieurs versions de génération de fréquents par niveau ont vu le jour [7, 46].

4.3 Ramification des supports crédibilistes

Un des problèmes majeurs de l'approche d'estimation du support des itemsets évidentiels est la complexité. En effet pour calculer le support d'un itemset, il faut recourir au produit Cartésien (Équation 4.1) pour chaque transaction et qui complexifie le calcul. Dans ce qui suit, nous proposons une autre écriture du support crédibiliste afin de réduire la complexité du calcul pour un résultat identique.

Proposition 1 *Soit la base évidentielle $\mathcal{EDB}$ et l'itemset $X = x_1 \times \cdots \times x_n$ le produit des items (éléments focaux) x_i $(1 \leq i \leq n)$ appartenant aux cadres de discernement exclusifs Θ_i. Pour la transaction T_j, nous avons :*

$$Support_{T_j}(X) = \prod_{i \in [1 \ldots n]} Support_{T_j}(x_i) = \prod_{i \in [1 \ldots n]} Bel(x_i) \tag{4.4}$$

$$Support_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} Support_{T_j}(X) \tag{4.5}$$

Preuve 1 *Soient deux items évidentiels x_i tels que $i = 1, \cdots, n$ appartenant respectivement aux fonctions de masse m_i $(i = 1, \cdots, n)$ telles que $m_{1 \times \cdots \times n} = m_1 \times \cdots \times m_n$.*

$$Bel(\prod_{x_i \in \Theta_i, 1 \leq i \leq n} x_i) = \sum_{a \subseteq x_1 \times \cdots \times x_n} m_{1 \times \cdots \times n}(a)$$

$$Bel(\prod_{x_i \in \Theta_i, 1 \leq i \leq n} x_i) = \sum_{y_1 \subseteq x_1, \ldots, y_2 \subseteq x_n} m_1(y_1) \times \cdots \times m_n(y_2)$$

$$Bel(\prod_{x_i \in \Theta_i, 1 \leq i \leq n} x_i) = \sum_{y_1 \subseteq x_1} m_1(y_1) \times \cdots \times \sum_{y_n \subseteq x_n} m_n(y_n)$$

$$Bel(\prod_{x_i \in \Theta_i, 1 \leq i \leq n} x_i) = Bel(x_1) \times \cdots \times Bel(x_n) = \prod_{i \in [1 \ldots n]} Bel(x_i)$$

Malgré l'apport indéniable de la ramification du support crédibiliste en terme de rapidité de calcul, d'autres limites existent. Parmi ces limites, on trouve la précision de la mesure. C'est dans ce cadre, que nous détaillerons dans la suite une nouvelle mesure de support que nous appellerons *la mesure du support précis*.

4.4 Support précis

Le support crédibiliste présente plusieurs limites. Dans la section 4.3, nous avons souligné la complexité de calcul afin de retrouver le support d'un itemset. Cette limite réside essentiellement dans l'écriture du support qui nécessite de calculer le produit Cartésien pour chaque transaction. A notre avis, la contrainte la plus évidente du support crédibiliste réside dans sa formulation. Dans l'exemple 13, nous donnons un exemple pratique de la limite observée.

Exemple 13 *Soit la base de données évidentielles $\mathcal{EDB}$ du tableau 4.1, on se propose de retrouver la support crédibiliste de l'item A_1 :*

$$Support_{\mathcal{EDB}}(A_1) = \frac{1}{2}\sum_{j=1}^{2} Support_{T_j}(A_1)$$

$$Support_{\mathcal{EDB}}(A_1) = \frac{1}{2}(Bel_1(A_1) + Bel_2(A_1)) = 0.35$$

Le support, ci-dessus indiqué, est retrouvé à partir des croyances allouées à A_1. Ceci est dû à la nature pessimiste de la fonction de crédibilité $Bel(.)$. Or, en réalité, cette assertion, n'est pas vraie car une part de croyance de A_1 réside dans des hypothèses plus larges. Dans notre exemple, par définition de l'hypothèse ignorance Θ_A, intègre A_1 où $A_1 \subseteq \Theta_A$.

Soit la base de données évidentielles $\mathcal{EDB}$ et un itemset $X = x_1 \times \cdots \times x_n$, le produit de plusieurs items (éléments focaux) x_i $(1 \leq i \leq n)$ du cadre de discernement Θ_i. Le degré de présence de l'item x_i dans une transaction T_j peut être mesuré de la façon suivante :

$$Pr : 2^{\Theta} \rightarrow [0, 1] \tag{4.6}$$

tel que :

$$Pr(x_i) = \sum_{x \subseteq \Theta_i} \frac{|x_i \cap x|}{|x|} \times m(x) \qquad \forall x_i \in 2^{\Theta_i} \tag{4.7}$$

où $Pr(.)$ calcule une probabilité sur une fonction de masse unique. Nous pouvons clairement voir que la fonction Pr est équivalente à la probabilité pignistique dans le cas où $x_i \in \Theta_i$. Le support évidentiel pour un itemset $X = \prod_{i \in [1...n]} x_i$ est alors calculé de la manière suivante :

$$Support^{Pr}_{T_j}(X) = \prod_{X_i \in \Theta_i, i \in [1...n]} Pr(X_i) \tag{4.8}$$

$$Support_{\mathcal{EDB}}(X) = \frac{1}{d}\sum_{j=1}^{d} Support^{Pr}_{T_j}(X). \tag{4.9}$$

De cette façon, la mesure de support précis considère plus d'éléments focaux dans son calcul que l'approche crédibiliste [7, 46]. En effet, contrairement à la fonction crédibilité, la fonction $Pr(.)$ assimile dans son estimation du support de X l'ensemble lui même ainsi que les ensembles qui lui sont plus larges. Ainsi, une partie des croyances des hypothèses plus larges que X est additionnée à la mesure du support. Cette écriture de support généralise plusieurs autres mesures proposées dans le cadre binaire, probabiliste et même flou [89][2].

Proposition 2 *La mesure du support précis vérifie la propriété de l'anti-monotonie. Soient deux itemsets evidentiels A et B tels que $A \subseteq B$:*

$$Support(A) \leq Support(B) \tag{4.10}$$

Preuve 2 *Soit une base évidentielle $\mathcal{EDB}$, soient les deux items évidentiels A et $A \times X$ tel que $A \subset A \times X$ vérifiant $\forall x \in A,\ x \in A \times X$. Pour que $Pr(.)$ soit anti-monotone, il faut vérifier l'assertion suivante : $Support(A \times X) \leq Support(A)$.*

$$Support_{T_j}(A \times X) = Pr(A) \times Pr(X)$$

$$Support_{T_j}(A \times X) \leq Support_{T_j}(A) \text{ où } Pr(X) \in [0,1] \text{ alors}$$

$$Support_{\mathcal{EDB}}(A \times X) \leq Support_{\mathcal{EDB}}(A)$$

En plus de l'apport en précision qu'apporte le support précis, il offre aussi une simplicité dans le calcul. En effet, le support précis s'écrit comme le produit des $Pr(.)$ des éléments focaux comme la ramification crédibiliste. Ainsi, nous évitons de calculer le produit Cartésien des fonctions de masse. Afin d'optimiser le calcul, il est intéressant de stocker les $Pr(.)$ des éléments focaux évitant de retourner à la base de données à chaque fois qu'il est nécessaire. Ce tableau est appelé la *table Pr*.

Exemple 14 *Le tableau 4.2 constitue la table Pr construite à partir de la base évidentielle $\mathcal{EDB}$ (tableau 4.1).*

Le support de l'itemset $A_1 \times B_1$ est calculé de la manière suivante :

$$Support^{Pr}_{T_1}(A_1 \times B_1) = Pr^{\Theta_A}(A_1) \times Pr^{\Theta_B}(B_1) = 0.51$$

$$Support^{Pr}_{T_2}(A_1 \times B_1) = Pr^{\Theta_A}(A_1) \times Pr^{\Theta_B}(B_1) = 0.35$$

$$Support_{\mathcal{EDB}}(A_1 \times B_1) = \frac{1}{2} \sum_{j=1}^{2} Support^{Pr}_{T_j}(A_1 \times B_1) = \frac{0.51+0.35}{2} = 0.43$$

Le support précis apporte un gain supplémentaire par rapport au support crédibiliste en terme de précision de calcul. En effet, la mesure du support est dorénavant

[2] Le papier prouvant l'avantage de la généralisation du support précis est présenté dans l'annexe B.2

Transaction	Support transactionnel
T1	$Pr^{\Theta_A}(A_1) = 0.85$
	$Pr^{\Theta_A}(A_2) = 0.15$
	$Pr^{\Theta_A}(\Theta_A) = 1.00$
	$Pr^{\Theta_B}(B_1) = 0.60$
	$Pr^{\Theta_B}(B_2) = 0.40$
	$Pr^{\Theta_B}(\Theta_B) = 1.00$
T2	$Pr^{\Theta_A}(A_1) = 0.35$
	$Pr^{\Theta_A}(A_2) = 0.65$
	$Pr^{\Theta_A}(\Theta_A) = 1.00$
	$Pr^{\Theta_B}(B_1) = 1.00$
	$Pr^{\Theta_B}(B_2) = 0.00$
	$Pr^{\Theta_B}(\Theta_B) = 1.00$

TAB. 4.2: Le tableau Pr déduit à partir de la base évidentielle $\mathcal{EDB}$ présentée dans le tableau 4.1

calculée en prenant en compte tous les éléments focaux présents. En plus, il vérifie la propriété d'anti-monotonie qui permet d'envisager la conception d'algorithme d'extraction de motifs fréquents à base d'élagage. Dans la section suivante, nous introduirons un algorithme d'extraction de motifs évidentiels fréquents reposant sur les bases de l'algorithme Apriori.

4.5 Algorithme Evidentiel de Data Mining (EDMA-p)

L'algorithme Apriori représente un des algorithmes de référence dans la fouille de données. Malgré les limites qu'il affiche (détaillées dans le chapitre 2), il est capable de retrouver tous les motifs fréquents grâce à une heuristique simple par niveau. C'est pour cela, qu'il est la base de la majorité des algorithmes d'exploration des fréquents dans les bases de données imparfaites (probabiliste, floue, évidentielle) [48, 19, 6]. C'est dans cette perspective que nous avons également employé le principe d'Apriori pour concevoir notre propre algorithme de recherche de motifs évidentiels fréquents. Cet algorithme, nous l'avons nommé EDMA-p[3]. Dans ce qui suit, nous détaillons les différents modules le constituant.

Un premier pré-traitement sur la base de données est nécessaire avant d'entamer la recherche des fréquents. Cette étape consiste à construire la Table Pr correspon-

[3]en anglais Evidential Data Mining Algorithm-pattern version

dante à la base de données étudiée en vue d'optimiser par la suite le temps de calcul du support. L'algorithme 2 présente la méthode entreprise pour la construction de la Table Pr. L'algorithme contient la fonction *compute_Pr* prenant comme paramètres une fonction de masse BBA et un item évidentiel (élément focal) et retourne son support précis transactionnel (i.e., $Pr_{T_j}(A)$). Le calcul se fait par exploration de tout l'espace de recherche des éléments focaux ayant une intersection non nulle avec A. La valeur finale du support transactionnel de A n'est autre que l'application de l'équation 4.7. Le même calcul est répété pour tous les items de la base de données. La Table Pr finale est stockée dans la variable PT.

Algorithm 2 Algorithme de construction de la Table Pr

Require: $\mathcal{EDB}, size\_\mathcal{EDB}, size\_attri$

Ensure: PT

function COMPUTE_PR(BBA , A)
  $Pr \leftarrow 0$
  for all $foc\_elt \in BBA.focal\_element$ **do**
    if $foc\_elt \cap A \neq \emptyset$ **then**
      $Pr \leftarrow Pr + \frac{|A \cap foc\_elt|}{|foc\_elt| \times (1-K)} \times BBA.mass$
    end if
  end for
  return Pr
end function
for $i = 1$ *to* $size\_\mathcal{EDB}$ **do**
  for $j = 1$ *to* $size\_attri$ **do**
    for all $foc\_elt \in \mathcal{EDB}(i,j).focal\_element$ **do**
      $PT(i).focal\_element \leftarrow foc\_elt$
      $PT(i).Pr\_value \leftarrow Compute\_Pr(\mathcal{EDB}(i,j), foc\_elt)$
    end for
  end for
end for

L'algorithme 3 introduit la fonction d'estimation du support précis. La fonction *Support_estimation* prend comme paramètre la table Pr PT, l'itemset I pour lequel nous cherchons le support et finalement le nombre de transactions dans la base de données d. Puisque cette fonction nécessite la table Pr, il est impératif de commencer par l'algorithme 2. En effet, le support transactionnel est fourni par la variable PT qui est consultée chaque fois qu'un item constituant notre itemset I est étudié. Finalement, le support final n'est autre que la somme des supports transactionnels de I que nous normalisons par d.

EDMA-p, détaillé dans l'algorithme 4, décrit l'approche employée pour la re-

Algorithm 3 Fonction d'estimation du support précis

```
1: function SUPPORT_ESTIMATION(PT , I, d)
2:     Sup_I ← 0
3:     for j=1 to d do
4:         Sup_Trans ← 1
5:         for all i ∈ Pr(j).focal_element do
6:             if Pr(j).focal_element ∈ I then
7:                 Sup_Trans ← Sup_Trans × Pr(j).value
8:             end if
9:         end for
10:        Sup_I ← Sup_I + Sup_Trans
11:    end for
12:    return Sup_I / d
13: end function
```

cherche des motifs évidentiels fréquents. Tout comme Apriori [3], EDMA-p explore les fréquents niveau par niveau profitant de l'anti-monotonie de la mesure du support précis. L'algorithme débute par la génération des fréquents de niveau 1, ensuite la fonction *Frequent_itemset* retourne tous les fréquents de taille k. Ainsi, cette fonction prend en argument la table Pr, un minsup et les candidats.

4.6 Expérimentations et résultats

Dans cette partie, nous étudions les résultats de l'approche d'exploration des motifs évidentiels fréquents. Pour expérimenter EDMA-p, il est impératif de disposer d'une base de données évidentielles. L'inconvénient majeur en travaillant sur cette thématique est l'absence d'un benchmark permettant d'effectuer les tests nécessaires. Même dans les travaux de référence dans le domaine comme [7, 46], l'absence d'une base de données évidentielles a été l'une des difficultés majeures. En effet, dans [46], les auteurs ont utilisé une base de données militaires évidentielles. Dans [7], les auteurs ont expérimenté l'algorithme de recherche de motifs fréquents sur une base construite à partir d'un questionnaire conduit sur des étudiants. Cette base de données est considérée comme partiellement évidentielle vu qu'un seul attribut est représenté avec des fonctions de masse. Dans ce qui suit, nous présentons une méthode permettant de construire une base de données évidentielles à partir de benchmark. Les bases évidentielles, ainsi obtenues, serviront de tests par la suite.

Algorithm 4 Algorithme EDMA-p

Require: $\mathcal{EDB}, minsup, PT, Size\_\mathcal{EDB}$

Ensure: $\mathcal{EIFF}$

function FREQUENT_ITEMSET($candidate$, $minsup$, PT, $Size\_\mathcal{EDB}$)
  $frequent \leftarrow \emptyset$
  for all x in $candidate$ **do**
    if $Support\_estimation(PT, x, Size\_\mathcal{EDB}) \geq minsup$ **then**
      $frequent \leftarrow frequent \cup \{x\}$
    end if
  end for
  return $frequent$
end function
$\mathcal{EIFF} \leftarrow \emptyset$
$size \leftarrow 1$
$candidate \leftarrow candidate\_apriori\_gen(\mathcal{EDB}, size)$
While $(candidate \neq \emptyset)$
$freq \leftarrow Frequent\_itemset\ (candidate, minsup, PT, Size\_\mathcal{EDB})$
$size \leftarrow size + 1$
$\mathcal{EIFF} \leftarrow \mathcal{EIFF} \cup freq$
$candidate \leftarrow candidate\_apriori\_gen(\mathcal{EDB}, size, freq)$
End While

4.6.1 Construction de la base de données évidentielles

Les bases de données évidentielles représentent un pas en avant en terme de représentation des connaissances car elles permettent de stocker des opinions par rapport à des questions. Comme exemple réel, nous pouvons imaginer le contexte d'une base de données où sont stockées des informations sur des patients. Les informations peuvent être : les diagnostics et les symptômes pour l'ensemble des patients. Une telle base de connaissance peut servir par la suite pour retrouver une corrélation entre les symptômes et le diagnostic de la maladie d'un patient.
Dans cette partie, nous détaillons la méthode de construction d'une base de données évidentielles à partir d'une base de données numériques [40]. Pour cela, nous nous fondons sur l'algorithme ECM [73], qui partitionne un cadre de discernement et détermine l'appartenance d'un vecteur aux différentes hypothèses par une fonction de masse. ECM repose sur le même principe que celui proposé dans l'algorithme FCM [13]. Mais dans le cas de l'algorithme ECM, la fonction objective à minimiser est :

$$J_{ECM}(M,V) \triangleq \sum_{i=1}^{d} \sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} c_j^{\alpha} m_{ij}^{\beta} dist_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta} \tag{4.11}$$

avec pour contrainte :

$$\sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} m_{ij} + m_{i\emptyset} = 1 \quad \forall i = 1, d \tag{4.12}$$

où $m_{i\emptyset}$ et m_{ij} représente respectivement $m_i(\emptyset)$ et $m_i(A_j)$. M est la partition crédale $M = (m_1, \ldots, m_d)$ et V contient les centres des partitions. c_j^{α} est un coefficient de poids et $dist_{ij}$ est la distance Euclidienne. Dans notre expérimentation, les paramètres α, β et δ ont été fixés à 1, 2 et 10.
Dans notre approche de construction des bases évidentielles de test, nous nous sommes fondés uniquement sur quelques benchmark (dataset) de l'UCI [40]. Les bases de données étudiées sont détaillées dans le tableau 4.3. Pour chaque dataset, nous avons appliqué l'algorithme ECM et le nombre de partitions a été identifié par la minimisation de la fonction objective J_{ECM}.

Data set	**#Instances**	**#attributs**	**#éléments focaux**
Iris	150	4	32
Vertebral Column	310	6	64
Diabetes	767	9	144
Abalone	4177	9	40

TAB. 4.3: Les caractéristiques des datasets utilisés

Support	Iris		Diabete		Vertebral Column		Abalone	
	EDMA-Pr	EDMA-Bel	EDMA-Pr	EDMA-Bel	EDMA-Pr	EDMA-Bel	EDMA-Pr	EDMA-Bel
0.9	15	15	319	191	63	63	767	511
0.8	23	15	1503	319	95	63	767	511
0.7	47	15	5055	671	415	63	767	511
0.6	95	31	9074	1407	799	95	1919	511

TAB. 4.4: Résultats comparatifs en terme de nombre de motifs fréquents extraits

4.6.2 Résultats comparatifs

Nous avons comparé les performances du support précis à partir de l'algorithme EDMA-p à la mesure de support crédibiliste de [7, 46]. Le tableau 4.4 récapitule les performances des mesures de support en terme de nombre de motifs extraits. La colonne $EDMA-Pr$ est relative au support précis alors que $EDMA-Bel$ réfère à la définition crédibiliste du support. Dans notre expérimentation, nous avons intégré la ramification du support crédibiliste qui a été intégré dans EDMA à la place du support précis. Dans un effort d'intégration du support crédibiliste dans EDMA, nous avons dû créer une table contenant toutes les valeurs du support crédibiliste des éléments focaux. Cette table, nous l'avons appelée la table BT[4]. La table BT a la même structure que la table Pr.

Le tableau 4.4 récapitule le nombre de motifs fréquents retrouvés à partir de la base évidentielle. Nous pouvons clairement voir que le support précis permet d'extraire plus de motifs évidentiels que le support crédibiliste. Ce résultat est attendu car le support précis explore plus d'éléments focaux que le support crédibiliste. Par ailleurs, le nombre de motifs dépend étroitement de la valeur de *minsup* choisie. Plus cette valeur diminue plus le nombre de motifs fréquents extraits augmente.

Nous avons aussi effectué des tests comparatifs entre les algorithmes en terme de temps d'exécution. Pour cela, nous avons étudié l'approche fondée sur le calcul du produit Cartésien (Cart-Bel dans le tableau 4.5). L'approche de calcul du support par produit Cartésien détermine le support en effectuant le produit de toutes les BBA intervenant dans l'itemset. La complexité d'une telle approche est exponentielle en fonction du nombre d'éléments focaux. En effet, pour une base de données évidentielles de k attributs chacun contenant n éléments focaux et d transactions, la complexité arithmétique de ce produit vaut :

$$\mathcal{C} = d \times n^k = O(n^k) \tag{4.13}$$

Le tableau 4.5 est un comparatif de performance entre EDMA-p et un algorithme reposant sur le produit Cartésien (Cart-Bel). Les résultats de EDMA-p sont nettement meilleurs que ceux obtenus avec le produit Cartésien. En revanche, les performances de $EDMA-Bel$ sont meilleures que celles de $EDMA-Pr$. Ceci

[4] en anglais Belief Table.

Support	Iris			Diabete			Vertebral Column			Abalone		
	EDMA-Pr	EDMA-Bel	Cart-Bel ≈	EDMA-Pr	EDMA-Bel	Cart-Bel ≈	EDMA-Pr	EDMA-Bel	Cart-Bel ≈	EDMA-Pr	EDMA-Bel	Cart-Bel ≈
0.9	0.13	0.10	96172	4.72	1.65	6.43E+48	0.74	0.24	3.96E+24	79.71	16.35	9.13E+41
0.8	0.13	0.11	96172	175.94	3.00	6.43E+48	1.11	0.24	3.96E+24	75.77	16.18	9.13E+41
0.7	0.35	0.15	96172	21188	12.69	6.43E+48	15.21	0.24	3.96E+24	77.23	16.24	9.13E+41
0.6	1.01	0.25	96172	12.21E+4	100.56	6.43E+48	116.32	0.33	3.96E+24	337.87	16.21	9.13E+41

TAB. 4.5: Résultats comparatifs en terme de temps d'exécution (secondes)

est justifiable. En effet, le support précis effectue une recherche dans un espace beaucoup plus large que l'approche crédibiliste. D'un autre coté, la génération d'un grand nombre de fréquents entraine un nombre de candidats encore plus important au niveau suivant.

4.7 Conclusion

Dans ce chapitre, nous avons étudié les relations qui peuvent exister entre la fouille de données et la théorie des fonctions de croyance. En effet, nous avons vu l'apport de la théorie des fonctions de croyance pour la modélisation des données imparfaites dans les bases données. Ces bases sont appelées bases de données évidentielles. Nous avons consacré la première partie de ce chapitre à étudier la fouille de données sur ce type de base. La littérature manque de travaux qui traitent cette problématique. Toutefois, nous avons cité un travail, en particulier, qui s'est illustré en proposant une mesure de support qui se base sur la crédibilité. Nous avons aussi mis en évidence les limites majeures de cette mesure. La première est la complexité temporelle importante que nécessite le calcul du support. Dans cette optique, nous avons proposé une nouvelle réécriture du support en nous affranchissant de l'emploi du produit Cartésien. L'expérimentation a démontré que le gain en temps est très important et valide cet avantage théorique. La deuxième limite traitée est le manque de précision du support crédibilite. Nous avons introduit une nouvelle mesure que nous avons appelée mesure de support précis. Contrairement à la mesure précédente, elle étudie plus en profondeur les autres éléments focaux oubliés par la méthode crédibiliste. Nous avons prouvé, expérimentation à l'appui que nous retrouvons beaucoup plus de motifs fréquents évidentiels que la méthode proposée dans la littérature.

Bien que nous arrivions à trouver d'avantage de motifs fréquents, cela nous informe aucunement sur la qualité des éléments retrouvés. Il est important d'étudier si les motifs retrouvés sont représentatifs de la réalité. C'est généralement la confiance à travers la classification associative qui est un indicateur de la qualité des motifs. C'est dans ce cadre que nous nous intéressons, dans le chapitre suivant, à l'élaboration d'une mesure de confiance à base du support précis. Cette mesure sera employée dans un classifieur associatif ce qui permettra de mettre en avant la qualité des me-

sures proposées.

Chapitre 5

Classification Associative dans les bases évidentielles

Sommaire

5.1 Introduction

La classification par les règles d'association (classification associative) a fourni des résultats probants par rapport aux autres méthodes usuelles [15]. En effet, les méthodes de classification supervisée à base de règles d'association sont connues pour être efficaces, performantes et facilement interprétables. Bringmann et al. [16] proposent une vue d'ensemble des avantages offerts par ces approches. Jusqu'à ce jour, plusieurs algorithmes ont été proposés comme CBA [68], CMAR [66], CPAR [111] ou bien même GARC [15].
Avec l'apparition de nouvelles bases de données permettant de gérer les informations imparfaites, une extension des approches de fouille de données classiques a été

lancée [19, 46]. Pour chaque famille de base de données, de nouvelles mesures de support et de confiance ont été élaborées (voir chapitre 2 et 4). Cela a également été le cas pour les bases de données évidentielles pour lesquelles nous avons proposé une nouvelle mesure de support. Cette mesure de support a apporté plus de précision dans l'estimation du support des itemsets évidentiels. Dès lors, nous pouvons nous interroger sur la possibilité de créer, à partir de cette mesure de support, une mesure de confiance en relation avec la mesure précise et cohérente avec l'état de l'art du domaine.
Ainsi, dans ce chapitre, nous proposons une mesure de confiance fondée sur le support précis. Cette mesure de confiance est utilisée pour la conception d'un nouvel algorithme de classification associative sur les bases de données évidentielles. Cet algorithme EDMA-c est proposé sous deux versions selon les règles générées. Ainsi, deux heuristiques différentes sont développées afin de réduire le nombre important de règles de classification en distinguant deux types : les *règles génériques* et les *règles précises*. Dans la deuxième partie de ce chapitre, nous définissons une mesure de fiabilité des règles afin de n'utiliser que les règles pertinentes. Une autre version de l'algorithme EDMA-c est développée prenant en compte la fiabilité des règles évidentielles. Cet algorithme est appelé EDMA-wc. Les approches proposées sont finalement testées sur des bases évidentielles et les résultats de classification obtenus sont comparés.

5.2 État de l'art de la classification associative dans les bases évidentielles

Dans cette section, nous détaillons les approches d'estimation de la confiance d'une règle d'association. Bien que dans le domaine de la fouille de données évidentielles, il n'y ait pas beaucoup de travaux, nous citons les quelques mesures déjà développées. Il est important de noter que la majorité des mesures de confiance proposées dans le domaine de la fouille de données binaires ou bien imparfaites repose sur la probabilité conditionnelle. En effet, la confiance n'est autre que la probabilité d'avoir un itemset sachant l'information de présence d'un autre. D'ailleurs, il est simple de vérifier cela à partir de l'écriture de la confiance dans les bases de données binaires. En effet, la confiance allouée à la règle R issue d'une base de données binaires s'écrit de la manière suivante :

$$Confiance(R) = P(R_c|R_a) = \frac{\sum_{i=1}^{d} P_{T_i}(R_a \cap R_c)}{\sum_{i=1}^{d} P_{T_i}(R_a)} \tag{5.1}$$

où R_a est la prémisse de la règle R, et R_c la conclusion (i.e., $R : R_a \rightarrow R_c$). $P(.)$ est la mesure de support qui peut être aussi vue comme une probabilité de présence d'un itemset dans une base de données (voir la section 2.2.3).

Dans ce qui suit, nous présentons la seule mesure de confiance proposée dans la fouille de données évidentielles. Cette mesure est à attribuer à Hewawasam et al. [46]. Les auteurs ont défini la mesure de confiance comme une crédibilité conditionnelle en cohérence avec leur mesure de support qui repose, elle aussi, sur la fonction de crédibilité. D'une manière générale, la confiance d'une règle d'association évidentielle R appartenant à l'ensemble des règles $\mathcal{R}$, i.e., $R \in \mathcal{R}$ s'écrit de la façon suivante :

$$Confiance(R) = Bel(R_c|R_a) \tag{5.2}$$

où $Bel(.|.)$ est la crédibilité conditionnelle proposée par Dempster. Cette forme de confiance est générale car on retrouve, dans la littérature sur la théorie des fonctions de croyance, plusieurs interprétations de la crédibilité conditionnelle. Dans le chapitre 1, nous avons vu la crédibilité conditionnelle proposé par Dempster [25] qui s'écrit ainsi :

$$Bel(R_c|R_a) = \frac{Bel(R_c \cup \overline{R_a}) - Bel(\overline{R_a})}{1 - Bel(\overline{R_a})}. \tag{5.3}$$

Cette écriture de la crédibilité conditionnelle généralise la probabilité conditionnelle. Elle a souvent donné des résultats contre-intuitifs [114, 81, 30]. Une autre définition de la crédibilité conditionnelle a été proposée par Fagin et al. [38] :

$$Bel(R_c|R_a) = \frac{Bel(R_a \cap R_c)}{Bel(R_a \cap R_c) + Pl(R_a \cap \bar{R_c})}. \tag{5.4}$$

Les deux versions de la crédibilité conditionnelle affichent quelques limites dans des cas très particuliers. L'exemple qui suit illustre les difficultés d'obtention de la confiance des règles d'association évidentielles par les méthodes décrites précédemment.

Transaction	Attribut A	Attribut B
T1	$m_{11}(A_1) = 0.7$	$m_{21}(B_1) = 0.4$
	$m_{11}(\Theta_A) = 0.3$	$m_{21}(B_2) = 0.2$
		$m_{21}(\Theta_B) = 0.4$
T2	$m_{12}(A_2) = 0.3$	$m_{22}(B_1) = 1$
	$m_{12}(\Theta_A) = 0.7$	

TAB. 5.1: Exemple de base de données évidentielles $\mathcal{EDB}$

Exemple 15 *A travers cet exemple, nous mettons en lumière l'utilisation inadéquate de la crédibilité conditionnelle définie dans [25]. Considérons la transaction 1 de la base de données du tableau 5.1. La confiance de la règle d'association évidentielle $A_2 \rightarrow B_1$ (i.e., $Bel(B_1|A_2)$) par la crédibilité conditionnelle de l'équation 5.3 donne le résultat suivant :*

$$Bel(B_1|A_2) = \frac{Bel(B_1 \cup \overline{A_2}) - Bel(\overline{A_2})}{1 - Bel(\overline{A_2})} = \frac{Bel(B_1)}{1} = 0.4$$

La crédibilité d'avoir B_1 sachant que A_2 est vraie et égale à $Bel(B_1)$. Ceci est dû à l'indépendance des hypothèses A_2 et B_1. Ce résultat signifie que l'occurrence de B_1 ne dépend aucunement de A_2. Il en va de même avec la crédibilité conditionnelle de Fagin et al. [38] définie par l'équation 5.4 à partir de laquelle, nous obtenons :

$$Bel(B_1|A_2) = \frac{Bel(A_2 \cap B_1)}{Bel(A_2 \cap B_1) + Pl(A_2 \cap \bar{B_1})} = \frac{Bel(\emptyset)}{Bel(\emptyset) + Pl(\emptyset)} = 0.$$

La difficulté à déterminer avec exactitude la valeur de la confiance d'une règle d'association évidentielle entrave la mise au point d'un algorithme de classification associative. Dans la section suivante, nous proposons une nouvelle mesure de confiance qui surmonte les limites affichées. De plus, cette mesure est cohérente avec celle définie dans le cadre de la fouille de données binaires.

5.3 Confiance probabilistique des règles associatives

Dans le chapitre 4, nous avons présenté la mesure de support précis permettant de retrouver les motifs évidentiels fréquents. La mesure introduite repose sur une étude de tout l'espace du cadre du discernement d'une fonction de masse m. De ce fait, cette mesure se distingue de la mesure du support crédibiliste. Dans ce qui suit, nous présentons notre approche de mesure de confiance qui repose sur la notion de support précis défini précédemment. Pour une règle $R : R_a \rightarrow R_c$, la confiance [86] s'écrit sous cette forme :

$$Confiance(R) = \frac{\sum_{j=1}^{d} Pr_{T_j}(R_a) \times Pr_{T_j}(R_c)}{\sum_{j=1}^{d} Pr_{T_j}(R_a)} \tag{5.5}$$

où d est le nombre de transactions dans la base évidentielle. La confiance peut être écrite avec le support :

$$Confiance(R) = \frac{Support_{\mathcal{EDB}}(R_a \times R_c)}{Support_{\mathcal{EDB}}(R_a)} \tag{5.6}$$

Cette écriture de la confiance soutient la mesure initiale proposée dans le cadre binaire [3]. Cette formulation a aussi l'avantage de tirer profit de la mesure de support précis. Cela veut dire que la mesure de confiance exploite tout le cadre de discernement pour l'estimation de la confiance d'une règle d'association.

Exemple 16 *Soit la base de données évidentielles du tableau 5.1. La confiance de la règle d'association évidentielle* $R_1 : A_1 \rightarrow B_1$ *est calculée ainsi :*

$$Confiance(R_1) = \frac{Pr_{T_1}(A_1) \times Pr_{T_1}(B_1) + Pr_{T_2}(A_1) \times Pr_{T_2}(B_1)}{Pr_{T_1}(A_1) + Pr_{T_2}(A_1)} = 0.75$$

La confiance des règles d'association est l'information la plus intéressante que l'on peut extraire à partir d'une base de données. Afin de tester sa qualité, plusieurs méthodes existent. Dans le cadre de nos travaux, nous avons privilégié de tester la pertinence des règles générées à travers la classification associative.

5.4 Classification associative par les règles précises et génériques

L'une des limites de l'algorithme Apriori est le nombre important de règles d'association qu'il génère. En effet, pour un itemset I de taille k, $2^k - 2$ règles peuvent être déduites. Un ensemble d'opérations est à prévoir afin de réduire le nombre de règles pour ne retenir que les plus intéressantes. Dans cette section, nous proposons un ensemble de méthodes afin de réduire le nombre de règles. Les règles retenues formeront l'ensemble des règles du classifieur associatif.

Une première idée, pour élaguer des règles, serait de ne retenir que celles ayant une terminaison contenant une classe. Formellement, pour une règle $\prod_{i \in I} X_i \rightarrow \prod_{j \in J} Y_j$, nous ne gardons que celles ayant une conclusion ne contenant qu'une classe c'est à dire $Y_j \in \Theta_C$ où Θ_C est le cadre de discernement.

Exemple 17 *Soit l'ensemble des règles d'association* $S = \{A_1 \rightarrow C_1; A_1, B_2 \rightarrow C_1; A_1 \rightarrow B_1\}$ *et l'ensemble des classes tel que* $\Theta_C = \{C_1, C_2\}$*. Après réduction, l'ensemble* S *devient* $S' = \{A_1 \rightarrow C_1; A_1, B_2 \rightarrow C_1\}$*.*

5.4.1 Règles d'association précises et génériques

Afin de poursuivre la réduction du nombre de règles d'association évidentielles, nous proposons deux heuristiques. La première consiste à ne retenir que les règles avec une prémisse minimale. Ce type de règle est appelé *règle générique*. La deuxième heuristique ne retient que les règles avec une prémisse maximale et elles sont appelées les *règles précises*.

L'idée sous-jacente des règles génériques est de minimiser le nombre de règles générées en un ensemble ne contenant que quelques exemples pouvant couvrir l'ensemble des cas de classification. Cette méthode s'inspire de la méthode d'élagage des règles redondantes dans les bases de données binaires (voir chapitre 3). En effet, une règle R_1 est considérée comme redondante si elle n'apporte aucune information supplémentaire par rapport à une autre règle R_2. Dans ce cas de figure, R_1 est dite

une règle redondante par rapport à R_2. De même dans le cadre évidentielle, nous formons à partir de l'ensemble de règles d'association évidentielles $\mathcal{R}$ l'ensemble $\mathcal{RC}$ tel que :

$$\mathcal{RC} = \{R : R_a \rightarrow R_c \in \mathcal{R} |\ \nexists R' : R'_a \rightarrow R_c, R'_a \subset R_a\}$$

Exemple 18 *Soit l'ensemble des règles d'association* $S = \{A_1 \rightarrow C_1; A_1, B_2 \rightarrow C_1; A_1 \rightarrow B_1\}$. *Le filtrage de* S *par les règles génériques donne* $S' = \{A_1 \rightarrow C_1; A_1 \rightarrow B_1\}$.

Contrairement aux règles génériques, les règles précises requièrent une prémisse maximale. Avec une telle construction, les règles sont précises et ne s'appliquent que lorsque toute la partie prémisse est vérifiée. Dès lors, contrairement aux règles génériques, les règles précises sont très spécifiques et leur application est restreinte. En revanche, dans le cas où leur application est possible, la confiance de la règle utilisée est pertinente. Ces règles sont obtenues de la façon suivante :

$$\mathcal{RC} = \{R : R_a \rightarrow R_c \in \mathcal{R} |\ \nexists R' : R'_a \rightarrow R_c, R_a \subset R'\}$$

Exemple 19 *On considère l'ensemble des règles* $S = \{A_1 \rightarrow C_1; A_1, B_2 \rightarrow C_1; A_1 \rightarrow B_1\}$. *L'ensemble* S' *des règles précises est* $S' = \{A_1, B_2 \rightarrow C_1\}$.

5.4.2 Algorithme Evidentiel de Data Mining (EDMA-c) : classification associative

Afin d'évaluer la qualité de nos règles d'association évidentielles, nous traitons le problème de la classification. Soit une instance X à classer telle que :

$$X = \{m_i | m_i \in X, x_i^j \in \Theta_i\} \tag{5.7}$$

où x_i^j est un élément focal de la fonction de masse m_i. Chaque règle d'association retenue représente une source d'information sur l'appartenance de X aux différentes classes considérées. Les règles pouvant informer sur la classe de X sont celles qui ont une intersection non nulle avec X. Elles sont retrouvées de la manière suivante :

$$\mathcal{RI} = \{R \in \mathcal{RC}, \exists x_i^j \in \Theta_i,\ x_i^j \in R_a\}. \tag{5.8}$$

Chaque règle $R \in \mathcal{RI}$ est une information quant à l'appartenance de X. Il est important de souligner que la cardinalité de l'ensemble est variable. En effet, si $|\mathcal{RI}| = 0$ alors l'instance X ne peut être classée. Dans le cas où $|\mathcal{RI}| > 1$, plusieurs règles doivent être prises en compte pour décider de l'appartenance de X. L'une des solutions serait de concevoir un système de fusion d'information par la théorie des fonctions de croyance. Toutes les règles de $\mathcal{RI}$ seront des sources d'information. Dans ce cas quel type de règles choisir : génériques ou précises ?
Dans ce travail, nous distinguons les deux méthodes de classification. La première fondée sur des règles précises et la seconde élaborée à partir des règles génériques.

Chaque règle $R_l \subset \mathcal{RI}$, $l \in [1 \dots L]$ et $L < |\mathcal{RI}|$ est transformée en une fonction de masse dans le cadre de discernement Θ_C (cadre de discernement de R_cl). La construction de la fonction de masse se fait de la manière suivante :

$$\begin{cases} m_{R_l}^{\Theta_C}(\{R_{cl}\}) = Confiance(R_l) \\ m_{R_l}^{\Theta_C}(\Theta_C) = 1 - Confiance(R_l) \end{cases} \tag{5.9}$$

où R_{cl} est la partie conclusion de la règle R_l.

Les L fonctions de masses construites sont alors fusionnées à l'aide de l'opérateur de combinaison de Dempster [25] :

$$m_{\oplus} = \oplus_{l=1}^{L} m_{R_l}^{\Theta_C}. \tag{5.10}$$

Dans ce qui suit, nous présentons l'algorithme de classification associative évidentielle EDMA-c[1] [88]. Il est détaillé dans l'algorithme 6. L'algorithme est construit autour de la fonction $Construct\_Rule(x, \Theta_C)$ qui génère les règles d'association évidentielles en ne retenant que celles de classification. L'algorithme 6 implémente la fonction $Find\_Confidence(R, Pr\_Table)$ qui est détaillée dans l'algorithme 5. Finalement, la fonction $Redundancy(\mathcal{R}, R)$ construit l'ensemble de toutes les règles de classification $\mathcal{RC}$ non redondantes et ayant une confiance supérieure ou égale à $minconf$.

5.5 Règles associatives évidentielles affaiblies

Dans l'équation 5.10, toutes les règles d'association sont combinées avec le même poids. Cela veut dire que le classifieur ne fait pas de distinction entre les règles. Elles sont toutes équivalentes en terme d'informations apportées. En effet, chaque règle ayant une intersection non nulle avec X est fusionnée. Malheureusement, certaines règles sont moins informatives que d'autres concernant X. Il est alors intéressant de faire la distinction entre ces règles en leur associant une mesure de fiabilité. Dans ce qui suit, nous présentons l'intérêt de la pondération de ces règles ainsi que les différents critères qui nous permettront de définir les valeurs des poids.

5.5.1 Motivation

Le seul critère de sélection des règles à fusionner repose sur l'intersection avec l'instance à classer. Ceci ne garantit pas la sélection des règles pertinentes. En effet, dans la méthode précédente, une règle R_1 ayant un seul élément en commun avec X (i.e., $|R_1 \cap X| = 1$) est fusionnée avec le même poids qu'une règle R_2 ayant plusieurs éléments en commun avec l'instance (i.e., $|R_1 \cap X| > 1$). Ajoutons à cela, que le résultat de la combinaison peut être détérioré avec l'intégration de règles qui pour certaines apportent soient peu d'informations soient des informations contradictoires sur l'appartenance de X. Les propriétés P_1 et P_2, présentées ci-dessous, définissent

[1] en anglais Evidential Data Mining Algorithm- classification version

Algorithm 5 Algorithme de la fonction d'estimation de la confiance

```
1: function FIND_CONFIDENCE(R, Pr)
2:     numer ← 0
3:     denom ← 0
4:     for j=1 to d do
5:         num ← 1
6:         den ← 1
7:         for all i ∈ Pr(j).focal_element do
8:             if Pr(j).focal_element ∈ R.premise then
9:                 num ← num × Pr(j).val
10:                den ← den × Pr(j).val
11:            else
12:                if Pr(j).focal_element ∈ R.conclusion then
13:                end if
14:            end if
15:        end for
16:        numer ← numer + num
17:        denom ← denom + den
18:    end for
19:    return numer/denom
20: end function
```

deux conditions qui doivent être vérifiées afin de distinguer les règles d'association pertinentes.

Propriété 6
- P_1 : *Un poids plus important doit être accordé à la règle la plus précise par rapport à* X.
- P_2 : *La mesure de croyance accordée aux éléments focaux de l'instance* X *est un critère de pondération des règles d'association.*

Exemple 20 *Dans cet exemple, nous détaillons les propriétés* P_1 *et* P_2 *par des exemples illustratifs. Soit l'instance* X *à classer du tableau 5.2.*

Soit l'ensemble des règles d'association de classification suivante $\mathcal{RI} = \{R_1 : A_1, B_1 \rightarrow C_1; R_2 : \Theta_A, B_1 \rightarrow C_1; R_3 : A_1 \rightarrow C_1; R_4 : B_2 \rightarrow C_2\}$. *Nous avons :*

- *D'après la propriété* P_1, R_1 *devrait avoir un poids plus important que* R_2 *car* $R_{1a} \subset R_{2a}$.
- *D'après la propriété* P_2, R_4 *n'est pas une règle pertinente car la croyance de* B_2 *(i.e.,* $m(B_2)$*) est faible.*

Algorithm 6 Algorithme EDMA-c

Require: $Pr\_Table, minconf, \mathcal{FI}, \Theta_C$

Ensure: $\mathcal{RC}$

1: **for all** $x \in \mathcal{FI}$ **do**
2: $R \leftarrow Construct\_Rule(x, \Theta_C)$
3: **if** $R \neq \emptyset$ **then**
4: $Conf \leftarrow Find\_Confidence(R, Pr\_Table)$
5: **if** $Conf > minconf$ **then**
6: $\mathcal{RC} \leftarrow Redundancy(\mathcal{RC}, R)$
7: **end if**
8: **end if**
9: **end for**
10: **function** CONSTRUCT_RULE(X , Θ_C)
11: **for all** $x \in X$ **do**
12: **if** $x \notin \Theta_C$ **then**
13: $prem \leftarrow prem + \{x\}$
14: **else**
15: $concl \leftarrow concl + \{x\}$
16: **end if**
17: **end for**
18: $R.premise \leftarrow prem$
19: $R.conclusion \leftarrow concl$
20: **return** R
21: **end function**
22: **function** REDUNDANCY($\mathcal{RC}$,R)
23: **for all** $rule \in \mathcal{RC}$ **do**
24: **if** $R.premise \subset rule.premise$ & $R.conclusion = rule.conclusion$ **then**
25: $\mathcal{RC} \leftarrow \mathcal{RC} \backslash rule$
26: $\mathcal{RC} \leftarrow \mathcal{RC} \cup R$
27: **end if**
28: **end for**
29: **return** $\mathcal{RC}$
30: **end function**

Attribut A	Attribut B
$m(A_1) = 0.6$	$m(B_1) = 0.5$
$m(A_2) = 0.2$	$m(B_2) = 0.1$
$m(\Theta_A) = 0.2$	$m(\Theta_B) = 0.4$

TAB. 5.2: Instance évidentielle X à classer

5.5.2 Détermination des facteurs de poids par des règles d'association évidentielles

Dans cette partie, nous introduisons une nouvelle approche permettant la pondération des règles d'association évidentielles. Cette méthode prend en considération les deux critères cités dans l'exemple 20. Soit une règle d'association évidentielle $R : R_a \rightarrow R_c$ dont nous essayons de déterminer la pertinence par rapport à une instance X à classer. Pour chaque item constituant la partie prémisse de R i.e., $\{ x_i^j \in R_a | x_i^j \in \Theta_i, i \in n, j \in J\}$, nous calculons la distance le séparant de X. A partir de chaque item (élément focal) $\{x_i^j \in R_a, i \in n, j \in J\}$, nous construisons une fonction de masse catégorique $m_i^c(\{x^j\})$. Cette fonction de masse est alors comparée avec m_i afin d'estimer la distance les séparant. Cette approche permet de distinguer les règles d'association par rapport au critère P_1 (voir la section 5.5.1). Formellement, la distance est calculée comme suit :

$$d_i(m_i^c, m_i) = \sqrt{\frac{1}{2}(m_i^c - m_i)^t . D . (m_i^c - m_i)} \tag{5.11}$$

où :

$$D(A,B) = \begin{cases} 1 \quad \text{if} \quad A = B = \emptyset \\ \frac{|A \cap B|}{|A \cup B|} \quad \text{if} \quad A, B \subseteq 2^\Theta, \end{cases} \tag{5.12}$$

d_i est la distance de Jousselme [53]. La matrice $D(A,B)$ formalise les relations d'inclusions entre les éléments focaux. Ainsi, le poids d'une règle R est déterminé à partir de l'ensemble des distances retrouvées :

$$dist(R) = \frac{\sum\limits_{i \in I} d_i}{I}. \tag{5.13}$$

Ainsi, pour la règle R_l, l'équation 5.9 devient :

$$\begin{cases} {}^{\alpha}m_{R_l}^{\Theta_C}(\{R_{cl}\}) = (1 - dist(R_l)) \times m_{R_l}^{\Theta_C}(\{R_{cl}\}) \\ {}^{\alpha}m_{R_l}^{\Theta_C}(\Theta_C) = (1 - dist(R_l)) \times m_{R_l}^{\Theta_C}(\Theta_C) + dist(R_l). \end{cases} \tag{5.14}$$

Exemple 21 *Soit les mêmes données que celles de l'exemple 20. Dans le tableau 5.3, nous présentons un exemple applicatif sur la détermination des fiabilités des règles*

d'association évidentielles et le résultat de leur fusion.

Règle	Distance $dist()$	Fiabilité (α)	$m_{R_l}^{\Theta_C}$	${}^{\alpha}m_{R_l}^{\Theta_C}$	$m_{\oplus}$	BetP
$R_1 : A_1, B_1 \rightarrow C_1$	0.34	0.66	$m_{R_1}^{\Theta_C}(C_1) = 0.59$ $m_{R_1}^{\Theta_C}(\Theta_C) = 0.41$	${}^{\alpha}m_{R_1}^{\Theta_C}(C_1) = 0.39$ ${}^{\alpha}m_{R_1}^{\Theta_C}(\Theta_C) = 0.61$	$m_{\oplus}(C_1) = 0.59$	
$R_2 : \Theta_A, B_1 \rightarrow C_1$	0.40	0.60	$m_{R_2}^{\Theta_C}(C_1) = 0.32$ $m_{R_2}^{\Theta_C}(\Theta_C) = 0.68$	${}^{\alpha}m_{R_2}^{\Theta_C}(C_1) = 0.19$ ${}^{\alpha}m_{R_2}^{\Theta_C}(\Theta_C) = 0.81$	$m_{\oplus}(C_2) = 0.06$	$BetP(C_1) = 0.77$
$R_3 : A_1 \rightarrow C_1$	0.31	0.69	$m_{R_3}^{\Theta_C}(C_1) = 0.32$ $m_{R_3}^{\Theta_C}(\Theta_C) = 0.68$	${}^{\alpha}m_{R_3}^{\Theta_C}(C_1) = 0.22$ ${}^{\alpha}m_{R_3}^{\Theta_C}(\Theta_C) = 0.78$	$m_{\oplus}(\Theta_C) = 0.35$	$BetP(C_2) = 0.23$
$R_4 : B_2 \rightarrow C_2$	0.73	0.27	$m_{R_4}^{\Theta_C}(C_2) = 0.66$ $m_{R_4}^{\Theta_C}(\Theta_C) = 0.34$	${}^{\alpha}m_{R_4}^{\Theta_C}(C_2) = 0.18$ ${}^{\alpha}m_{R_4}^{\Theta_C}(\Theta_C) = 0.82$		

TAB. 5.3: Exemple de la détermination des facteurs de poids pour des règles d'association évidentielles

Le tableau 5.3 est un exemple d'un calcul numérique de la fusion de règles d'association évidentielles avec affaiblissement. La colonne Fiabilité illustre la fiabilité de chaque règle d'association calculée à partir de la mesure distance. Les résultats numériques retrouvés sont conformes à nos attentes. Par exemple, la règle R_1 affiche une meilleure fiabilité que R_2. Ceci supporte la propriété P_1. D'un autre coté, R_3 est plus fiable que R_4 selon la propriété P_2. Ensuite, les règles sont modélisées en fonctions de croyance selon le procédé de l'équation 5.14. La décision prise après fusion par la maximum de la probabilité pignistique donne la classe C_1 comme la classe la plus probable ce qui est réellement le cas (la majorité des règles soutient la classe C_1).

5.5.3 Algorithme Evidentiel de Data Mining (EDMA-wc) : classification associative pondérée

Dans cette partie, nous détaillons une variante de l'algorithme EDMA pour la classification associative que nous appelons EDMA-wc[2]. L'algorithme 8 s'inspire de l'algorithme 6 celui-ci se distingue par un module supplémentaire permettant d'estimer la fiabilité des règles d'association. Cet algorithme implémente des fonctions de EDMA-p telle que $Frequent\_itemset()$ qui filtre tous les itemsets d'une taille donnée pour ne retenir que les fréquents. Une fois que tous les fréquents sont générés, EDMA-wc extrait les règles de classification intéressantes au travers de leurs confiances. La confiance est calculée par le biais de la fonction $Find\_Confidence()$.

[2]en anglais Evidential Data Mining Algorithm- weigthed classification version.

Les règles sont filtrées en retirant les règles redondantes décrites précédemment dans l'algorithme 6. Pour chaque règle retenue R, nous définissons la fonction de masse correspondante à l'aide de la fonction $construct\_BBA()$. Ensuite, la fonction de masse trouvée est utilisée afin de calculer le facteur de poids de la règle (facteur de fiabilité de la règle). Ce calcul est effectué par la fonction $compute\_reliability()$ qui implémente la distance de Jousselme [53]. Cette fonction est détaillée dans l'algorithme 7.

Algorithm 7 Algorithme de la fonction d'estimation de la fiabilité d'une règle d'association

1: **function** COMPUTE_RELIABILITY(R , X)
2: $d \leftarrow 0$
3: **for all** $r \in R_a$ **do**
4: $d \leftarrow d + Jousselme\_distance(m_r, X^i)$
5: **end for**
6: $\alpha_I \leftarrow \frac{d}{sizeof(R_a)}$
7: **return** α_I
8: **end function**

5.6 Expérimentation

Dans ce qui suit, nous étudions les performances de classification des règles évidentielles précises et génériques. Le tableau 5.4 présente un comparatif des performances entre les règles précises et celles affaiblies. Les résultats prouvent l'importance de l'affaiblissement pour les règles précises puisque nous avons amélioré la classification des bases de données Iris_EDB et Vertibral_column_EDB. Nous avons également obtenu un taux de bonne classification de 100%, avec l'approche intégrant l'affaiblissement pour la base de données Wine_EDB. En revanche, pour la base de données Diabete_EDB, les résultats de la classification sont proches.

Le même comparatif a été mené sur les règles génériques. Le tableau 5.5 montre l'efficacité de l'affaiblissement pour ce type de règles. En effet, nous avons considérablement amélioré les résultats de classification pour les bases de données Wine_EDB et Iris_EDB. Aussi, nous avons maintenu les mêmes performances en terme de classification pour les bases de données Vertibral_column_EDB et Diabete_EDB. L'amélioration des résultats de classification par les règles génériques affaiblies est d'autant plus intéressante au regard de la taille des prémisses. En effet, pour la majorité des règles génériques, la prémisse ne dépasse pas deux éléments (itemset de taille 2). Dans ce cas de figure, seule la propriété P_2 de la sous-section 5.5.1 est prise en compte.

Les règles évidentielles précises (affaiblies et non affaiblies) affichent de meilleurs résultats de classification que les règles génériques. En effet, plus la prémisse est

large plus la règle est fiable quant à l'appartenance de l'instance à classer. Les performances plus faibles obtenues avec les règles génériques s'expliquent par la quantité des règles fusionnées. Dans le cas de classification avec les règles ayant des prémisses minimales, plusieurs règles sont sécantes avec l'instance à classer. Dans ce cas de figure, plus nous avons de règles pour classer plus le risque de contradiction subsiste. Ce n'est pas le cas des règles précises, où dans la majorité des situations, seules quelques règles sont fusionnées et celles-ci sont souvent cohérentes au niveau de leurs parties conclusion. La Figure 5.1 présente le nombre élevé de règles fusionnées qui dépend étroitement de la valeur de *minsup*.

Base de données	Iris_EDB	Vertebral Column_EDB	Diabetes_EDB	Wine_EDB
Règles précises	80.67%	88.38%	83.20%	100%
Règles précises affaiblies	82.00%	89.03%	82.81%	100%

TAB. 5.4: Résultats comparatifs de la classification associative avec les règles précises évidentielles.

Base de données	Iris_EDB	Vertebral Column_EDB	Diabetes_EDB	Wine_EDB
Règles génériques	78.67%	67.74%	65.10%	51.68%
Règles génériques affaiblies	80.00%	67.74%	65.10%	76.40%

TAB. 5.5: Résultats comparatifs de la classification associative avec les règles génériques évidentielles.

5.7 Conclusion

Dans ce chapitre, nous avons étudié la classification associative dans la fouille de données évidentielles. Nous avons proposé une nouvelle mesure de confiance qui mesure la pertinence des règles d'association évidentielles. La mesure proposée est cohérente avec les mesures proposées précédemment et notamment dans le cadre binaire. Par ailleurs, elle ne présente pas les limites de celles déjà existantes dans l'état de l'art. Afin d'estimer la qualité de la mesure de support, nous avons introduit un nouveau classifieur évidentiel appelé EDMA-c. Ce classifieur repose sur les règles d'association évidentielles de confiance et sur leur fusion afin de déterminer la classe d'une instance évidentielle. Malheureusement, nous avons remarqué que de nombreuses règles de confiance peuvent être générées. Pour éviter cette situation, nous proposons de distinguer deux types de règles d'association évidentielles. La première famille de règles représente les règles génériques qui ont des prémisses minimales. La deuxième, les règles précises, ont une prémisse maximale.
Notre classifieur repose sur la fusion de règles d'association. Dans la majorité des cas leur nombre peut être très important et ainsi remettre en question la qualité de la fusion. Afin d'optimiser les résultats de la classification, nous introduisons une

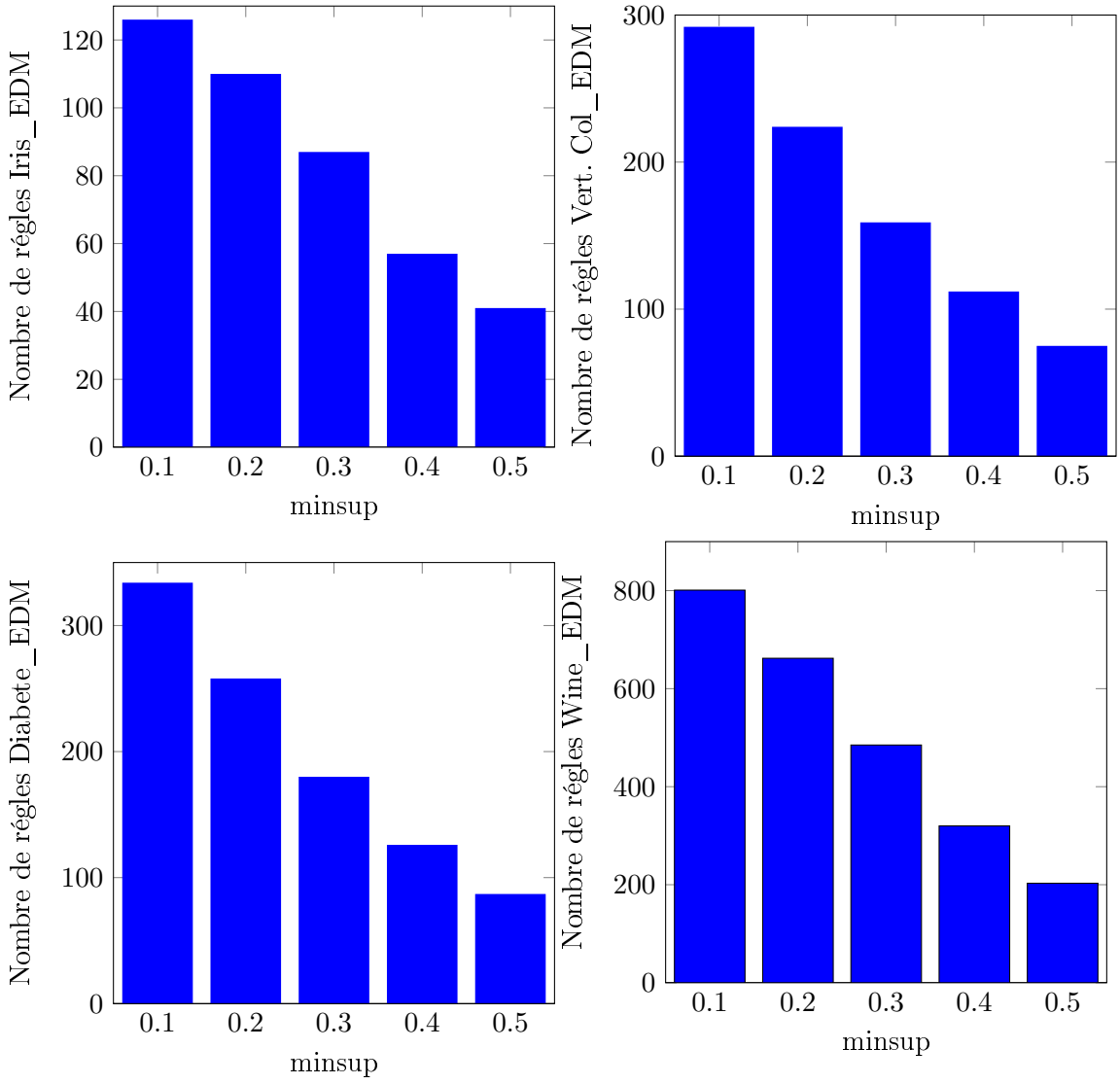

FIG. 5.1: Evolution du nombre de règles génériques générées par rapport au *minsup*

mesure de fiabilité qui estime la pertinence d'une règle d'association évidentielle par rapport à l'instance à classer. Cet algorithme est appelé EDMA-wc. L'expérimentation a validé l'apport de cette mesure de fiabilité et a donné des résultats très encourageants.

Algorithm 8 Algorithme EDMA-wc

Require: $\mathcal{EDB}, minsup, Pr\_Table, minconf, Size\_\mathcal{EDB}, \Theta_C, X$

Ensure: $\mathcal{R}, \mathcal{EIFF}, m$

$\mathcal{EIFF} \leftarrow \emptyset$
$size \leftarrow 1$
$candidate \leftarrow candidate\_apriori\_gen(\mathcal{EDB}, size)$
While $(candidate \neq \emptyset)$
$freq \leftarrow Frequent\_itemset\ (candidate, minsup, Pr\_Table, Size\_\mathcal{EDB})$
$size \leftarrow size + 1$
$\mathcal{EIFF} \leftarrow \mathcal{EIFF} \cup freq$
$candidate \leftarrow candidate\_apriori\_gen(\mathcal{EDB}, size, freq)$
End While
for all $x \in \mathcal{EIFF}$ **do**
    $R \leftarrow Construct\_Rule(x, \Theta_C)$
    if $R \neq \emptyset$ **then**
        $Confidence \leftarrow Find\_Confidence(R, Pr\_Table)$
        $\mathcal{R} \leftarrow Redundancy(\mathcal{R}, R, confidence)$
    end if
end for
for all $R \in \mathcal{R}$ **do**
    if $x \cap R \neq \emptyset$ **then**
        $RI \leftarrow RI \cup R$
    end if
end for
for all $R \in RI$ **do**
    $BBA \leftarrow construct\_BBA(R)$
    $\alpha_I \leftarrow compute\_reliability(R, X)$
    $BBA^{\alpha} \leftarrow discounting(BBA, \alpha_I)$
    $m \leftarrow m \oplus BBA^{\alpha}$
end for
function CONSTRUCT_BBA(R)
    $m_R(R.conclusion) \leftarrow d \times R.confidence$
    $m_R.(\Theta) \leftarrow 1 - d \times R.confidence$
    $BBA \leftarrow m_R$
    return BBA
end function

Conclusion et perspectives

Dans ce travail de thèse, nous nous sommes intéressés à deux domaines différents. Le premier est la théorie des fonctions de croyance, qui permet de modéliser et de fusionner des informations imparfaites. Le second, la fouille de données, est une discipline qui vise à extraire des connaissances à partir d'un gros volume de données. Nous avons étudié les possibles interactions, dans les deux sens, entre ces deux domaines. La première partie des contributions de ce manuscrit a été dédiée à l'étude de l'impact de la fouille de données dans la résolution de problèmes propres à la théorie des fonctions de croyance. Dans un deuxième volet, nous nous sommes intéressés à la contribution de la théorie des fonctions de croyance dans la représentation des connaissances imparfaites au sein de la fouille de données.

Synthèse des travaux entrepris

Après avoir présenté les outils de la théorie des fonctions de croyance dans le cadre du Modèle des Croyances Transférables, nous avons introduit les concepts fondamentaux de fouille de données en se fondant sur les connexions de Galois. Ainsi, en plus des outils classiques, tels que le support et la confiance, nous avons exposé les opérateurs de fermeture qui sont à l'origine des approches de représentation condensée des motifs et des règles d'association. Suite à l'introduction de ces principes, nous avons proposé une présentation et un comparatif entre les différentes bases de données imparfaites.
Notre première contribution se matérialise par l'apport de la fouille de données dans la théorie des fonctions de croyance. En effet, nous traitons le cas d'une fonction de masse conflictuelle issue de la combinaison conjonctive de plusieurs sources contradictoires. Nous supposons aussi, la présence d'une autre source d'information contenue dans une base de données. Cette dernière information révèle des connaissances différentes mais complémentaires aux sources fusionnées. Cette base de données est étudiée par les outils de fouille de données afin d'extraire des règles d'association génériques. Ces règles présentent la particularité d'être autant informatives que les bases de règles usuelles mais leur nombre est plus restreint. Ces règles d'association sont utilisées afin de guider la gestion de conflit en transférant les croyances vers les hypothèses les plus probables. Cette approche de gestion de conflit par les règles d'association, appelé ACM, est appliquée sur un problème de classification de couronnes d'arbre. Les résultats fournis, comparés avec les approches usuelles, ont démontré l'intérêt d'utiliser les règles d'association pour la gestion de conflit.
Notre seconde contribution concerne l'apport de la théorie des fonctions de croyance dans le domaine de la fouille de données. La théorie des fonctions de croyance est utilisée pour modéliser les informations imprécises et incertaines dans les bases de données. La base de données résultante, qui représente des opinions par des fonctions de masse, est appelée base de données évidentielles. Nous avons souligné les limites

de la mesure de support existante et nous avons introduit des améliorations. La première amélioration consiste en une nouvelle réécriture de cette mesure du support qui réduit considérablement le temps d'exécution déjà exponentiel. Dans un deuxième temps, nous avons proposé une alternative à cette mesure de support que nous avons appelé *support précis*. Cette appellation vient de la spécificité de cette mesure qui explore d'avantage l'espace des éléments focaux et les intègre dans le calcul. Cette mesure a prouvé son efficacité tant en temps d'exécution qu'en nombre de motifs fréquents extraits.
Par la suite, nous avons conçu un classifieur associatif évidentiel. Pour cela, nous avons introduit une mesure de confiance permettant d'estimer la pertinence d'une règle d'association évidentielle. Afin d'éviter le recours à une approche classique de génération de règles d'association, qui produit une base importante en nombre et souffre de redondance, nous avons proposé trois types de règles :

- les règles de classification : les règles qui ont une classe dans la partie conclusion.
- les règles génériques : les règles ayant une prémisse minimale.
- les règles précises : qui correspondent aux règles avec une prémisse maximale.

Nous avons construit deux variantes de l'algorithme de classification associative EDMA-c. La première utilise les règles précises de classification et la seconde les règles génériques de classification. Les tests ont été effectués sur des bases évidentielles construites à partir de datasets.
Nous avons aussi mis en avant l'intérêt d'une mesure permettant de distinguer les différentes règles utilisées pour la classification. C'est dans ce sens que nous avons élaboré une mesure d'estimation de la fiabilité des règles d'association qui permet de différencier les règles les plus pertinentes des autres. L'algorithme EDMA-wc, implémentant cette approche, montre de meilleures performances comparativement à celui qui en est démuni.

Perspectives

Deux types de perspectives selon leur faisabilité dans le temps sont différenciés : des perspectives à court terme et d'autres à long terme. Les perspectives à court terme nécessite une réflexion immédiate sur la résolution de la problématique. Celles à long terme sont des perspectives qui peuvent être étudiées sur la durée.

Bien que les résultats obtenus soient encourageants, un certain nombre de limites sont à citer. La plus préoccupante concerne le temps d'exécution. En effet, une représentation plus juste de l'imperfection des données, réalisée à l'aide des fonctions de croyance, nécessite, en contre partie, un temps plus important pour le calcul des mesures de support et de confiance. Cette limite a été constatée dans le chapitre 4 et 5 où les performances d'exécution ne donnent pas entière satisfaction. Il serait judicieux, à court terme, d'apporter des améliorations à toutes les variantes de l'algorithme EDMA afin qu'elles soient plus en adéquation avec le domaine dans lequel

elles évoluent en terme de temps de réponse. Une autre possibilité d'amélioration du temps d'exécution de nos algorithmes serait d'intégrer des méthodes propres au domaine de la logique comme la symétrie ou bien la théorie des graphes. En effet, il est possible de trouver d'autre ramification dans le calcul en retrouvant des symétries dans la base de données [50].

C'est sans doute la variété des domaines d'applications de la fouille de données qui a fait la notoriété de cette discipline. Cette diversité nous encourage à explorer d'autres champs d'applications pour la fouille de données évidentielles. En effet, plusieurs domaines comme l'optimisation ou la conception de chaînes logistiques peuvent être des applications pour notre approche d'extraction de motifs fréquents. La fouille de données peut être un moteur de prise de décision stratégique pour une entreprise à partir d'une base de données d'experts.

Les origines de la fouille de données sont récentes et il y a encore peu de travaux, dans la littérature, associant cette thématique avec les fonctions de croyance. Il est envisageable d'entrevoir la base de données évidentielles comme une généralisation de ses prédécesseurs en terme de représentation. Pour cela, à long terme, nous envisageons d'étendre les approches binaires qui ont fait de la fouille de données cet outil incontournable. En effet, dans le chapitre 3, nous avons vu le gain octroyé par les bases règles génériques. Il serait intéressant de revoir la faisabilité de ces approches dans le cadre évidentiel.

Annexe A

Théorie des sous-ensembles flous

A.1 Introduction

La théorie des sous-ensembles flous est une généralisation de la théorie des ensembles classiques (i.e., ensembles nets). L'idée de la logique floue, introduite en 1965 par Zadeh [112], est de permettre des gradations dans l'appartenance d'un élément à une classe, c'est-à-dire d'autoriser un élément à appartenir plus ou moins fortement à une classe. Dans ce qui suit, nous allons présenter les notions de base relatives à la théorie des sous-ensembles flous [35].

A.2 Définition

Définition 2 *Un sous-ensemble ordinaire (classique) A inclus dans U est défini par la fonction caractéristique $\mu_A : U \rightarrow \{0, 1\}$. Un élément $x \in U$ est un élément de A si et seulement si : $\mu_A(x) = 1$. Un élément $x_1 \in U$ n'est pas élément de A si et seulement si : $\mu_A(x_1) = 0$*

Définition 3 *Un sous-ensemble flou $\tilde{A}$ inclus dans U est défini par la fonction caractéristique $\mu_{\tilde{A}} : U \rightarrow [0, 1]$ où $\mu_{\tilde{A}}(x)$ désigne le degré avec lequel un élément $x \in U$ est un élément de $\tilde{A}$.*

A.3 Opérations sur les sous-ensembles flous

Définition 4 *Un sous-ensemble flou $\tilde{A} \in U$ est inclus dans un autre sous-ensemble flou $\tilde{B} \in U (\tilde{A} \subseteq \tilde{B})$ si et seulement si tout élément x de U qui appartient à $\tilde{A}$ appartient aussi à $\tilde{B}$ avec un de degré au moins aussi grand, i.e.,*

$$\forall x \in U, \mu_{\tilde{A}} \leq \mu_{\tilde{B}} \tag{A.1}$$

Définition 5 *L'intersection de deux sous-ensembles flous $\tilde{A}$ et $\tilde{B}$ de U est un sous-ensemble flou constitué des éléments de U affectés du plus petit de leurs deux degrés d'appartenance, donnés par $\mu_{\tilde{A}}$ et $\mu_{\tilde{B}}$. C'est le sous-ensemble $\tilde{C} = \tilde{A} \cap \tilde{B}$ de U tel que :*

$$\forall x \in U, \mu_{\tilde{C}}(x) = min\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\} \tag{A.2}$$

Définition 6 *L'union de deux sous-ensembles flous $\tilde{A}$ et $\tilde{B}$ de U est un sous-ensemble flou constitué des éléments de U affectés du plus grand de leurs deux degrés d'appartenance, donnés par $\mu_{\tilde{A}}$ et $\mu_{\tilde{B}}$. C'est le sous-ensemble $\tilde{C} = \tilde{A} \cup \tilde{B}$ de U tel que :*

$$\forall x \in U, \mu_{\tilde{C}}(x) = max\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\} \tag{A.3}$$

Les opérateurs d'union et d'intersection préservent presque toute la structure de la théorie des ensembles classiques. En effet, d'après les définitions données ci-dessus, nous pouvons retrouver les propriétés classiques de l'union et de l'intersection à savoir :

- Associativité et commutativité de $\cup$ et $\cap$,
- Distributivité dans les deux sens de $\cup$ et $\cap$,
- $\tilde{A} \cup \emptyset = \tilde{A}, \tilde{A} \cup U = U$,
- $\tilde{A} \cap U = \tilde{A}, \tilde{A} \cap \emptyset = \emptyset$.

D'autres opérateurs sont envisageables. Ces opérateurs sont définis à l'aide d'une norme triangulaire et d'une conorme triangulaire définies comme suit :

Définition 7 *Une norme triangulaire "t-norme" est une fonction $T : [0,1]\times[0,1] \rightarrow [0,1]$ véfiant, pour tout x et y dans $[0,1]$, les propriétés suivantes :*

- *T est commutative $T(x,y) = T(y,x)$,*
- *T est associative $T(x,T(y,z)) = T(T(x,y),z)$,*
- *T est croissante $T(x,y) \leq T(z,t)$ si $x \leq z$ et $y \leq t$,*
- *$T(x,1) = x$.*

D'une manière générale, l'opérateur d'intersection de deux sous-ensembles flous $\tilde{A}$ et $\tilde{B}$ peut être défini par une t-norme comme suit :

$$\mu_{\tilde{A}\cap_T\tilde{B}}(x) = T(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)) \tag{A.4}$$

Définition 8 *Une conorme triangulaire "t-conorme" est une fonction $\perp : [0,1] \times [0,1] \rightarrow [0,1]$ véfiant, pour tout x et y dans $[0,1]$, les propriétés suivantes :*

- *$\perp$ est commutative $\perp(x,y) = \perp(y,x)$,*
- *$\perp$ est associative $\perp(x,\perp(y,z)) = \perp(\perp(x,y),z)$,*
- *$\perp$ est croissante $\perp(x,y) \leq \perp(z,t)$ si $x \leq z$ et $y \leq t$,*
- *$\perp(x,0b) = x$.*

D'une manière générale, l'opérateur d'union de deux sous-ensembles flous $\tilde{A}$ et $\tilde{B}$ peut être défini par l'intermédiaire d'une t-conorme comme suit :

$$\forall x \in U, \mu_{\tilde{A}\cap_\perp\tilde{B}}(x) = \perp(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)) \tag{A.5}$$

Annexe B

Recueil de publications

B.1 Reliability estimation measure : Generic Discounting Approach

Cet article traite l'estimation de la fiabilité des sources d'information à travers des mesures de distance intrinsèque et extrinsèque. Un synthèse des ces mesures est intégrée dans un classifieur évidentiel reposant sur la distance. Le classifieur est testé sur des problèmes de classification et notamment sur un problème de classification d'images urbaines.

A. Samet, I. Hammami E. Lefevre et S. Ben Yahia : Reliability estimation measure : Generic Discounting Approach. *Annals of Mathematics and Artificial Intelligence (under review)*, 2014.

Ann Math Artif Intell manuscript No.
(will be inserted by the editor)

Reliability estimation measure: Generic Discounting Approach

Ahmed Samet · Imen Hammami · Eric Lefèvre · Sadok Ben Yahia

Received: date / Accepted: date

Abstract In the belief function theory, several measures of uncertainty have been introduced. One of their possible use is unreliable source discounting before the fusion stage. Two different measures of uncertainty exist which are the intrinsic and extrinsic ones. The intrinsic measure makes it possible to assess the source's confusion whereas the extrinsic one measures the contradiction between sources. In this paper, we associate both measures in order to estimate the global reliability of a source. This method, named Generic Discounting Approach (GDA), is proposed in two different versions: Weighted GDA and Exponent GDA. Those reliability measures are integrated into a classifier. The method was tested, against to some pioneer approaches, on several UCI datasets as well as on an urban image classification problem and showed very encouraging results.

Keywords Belief function theory · Discounting · Classification · Conflict management · Source confusion

1 Introduction

The belief functions theory, introduced by Dempster (1967) and formalized by Shafer (1976), has been shown to act as a powerful mathematical background within the information fusion domain as it allows one to express uncertainty and imprecision. In addition to uncertainty handling, this theory allows the extraction of the most likely proposition from multiple sources of provided information. The fusion ability of this formalism is granted by several

A. Samet I. Hammami S. Ben Yahia
University of Tunis El Manar, LIPAH Laboratory, Faculty of Sciences of Tunis, Tunisia
E-mail: firstname.lastname@fst.rnu.tn

A. Samet E. Lefevre
Univ. Lille Nord de France UArtois, EA 3926 LGI2A, F-62400, Béthune, France
E-mail: firstname.lastname@univ-artois.fr

combination rules; the oldest one is the Dempster's rule of combination. However, Zadeh (1994) has highlighted its counterintuitive behavior. As a result, many works have tackled this conflict management issue proposing different types of solutions, that could be split into two main family approaches: *(i)* Conflict management approaches based on discounting the unreliable sources (Klein and Colot, 2010; Martin et al, 2008; Schubert, 2011); *(ii)* Redistribution of the conflict after source's combination (Sentz and Ferson, 2002; Smets, 2007a).

The discounting approach has largely been addressed in the induced literature e.g., Smets (2007a). It relies on the fact that a conflict is generated by the unreliability of at least one source. The unreliable sources are then discounted by a coefficient affecting their consideration during the combination phase. Several works have been carried out in this stream in the sake of finding those discounting factors (Klein and Colot, 2010; Martin et al, 2008; Guo et al, 2006). However, comparatively to the redistribution family, the discounting approaches are less explored by research because of difficulty of measuring source reliability. Nevertheless, some interesting works have been proposed recently based on source's distance measure, providing some interesting results (Jousselme et al, 2001; Daniel, 2010; Smarandache et al, 2011). Indeed, all those works were based on the assumption that the more a source is distant from the other ones (i.e., source in contradiction with other ones), the higher its unreliability.

Several works have highlighted that the conflict resulting from the source combination phase is not necessarily the result of their contradiction (Shafer, 1976; Daniel, 2010; Liu, 2006). Indeed, the lack of informativity or the high confusion[1] of a source could be a sufficient reason for conflict appearance. In literature, several measures were proposed in order to estimate those conflict causes (Jousselme et al, 2001; Smarandache et al, 2011). However, even though the literature witnesses a huge number of conflict management approaches, only a little attention was paid to those considering both conflict managements.

In this paper, we distinguish two possible origins of conflicts and we take them into account. The *intrinsic conflict* caused by the confusion rate of a source to determine certain classes. The second considered conflict origin is the *extrinsic conflict* which indicates to what extent the obtained sources are in contradiction. To eliminate conflict and enhance the right hypothesis during the fusion process, we have to consider those two conflict causes. To achieve this purpose, two new methods are introduced to estimate the sources reliability, namely the Weighted Generic Discounting Approach (GDA-W) and Exponent Generic Discounting Approach (GDA-E). The proposed discounting approaches were integrated into a based belief function distance classifier (Denoeux, 1995). The proposed classifier is experimented within two different contexts. In the first stage, we carried out comparative tests between our classifier and several pioneer approaches on some benchmarks. We also have led

[1] A source is said to be confused if it is hard to pick a decision from its brought information

tests of our classifier performance in an urban image classification problem. The remainder of this paper is organized as follows: Basic concepts of belief functions are recalled in Section 2. Without being exhaustive, various measures of intrinsic and extrinsic conflict, developed in the framework of belief functions, are exposed in Section 3. In Section 4, we introduce both proposed variants of the Generic Discounting Approach (GDA) allowing source's reliability estimation. In Section 5, we present a based belief function classifier that integrated the GDA discounting for source fusion improvement. The proposed classifier is experimented on several benchmark datasets as well as on a high resolution urban image classification problem comparatively to some pioneering works. Finally, we conclude and we sketch issues of future work.

2 Belief functions theory

The Belief functions theory was initiated by the pioneering work of Dempster (1967) on the upper and lower Probabilities. The development of the theory formalism is owed to Shafer (1976). Shafer showed the benefits of belief functions theory in modeling uncertain knowledge. In addition, it allows to fuse information that was obtained through various sources. The belief functions theory is based on several concepts. In this part, we intend to present the main concepts of this theory. For more details, the interested reader may refer to Shafer (1976), Smets and Kennes (1994).

2.1 Frame of discernment

The frame of discernment is the set of possible answers for a treated problem and generally noted Ω. It is composed of N exhaustive and exclusive hypotheses:

$$\Omega = \{H_1, H_2, \ldots, H_N\}$$

The exhaustive assumptions indicate that the solution of the problem is necessarily one of the hypotheses H_i from the frame of discernment. The exclusivity condition support the unicity of the solution $H_i \cap H_j = \emptyset$, $\forall i \neq j$. From the frame of discernment Ω, we deduce the superset 2^Ω containing all the 2^N subsets A of Ω:

$$2^\Omega = \{A, A \subseteq \Omega\} = \{H_1, H_2, \ldots, H_N, H_1 \cup H_2, \ldots, \Omega\}$$

This set constitutes a reference to assess the veracity of any proposal.

2.2 Basic Belief Assignment

The Basic Belief Assignment (BBA) or the basic belief is function m is the mapping from elements of the power set 2^Ω into [0, 1] so that as:

$$m : 2^\Omega \longrightarrow [0, 1]$$

having as constraints:

$$\begin{cases} \sum\limits_{A \subseteq \Omega} m(A) = 1 \\ m(\emptyset) = 0. \end{cases} \tag{1}$$

where $m(A)$ is the confidence strictly assigned to A without is being able to be divided on the hypothesis which composes it. Each subset A of 2^{Ω}, fulfilling $m(A) > 0$, is called *focal element*. Constraining $m(\emptyset) = 0$ is the normalized form of a BBA and this corresponds to a closed-world assumption (Smets, 1990), while allowing $m(\emptyset) > 0$ corresponds to an open world assumption (Smets and Kennes, 1994).
From a BBA, another function can be defined. The plausibility, denoted $Pl(A)$, estimates the maximum potential support that could be given to A, if further evidence becomes available and is defined by:

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{2}$$

2.3 Discounting

Assuming that a source of information has a reliability rate equal to $(1 - \alpha)$ where $(0 \leq \alpha \leq 1)$, such a meta-knowledge can be taken into account using the discounting operation introduced by Shafer (1976), and is defined by:

$$\begin{cases} m^{\alpha}(B) = (1 - \alpha) \times m(B) & \forall B \subseteq \Omega \\ m^{\alpha}(\Omega) = (1 - \alpha) \times m(\Omega) + \alpha \end{cases} \tag{3}$$

A discount rate α equal to 1 means that the source is not reliable and the piece of information that is provided cannot be taken into account. On the contrary, a null discount rate indicates that the source is fully reliable and the piece of information that is provided is entirely acceptable. Thanks to discounting, an unreliable source's BBA is transformed into a function assigning a larger mass to Ω.

2.4 Combination rules

The combination rules are used to combine several belief functions provided by different sources in order to synthesize a single BBA. In this subsection, we survey some pioneer combination rules.

2.4.1 Conjunctive rule

The belief function theory makes it possible to combine some information modelized as BBA. Several operators were defined such as the conjunctive rule. This combination operator assigns the mass to propositions initially confirmed by the sources. For two sources S_1 and S_2 having respectively m_1 and m_2 as BBA, the conjunctive rule is defined by:

$$m_{\textcircled{\cap}} = m_1 \textcircled{\cap} m_2. \tag{4}$$

For an event A, $m_{\textcircled{\cap}}$ can be written as follows:

$$m_{\textcircled{\cap}}(A) = \sum_{B \cap C = A} m_1(B) \times m_2(C) \qquad \forall A \subseteq \Omega. \tag{5}$$

However, the conjunctive combination result does not fulfill the closed-world condition since it gives a conflictual mass.

2.4.2 Dempster's rule of combination

The normalized version of conjunctive rule, proposed by Dempster (1967), integrates a conflict management approach that redistributes the generated conflictual mass. The Dempster's rule of combination, so called orthogonal sum, is defined as follow:

$$m_{\oplus} = m_1 \oplus m_2. \tag{6}$$

For two sources S_1 and S_2, the aggregation of evidence can be written as follows:

$$m_{\oplus}(A) = \frac{1}{1 - m(\emptyset)} \sum_{B \cap C = A} m_1(B) \times m_2(C) = \frac{1}{1 - m(\emptyset)} m_{\textcircled{\cap}}(A) \qquad \forall A \subseteq \Omega, A \neq \emptyset. \tag{7}$$

where $m(\emptyset)$ is defined by:

$$m(\emptyset) = \sum_{B \cap C = \emptyset} m_1(B) \times m_2(C) = m_{\textcircled{\cap}}(\emptyset). \tag{8}$$

$m(\emptyset)$ represents the conflict mass between m_1 and m_2.

2.5 Pignistic probability

In the literature, we often come a cross the notion of pignistic probability. The pignistic probability, denoted $BetP$, was proposed by Smets (2005) within its Transferable Belief Model (TBM) approach. TBM is based on the differentiation between the knowledge representation and decision-making level. In the decision phase, the pignistic transformation consists in distributing equiprobably the mass of a proposition A on its included hypotheses. Formally, the pignistic probability is defined by:

$$BetP(H_n) = \sum_{A \subseteq \Omega} \frac{|H_n \cap A|}{|A|} \times m(A) \qquad \forall H_n \in \Omega. \tag{9}$$

3 Conflict measures

Within the framework of the belief functions theory, the measure $m(\emptyset)$ (equation 8) is often used as the only measure to quantify the conflict. However, it is not satisfactory because it does not consider all conflictual situations (Shafer, 1976). Recently, several works have proposed other measures (Liu, 2006; Klein and Colot, 2011; Martin, 2012; Destercke and Burger, 2012). In this section, several conflict measures (or discordance measures) developed within the framework of belief functions are presented. These measures can be classified into two categories:

- The measures which allow the evaluation of the *extrinsic conflict* (discordance between two bodies of evidence) and will be labeled *extrinsic measures*.
- The measures which allow the estimation of the *intrinsic conflict* (confusion rate of a source) and which will be called *intrinsic measures*.

3.1 Extrinsic conflict

Several measures of extrinsic conflict have been studied in order to model the disagreement between sources. Indeed, if one source opinion disagrees with another, then their fusion will lead to an important conflictual mass (Smets, 2007b). An extension of the Euclidean distance is given by Cuzzolin (2008). In (Tessem, 1993), Tessem introduced a distance measure between the pignistic probabilities that are associated to mass functions. Other distances were studied to define distance between two BBA as the sum of differences of conflicting normalized plausibility masses (Daniel, 2010). Some authors have directly defined distance between different mass functions such as (Jousselme et al, 2001) that has the advantage of taking into account the cardinality of focal elements. This distance complies with the metric axioms and is an appropriate measure of the contradiction between two BBAs. In the remainder, we have chosen it to measure the contradiction between two BBAs and it is formalized as:

$$d(m_1, m_2) = \sqrt{\frac{1}{2}(m_1 - m_2)^t . D . (m_1 - m_2)} \tag{10}$$

where:

$$D(A, B) = \begin{cases} 1 \quad \text{if} \quad A = B = \emptyset \\ \frac{|A \cap B|}{|A \cup B|} \quad \text{if} \quad A, B \subseteq 2^{\Omega}. \end{cases} \tag{11}$$

For further details, the interested reader may refer to (Jousselme and Maupin, 2012).

3.2 Intrinsic measures

The intrinsic conflict quantifies the consistency between the different focal elements inside the BBA. Several measures have been proposed in the literature, that take into account the inclusion relations between the focal elements

present in the BBA. Several intrinsic measures were proposed such as the confusion distance introduced by George and Pal (1996). In (Martin et al, 2008), the authors define the notion of auto-conflict (initially introduced by Yager (1992)) given by the weight assigned to the emptyset generated by the conjunctive combination between m and itself. Nevertheless, the auto-conflict is a kind of confusion measure that depends on the number of combination made. In (Smarandache et al, 2011), the authors introduced a contradiction measure that no longer depends on order that can be written as follows:

$$contr(m) = c \sum_{X \subseteq 2^\Omega} m(X) \cdot D(m, m_X) \tag{12}$$

where $m_X(X) = 1$, $X \in 2^\Omega$ is the categorical[2] BBA, c is normalization constant and D is the Jousselme et al (2001)'s distance (see subsection 3.1). Another approach was presented by Daniel (2010) using the normalized plausibility transformation to measure the internal conflict given by:

$$Pl\_IntC(m) = 1 - \max_{\omega \in \Omega} Pl(\omega). \tag{13}$$

Starting from this definition, there are many BBA without any internal conflicts: all BBA having $X \subseteq \Omega$, $Pl(X) = 1$. There are some examples of BBA having no internal conflict: categorical, simple support[3], consonant BBA[4]. Finally all BBA, whose all focal elements have non-empty intersection, have no internal conflict.

3.3 Conflict measures and discounting

Several works have been carried out to discount the BBAs (Elouedi et al, 2004, 2010). However, these studies are usually based on a learning database. Very few studies used measures of conflict to adapt the belief functions and in this case the used measures are extrinsic measures of conflict.

The first use of extrinsic measure to discount the belief functions was performed by Deng et al (2004). Initially, in this approach, a similarity matrix is build between belief functions. Next, the values of this matrix is used to weight the BBAs. This approach has been modified by choosing a new similarity matrix (Chen et al, 2005).

Martin et al (2008) propose using a function that quantifies the conflict between BBA. This function, called $Conf(.,.)$, is defined as:

$$Conf(i, E) = \frac{1}{M-1} \sum_{k=1; i \neq k}^{M} Conf(i, k) \tag{14}$$

[2] A BBA with only one focal element A is said to be categorical and is denoted $m(A) = 1$.
[3] A BBA is said to be simple if m has no more than two focal sets, Ω included.
[4] A BBA is said consonant if its focal elements are nested.

with M is the number of belief functions produced respectively by M sources called $S_1, \ldots, S_M$ and E is the set of BBAs such that $\{m_k | k = 1, \ldots, M \text{ and } k \neq i\}$. The function $Conf(i, k)$ is obtained using a BBA distance introduced by Jousselme (equation 10):

$$Conf(i, k) = d(m_i, m_k). \tag{15}$$

The value $Conf(i, E)$ quantifies the average conflict between the BBA m_i and the BBAs of the set E. Another proposition consists in comparing the BBA m_i with the BBA of the artificial expert representing the combined opinions of all the experts in E. The measure $Conf$ can be obtained by:

$$Conf(i, E) = d(m_i, m_*) \tag{16}$$

with m_* denoting combination of all BBAs of E. m_* can be obtained by using different combination rules and more precisely the conjunctive rule (equation 5). Once the conflict measure is obtained, the authors have proposed to compute reliability rates as follows:

$$\beta_i = f(Conf(i, E)) \tag{17}$$

where f is a decreasing function. The authors propose to use the function f defined as follows:

$$\beta_i = (1 - Conf(i, M)^{\lambda})^{1/\lambda} \tag{18}$$

with $\lambda > 0$. The authors in (Martin et al, 2008) recommend using $\lambda = 1.5$. Extensions of this work use the idea of sequential discount to manage the conflict when combining belief functions Klein and Colot (2010); Liu et al (2010). Schubert in (Schubert, 2011) uses the idea of sequential discount but the author employs the degree of falsity instead of the distance measure.

4 Intrinsic and Extrinsic based discounting factors

In the previous section, different measures were presented allowing the estimation of the two types of conflict. In spite of these measures, to our knowledge there is no reliability estimation method that links both types of conflict measure (see Section 3.3).

Example 1 The example, proposed in Table 1, presents the interest to use these two types of measures simultaneously. Initially in this example (Table 1-a), we study the fusion result of three involved sources but m_3 is inconsistent with the other ones. The discounting, based only on the extrinsic measure (equation 14), provides good result since the conflict decreases.

In the second part of this table (Table 1-b), the sources are now considered undecided. In this case, even through the discounting is based on an extrinsic measure, the conflict remains high.

	m_1	m_2	m_3	$\textcircled{\cap}$	discounting + $\textcircled{\cap}$
H_1	0.230	0.200	0.800	0.151	0.182
H_2	0.570	0.600	0.100	0.119	0.250
Ω	0.200	0.200	0.100	0.004	0.028
$\emptyset$	0.000	0.000	0.000	**0.726**	**0.540**

(a)

	m_1	m_2	m_3	$\textcircled{\cap}$	discounting + $\textcircled{\cap}$
H_1	0.450	0.400	0.470	0.156	0.158
H_2	0.450	0.500	0.430	0.174	0.177
Ω	0.100	0.100	0.100	0.001	0.001
$\emptyset$	0.000	0.000	0.000	**0.669**	**0.664**

(b)

Table 1 Extrinsic conflict measure and discounting.

The aim of the proposed approach is to anticipate the generation of a high conflictual source after the fusion phase. The discounting factors should be found by taking into consideration the ambiguity of each source (intrinsic measure) and the distance separating them (extrinsic measure). To achieve this purpose, we based our discounting approach on two distance measures criteria.

In this section, we propose two different discounting approaches that aim to discard the contradictory and non reliable sources that may eventually lead to an important conflict in the resulting fusion BBA. Let us assume the frame of discernment Ω containing all possible answers for a question Q relatively to the sources $S_1, \ldots, S_M$. In the fusion stage, to each processed BBA, a new discounting factor is assigned indicating its relevance and its global reliability.

We propose new methods for calculating discounting factors using the valuable information of the source such as confusion rate and belief function distance. Each proposed discounting function f must then fulfill the following constraints:

- f is an increasing function from $[0,1]^2 \rightarrow [0,1]$
- $f(1,1) = 1$ and $f(0,0) = 0$.

4.1 Weighted Generic Discounting Approach (GDA-W)

The GDA-W relies on a function f that gathers both aforementioned conflict measures. The function f can be written as follows:

$$(\delta, \beta) \rightarrow \frac{k.\delta + l.\beta}{k + l} \tag{19}$$

where $k > 0$ and $l > 0$ are the weight factors allowing the user to favor one distance measure rather than the other. In equation 19, δ denotes the internal conflict measure of the treated source indicating its confusion rate and β is the average distance between the treated source S_i and S_j with $j \in [1, \ldots, i-1, i+1, \ldots, M]$. We use the Jousselme distance defined in subsection 3.1, which is commonly used in belief measure works. Nevertheless, other Intrinsic and Extrinsic distances variants could be used. So, the values δ and β can be

defined by:

$$\delta = Pl\_IntC(m_i) \tag{20}$$

$$\beta = Dist(m_i) = \frac{\sum_{m_k \setminus m_i, k \in [1..M]} D(m_i, m_k)}{M-1}. \tag{21}$$

D is the Jousselme distance. Thus, the function f can be written as follows:

$$(\delta, \beta) \rightarrow \frac{k.Pl\_IntC(m_i) + l.\frac{\sum_{m_k \setminus m_i, k \in [1..M]} D(m_i, m_k)}{M-1}}{k+l}. \tag{22}$$

The classical discounting can be written as follows:

$$\begin{cases} m^{GDA-W}(B) = (1 - f(\delta, \beta)) \times m(B) & \forall B \subseteq \Omega \\ m^{GDA-W}(\Omega) = (1 - f(\delta, \beta)) \times m(\Omega) + f(\delta, \beta). \end{cases} \tag{23}$$

The determination of the weight factors can be found automatically by minimizing the following measures:

$$\begin{cases} k > 0 \\ l > 0 \\ E_{bet}(k, l) = \sum_{i=1}^{I} \sum_{n=1}^{N} (BetP^{(i)}(H_n) - U_n^i)^2 \end{cases} \tag{24}$$

where $BetP^{(i)}$ represents the pignistic probability of x^i vector from the learning base and U_n^i represents the x^i membership.

4.2 Exponent Generic Discounting Approach (GDA-E)

The GDA-E is also a discounting approach that estimates source's reliability based on both conflict origins. The GDA-E relies on a function g that can be written as follows:

$$(\delta, \beta) \rightarrow \beta^{(1-\delta)} \tag{25}$$

where β and δ are respectively the extrinsic measure and intrinsic measure defined, respectively, in equation 21 and 20. In (Samet et al, 2013), we proposed a different association between extrinsic and intrinsic measures such as the contradiction (see equation 12). The function g can be written as follows:

$$(\delta, \beta) \rightarrow (Dist(m_i))^{(1-Pl\_IntC(m_i))} \tag{26}$$

Thus, the discounting can be written as follows:

$$\begin{cases} m^{GDA-E}(B) = (1 - g(\delta, \beta)) \times m(B) & \forall B \subseteq \Omega \\ m^{GDA-E}(\Omega) = (1 - g(\delta, \beta)) \times m(\Omega) + g(\delta, \beta) \end{cases} \tag{27}$$

Table 2 shows the discounting value that could be associated to a BBA for extremum confusion and distance rates.

Table 2 The GDA-E discounting value for the extremum cases

Source	With Confusion $\delta = 1$	Without Confusion $\delta = 0$
Distant $\beta = 1$	$g(m) = 1$	$g(m) = Dist(m)$
Near $\beta = 0$	$g(m) = 1$	$g(m) = 0$

297 *Example 2* Let's consider the frame of discernment $\Omega = \{H_1, H_2\}$ and three
298 sources S_1, S_2 and S_3. The belief function values associated to those sources and their discounting values are computed in Table 3.

Table 3 Evaluation of discounting approach on an example

	S_1	S_2	S_3	$m^{\textcircled{\cap}}$	$m^{GDA-E}_{\textcircled{\cap}}$	$m^{GDA-W}_{\textcircled{\cap}}$
H_1	0.300	0.400	1.000	0.420	0.484	0.348
H_2	0.300	0.400	0.000	0.000	0.120	0.143
Ω	0.400	0.200	0.000	0.000	0.211	0.432
$\emptyset$	0.000	0.000	0.000	0.580	0.185	0.077
Intrinsic conflict: δ	0.300	0.400	0.000	-	-	-
Extrinsic conflict: β	0.319	0.305	0.524	-	-	-
GDA-E: $g(m)$	0.449	0.490	0.524	-	-	-
GDA-W: $f(m)(k = l = 1)$	0.307	0.352	0.262	-	-	-

299
300 As sketched by the statistics of Table 3, both GDA-W and GDA-E con-
301 sider S_3 as the most reliable source despite being distant from S_1 and S_2 (a
302 classical discounting approach would reject S_3 for being distant). Thanks to
303 its categorical constitution making it without any intrinsic conflict, S_3 is con-
304 sidered as a reliable source. The same explanation can be applied to S_1 and
305 S_2 where despite being close, neither of them can reinforce any hypothesis.
306 Even for GDA-E approach, the BBA's distance (extrinsic measure) is powered
307 by the confusion (intrinsic measure) that is why the GDA-E factor is equal to
308 the extrinsic measure when the source is not confused (source S_3). However,
309 it increases as far as the confusion of the source is increased. As it is shown
310 in Table 3, GDA-W and GDA-E have drastically decreased the conflict com-
311 paratively to the conjunctive sum, it falls from 0.580 to respectively 0.077 and
312 0.185.

313 In the following section, the GDA discounting approaches are integrated into
314 a belief based classifier.

315 5 GDA Classifier

316 In this section, we introduce the based belief function theory GDA classifier.
317 The proposed classifier relies on multi-source fusion and integrates the GDA
318 discounting approach for unreliable source detection. Several belief based clas-
319 sifiers exist, in which we can integrate our discounting approach such as the
320 likehood (Shafer, 1976) and the tree based classifiers (Elouedi et al, 2001).

In our case, we built the GDA classifier on the distance classifier (Denoeux, 1995) for its simplicity and combinatorial explosion avoidance. The properties of this classifier is detailed and compared in (Vannoorenbergue and Denoeux, 2001).

5.1 Distance Classifier

Introduced by Denoeux (1995), the Distance Classifier (DC) is a based belief function theory and multi-level fusion classification approach. It relies on a learning base in which we store the patterns x_i belonging to H_n^i class. Considering a vector x to classify, the application of the K Nearest Neighbors (KNN) algorithm on the learning base, provides k pieces of evidence. Indeed, each vector x_i, sufficiently close to x following a distance d brings information about x membership to H_n^i. This information is represented by a BBA m over the set Ω of classes. A fraction of the unit mass is assigned by m to the singleton H_n^i, and the rest is assigned to the whole frame of discernment Ω. The mass assigned to $m(\{H_n^i\})$ follows a decreasing function in distance d. For each neighbor x_i a BBA is modeled as follows:

$$\begin{cases} m_i(\{H_n\}) = \alpha^i \phi^i(d^i) \\ m_i(\Omega) = 1 - \alpha^i \phi^i(d^i) \end{cases} \tag{28}$$

where $0 < \alpha^i < 1$ is a constant. $\phi^i(.)$ is a decreasing function fulfilling $\phi^i(0) = 1$ and $lim_{d\to\infty}\phi^i(d) = 0$, d_i is the Euclidian distance between the vector x and the i-th prototype. The ϕ^i function might be an exponential function following this form:

$$\phi^s(d^s) = exp(-\gamma^s(d^s)^2) \tag{29}$$

where γ^s is a positive parameter associated to a prototype s and d^s is the distance between prototype s and x. A learning algorithm was proposed by Zouhal and Denoeux (1998) to determine the parameters γ^s in the equation (29) by optimizing an error criterion. The constructed BBAs are then fused following the Dempster's rule of combination as follows:

$$m = \oplus_{i\in[1,\ldots,I]} m_i \tag{30}$$

The aforementioned classification approach is the *mono-dimensional* variant of the distance classifier. In addition, the *multi-dimensional* strategy consists in modeling the information according to every characteristic x_j (with $j \in [1 : J]$) of the vector x to classify. The expression of equation 28 becomes:

$$\begin{cases} m_{ij}(\{H_n\}) = \alpha_j^i \phi_j^i(d_{ij}) \\ m_{ij}(\Omega) = 1 - \alpha_j^i \phi_j^i(d_{ij}) \end{cases} \tag{31}$$

where $0 < \alpha_j^i < 1$ is a constant and d_j^i represents the distance between the j-th component x_j of the vector x and its neighboring vector v_i ($i \in [1, K]$). The function ϕ_j^i can be expressed in the following way:

$$\phi_j^i(d) = exp(-\gamma_{ij}(d_{ij}^2) \tag{32}$$

The use of Dempster's combination operator makes it possible to merge those K belief functions. m_j is the resulting belief function and it is equal to:

$$m_j = \oplus_{i\in[1,k]} m_{ij}. \tag{33}$$

Thanks to its two hypothesis constructed BBA (see equation (31)), this model avoids combinatorial explosion resulting from several fusion processes. Thus, a unique belief function m is obtained by the application of the same fusion principle on those resulting J BBAs:

$$m = \oplus_{j\in[1,J]} m_j \tag{34}$$

with J the number of sources.

5.2 The GDA Classifier

The GDA discounting approach is integrated in the distance classifier in order to prune unreliable sources before the fusion phase. The level of the GDA integration differs following the chosen variant of the distance classifier. For the mono-dimensional distance classifier, the application of the GDA discounting formula (equation 23 or 27) is applied before the source's fusion (equation 30). However, for the multi-dimensional, the GDA discounting formula is applied after the neighbors fusion (equation 33) and before source's BBA fusion (equation 34). Figure 1 represents the proposed architecture for GDA classifier. The first part of the figure represents the belief function estimation based on distance. The GDA classifier part operates a discounting phase based on extrinsic and intrinsic conflict measures, source fusion and decision.

6 Experimentation and results

The experimentation of both GDA approaches were conducted at two stages. In the first stage the test was carried out on several UCI benchmarks (Frank and Asuncion, 2010). Indeed, the GDA classifier in its two version (with GDA-E and GDA-W discounting) was tested in a classification problem comparatively to several pioneer classifiers. In the second stage, we considered an urban image classification problem.

6.1 Data set classification experiments

The experimentation of the GDA classifier was carried out on several UCI benchmarks (Frank and Asuncion, 2010). The characteristics of these data

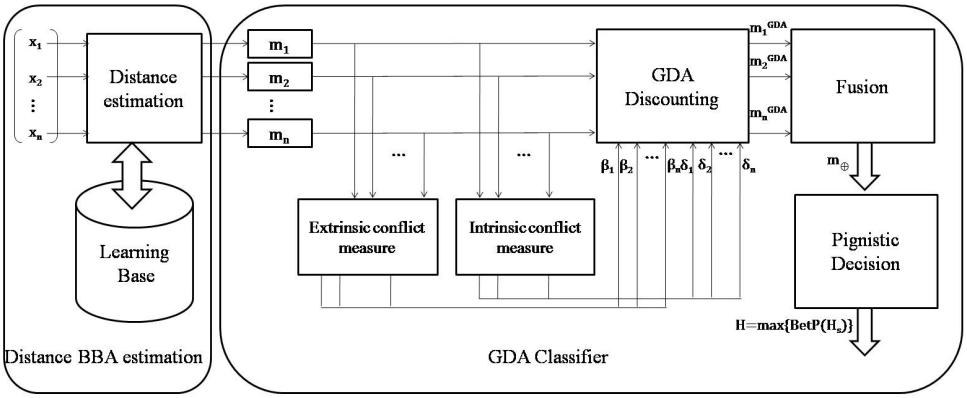

Fig. 1 The proposed GDA classifier architecture

Table 4 Data set characteristics description

Data set	**#Instances**	**#Attributes**	**#Classes**
Iris	150	4	3
Wine	178	13	3
ILPD (Indian Liver Patient Data set)	583	10	2
Diabetes	767	9	2

sets are summarized in Table 4. We carried out the tests using the mono-dimensional version of GDA classifier. For the classification task, we applied a cross-validation technique.

The results will be compared to several referenced works. Then, we are proposing an oriented discounting conflict management approach. We compare our results to Martin et al (2008)'s work (described in subsection 3.3 and denoted Mart) which has shown that it outperforms its predecessors. In order to get a general idea of our belief formalism classifier, we analyze the difference between the proposed approach and the Distance Classifier (DC). Since each tested method is based on the KNN algorithm, we fixed $K = 4$ for all of them. We also carried out a comparison with the $Naive\ Bayes\ Classifier$ (denoted as Bayes in Tables 5). Table 5 shows classification results of tested classifiers where '#' and '%' indicate, respectively, the number and the percentage of correct classified instances.

As highlighted by statistics shown in Table 5, we present better classification results than do Distance Classifier (DC) for all tested data sets. This fact shows the importance of using a discounting stage before fusion in order to discard unreliable sources. In addition, both classification accuracy are better than flagged out the Martin et al's approach in every treated data set. Since both classification methods (GDA and Martin et al's approach) differ only in the discounting method, this improvement highlights the importance

Table 5 Comparative results for datasets' classification

	DC		Mart		GDA-W		GDA-E		KNN		Bayes	
	#	%	#	%	#	%	#	%	#	%	#	%
Iris	147	98.00	147	98.00	**148**	**98.66**	**148**	**98.66**	143	95.33	144	96.00
Wine	154	86.51	151	84.83	**158**	**88.76**	**159**	**89.32**	169	94.94	**172**	**96.62**
ILPD	388	66.55	391	67.06	**392**	**67.23**	**392**	**67.23**	378	64.83	325	55.74
Diabete	538	70.05	538	70.05	**541**	**70.44**	**542**	**70.66**	539	70.18	**586**	**76.30**

of using the intrinsic measure to estimate the reliability of a source. We also compared other non based belief function theory classifiers such as KNN and $Naive\ Bayes\ Classifier$. By comparing GDA-E and GDA-W to the KNN classifier, we notice that we have also improved the obtained results for the Iris, ILPD and Diabete's data sets. This improvement can be interpreted as the contribution of uncertainty modeling and multi-source fusion. However, for the Wine data set, the KNN presents a better results. On the other hand, the $Naive\ Bayes\ Classifier$ presents the best classification results for Wine and Diabete data sets. This fact proves that Bayesian formalism handles better the uncertainty for those data sets.

In the following, we compare the belief based classification methods. Figures 2, 3, 4 and 5, respectively, show the different elements of Iris training base studied according to their petal width and length. In those figures, we illustrate the iso-pignistic curves where bright zones correspond to a high pignistic probability value whereas the dark one indicates the opposite. For Iris-setosa vectors (red colored in Figures), we notice that the pignistic probability is the same for every classification method thanks to the class uniformity of all extracted neighbors. For the two other classes, we noticed significant differences between studied approaches specially in the bordering area. In the DC approach Figure (see Figure 2), the class change is operated roughly leading to classification errors at the decision stage. However, GDA and Martin et al's approaches (Figure 3, 4 and 5) present a low pignistic probability in borders. This result is a natural consequence of discounted BBA that contributes to representing better the doubt between both classes. For Martin's and GDA-E discounting the doubt zone is the largest whenever compared to GDA-W which rejects fewer vectors in decision. Nevertheless, even though the reject zone is large for both methods, the GDA-E presents a high pignistic probability value than do Martin et al's approach.

6.2 Urban image classification experiments

The GDA classification was also tested on a Quickbird image covering urban areas of Strasbourg, taken in 2008, having four bands, each band has 2.44-2.88m/px as spatial resolution. This image contains a variety of objects: houses, parks, road, etc. Those objects can be reduced to three major concepts. For this reason, in the following, we are mainly interested in finding roads, buildings and vegetation which represent almost all possible ob-

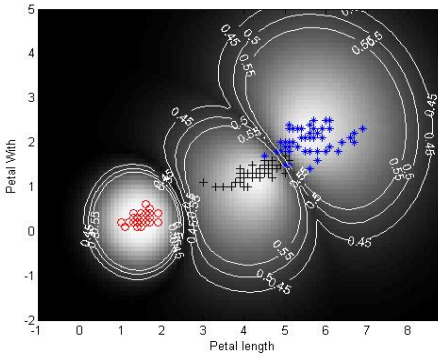

Fig. 2 Pignistic probability maxima for DC approach in Iris Data set.

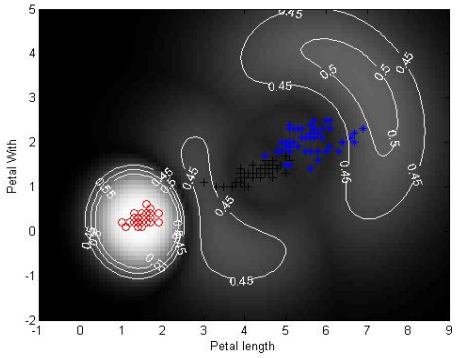

Fig. 3 Pignistic probability maxima for Martin's approach in Iris Data set

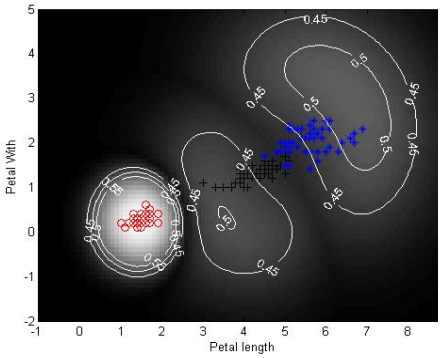

Fig. 4 Pignistic probability maxima for GDA-W approach in Iris Data set(k=0.5, l=2)

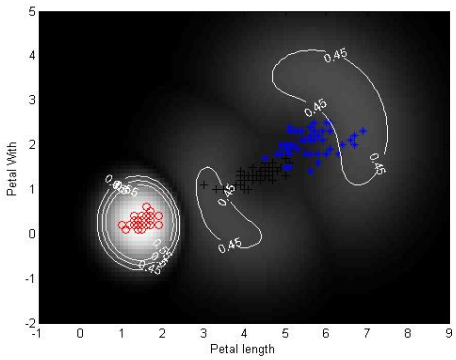

Fig. 5 Pignistic probability maxima for GDA-E approach in Iris Data set

jects in the image. Those three classes constitute our frame of discernment $\Omega = \{Roads, Building, Vegetation\}$. In order to extract those classes correctly, we have used five different sources. Some of the used sources correspond to a band from the image and others represent image products. The experiments were conducted on a basis of 8712 pixels. Indeed, we tested our approach on 2256 building pixels, 2926 road pixels and 3530 vegetation pixels. The considered sources are:

- R-G-B
- Near Infrared (NIR)
- Normalized Difference Vegetation Index (NDVI).

The NDVI denotes the vegetation index (Crippen, 1990) and is given by this formula:

$$NDVI = \frac{NIR - VIS}{NIR + VIS} \tag{35}$$

	<0.1	<0.2	<0.4	<0.6	≥ 0.6
Conjunctive	0.01%	0.33%	4.66%	24.20%	**70.80%**
Martin discounting + Conjunctive	0.01%	0.33%	4.98%	26.42%	**68.26%**
GDA-W approach + Conjunctive	0.01%	0.39%	12.92%	78.45%	**8.23%**
GDA-E approach + Conjunctive	0.56%	13.30%	51.11%	34.81%	**0.22%**

Table 6 Total conflict rate comparative result in urban image classification.

where VIS and NIR respectively stand for the radiometry measurements acquired in the visible (red) and near-infrared regions. Each source can identify the considered classes with a certain level of reliability. This fact makes from this context an adequate experimentation field for the GDA discounting approaches. Figure 6 shows the original high-resolution image that we tried to classify.

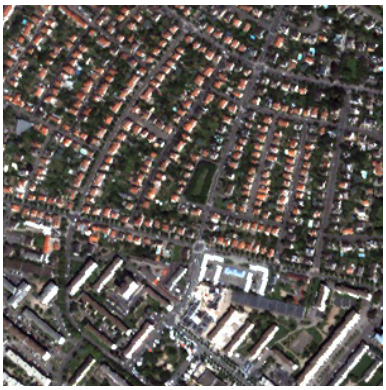

Fig. 6 Original high-resolution image of a Strasbourg site.

We compared the conflict rate of our proposed approach with those obtained respectively by conjunctive operator and (Martin et al, 2008) discounting approach in order to highlight conflict decrease. As it is shown in Table 6, GDA-E as well as GDA-W have improved the rates of conflict in the fused BBA comparatively to Martin et al's approach. The best improvement is carried by the GDA-E where conflict dropped from 70.80 to 0.22.

Figures 7 shows the classification results obtained with the use of only the extrinsic measure, where the vegetation, road, building classes are represented respectively by the colors green, gray and red. The GDA classifier with these settings is denoted the GDA-without and its classification performance are shown in Tables 7, 8 and 9. In addition, Figure 8 shows the GDA-W classification results obtained with the integration of both conflict origins. Indeed, the intrinsic conflict rate is an important piece of information about the total reliability of the source. The less the source is confused, the more it is reliable.

The proposed generic discounting is based on two pieces of information: intrinsic and extrinsic conflict measures (equation (25)). The integration of the

Fig. 7 Classification using GDA-W with only the extrinsic measure ($k = 0$).

Fig. 8 Classification using GDA-W with the extrinsic and intrinsic measures ($k = 0.2$ and $l = 1$).

	Mart	DC	GDA-W	GDA-E	GDA-without	KNN	Bayes
Building	**72.79%**	**68.25%**	**72.29%**	**77.14%**	**60.49%**	**92.40%**	**79.51%**
Road	25.90%	28.19%	26.48%	22.41%	23.24%	7.55%	20.49%
Vegetation	1.31%	3.56%	1.23%	0.55%	16.27%	0.05%	0.00%

Table 7 Comparative classification results for the building class.

	Mart	DC	GDA-W	GDA-E	GDA-without	KNN	Bayes
Building	15.21%	14.97%	14.97%	18.20%	19.15%	5.68%	4.34%
Road	**84.51%**	**81.69%**	**84.85%**	**81.03%**	**80.61%**	**93.97%**	**93.80%**
Vegetation	0.28%	0.34%	0.18%	0.77%	0.24%	0.35 %	1.86 %

Table 8 Comparative classification results for the road class.

	Mart	DC	GDA-W	GDA-E	GDA-without	KNN	Bayes
Building	0.00%	0.00%	0.21%	0.00%	0.34%	0.00%	0.12%
Road	0.36%	1.74%	0.15%	0.07%	2.35%	0.34 %	0.14%
Vegetation	**99.64%**	**98.26%**	**99.64%**	**99.93%**	**97.31%**	**99.66%**	**99.74%**

Table 9 Comparative classification results for the vegetation class.

intrinsic conflict rate for a source S_i constitutes the main added value of our method. In order to assess its contribution to classification improvement, we have tried to compare the classification results obtained with only the extrinsic information (denoted GDA-without) to the classifier integrating both conflict information (GDA-E and GDA-W). In addition, we compared ourselves to the KNN and the *Naive Bayes Classifier* (denoted *Bayes*).

Tables 7, 8 and 9 illustrate the confusion matrix of tested classification approaches for buildings, roads and vegetation detection. They also highlight the intrinsic measure integration contribution. Indeed, by comparing the GDA-W and GDA-E to GDA-without, all considered classes detection was improved with different proportions. The best improvement was the building class which means that fused information sources were confused when we tried to classify

them. In this case, the GDA-E and GDA-W consider discount the sources following their internal conflicts and their contradiction distances. Furthermore, road and vegetation class detection was slightly improved relatively to the building class. Nevertheless, the provided results were high. Indeed, since each source is specialized in detecting specific classes, the introduction of the GDA allowed us to avoid conflict generation by discounting those unreliable sources. The improvement of detection of all classes (comparatively to GDA-without) illustrates the compromises in terms of reliability that GDA seeks to find between source each time a pixel is under classification. For example, whenever handling a road pixel, the NDVI source (specialized in vegetation detection) is highly discounted since its assigned BBA is confused.

We also conducted comparative experiments to other based belief function works such as Martin et al (2008) discounting approach. We also compared our results to the Distance Classifier (DC) based on distance BBA estimation (see subsection 5.1) and pignistic decision (see subsection 2.5). As is shown in Table 7, 8 and 9, comparatively to the DC approach, we sharply improved the results for each class. We have also improved Martin et al's result for the road class with the GDA-W discounting and we maintained the same average for the vegetation. Interestingly enough, the GDA-E provides the best results for the building and the vegetation recognition considering all belief based classification approaches. In order to position our work in terms of pure classification, we compared our results to non belief based known classifiers such as KNN and $NaiveBayes$. In general, KNN provides the best classification results for building and road detection. For the vegetation class all classifiers provide almost the same good classification rates with a slight advantage for the Bayesian one. Despite the performance difference between our approaches and the KNN in this particular context of application, comparison can not be made directly since they rely on different formalism. In addition, we do provide the best result comparatively to belief classifiers and discounting based approach.

Figures 10, 11, 12 and 13 illustrate the classification of an urban site (Figure 9) with respectively the proposed DC, Martin, GDA-W and GDA-E classification approaches. We can notice that in this sample we improved through our approach the extraction of the road and building classes.

The figure 15 and 14 show the pixels (i.e, belief functions) where we registered an important conflict rate and the GDA approaches changed their initial classification (DC classification approach).

7 Conclusion

In this work, we introduced a conflict management approach named GDA. It allows to discount any information source following its estimated reliability. We introduced a new measure of reliability based on two conflict origins: Intrinsic and Extrinsic conflict. The discounting approach, proposed in two different versions GDA-W and GDA-E, was integrated into a based distance

Fig. 9 Original image.

Fig. 10 Classification using DC approach.

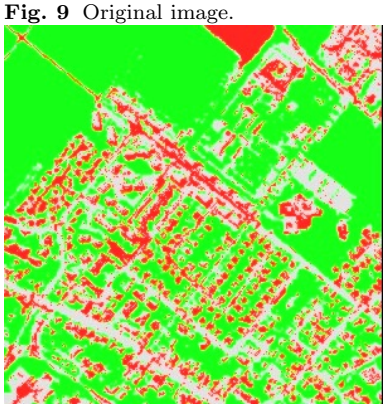

Fig. 11 Classification using Martin et al's discounting approach.

Fig. 12 Proposed GDA-W classification.

Fig. 13 Classification using GDA-E discounting approach.

belief function classifier. The GDA approach was thoroughly experimented on several UCI Data sets and on urban image classification problem. The provided results confirm the contribution of both conflict measures association. Even though very encouraging results were obtained, comparatively with other dis-

Fig. 14 Pixels where GDA-E approach changed the initial classification.

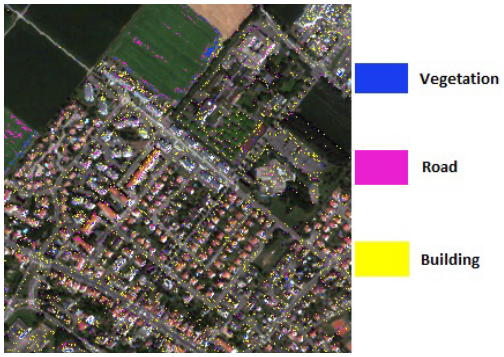

Fig. 15 Pixels where GDA-W approach changed the initial classification.

counting approaches, we aim to improve the introduced approach in further works by studying other conflict measures association. Additionally, further conflict origins can be studied like lying and insincere sources. Image classification improvement could be investigated by studying other approaches. In this work, we considered a pixel classification but in future work we plan to consider a region based approach. Indeed, a region based approach can provide valuable information such as : dimension, shape, etc.

References

Chen L, Shi W, Deng Y, Zhu Z (2005) A new fusion approach based on distance of evidences. Journal of Zhejian University Science 6A:476–482

Crippen RE (1990) Calculating the vegetation index faster. Remote Sensing of Environment 34(1):71 – 73

Cuzzolin F (2008) A geometric approach to the theory of evidence. IEEE transaction, Man, and Cybernetics-Part C: Application and reviews 38(4):522–534

Daniel M (2010) Conflicts within and between belief functions. in Proceedings of the International Conference on Information Processing and Management of Uncertainty, IPMU'10, Dortmund, Germany pp 696–705

Dempster A (1967) Upper and lower probabilities induced by multivalued mapping. AMS-38

Deng Y, Shi W, Zhu Z, Liu Q (2004) Combining belief functions based on distance of evidence. Decision Support Systems 38(3):489–493

Denoeux T (1995) K-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics 25(5):804–813

Destercke S, Burger T (2012) Toward an axiomatic definition of conflict between belief functions. IEEE Systems, Man and Cybernetics - part B

558 43(2):585–596
559 Elouedi Z, Mellouli K, Smets P (2001) Belief decision trees: theoretical foun-
560 dations. Int J Approx Reasoning 28(2-3):91–124
561 Elouedi Z, Mellouli K, Smets P (2004) Assessing sensor realibility for multi-
562 sensor data fusion within the Tranferable Belief Model. IEEE Transaction
563 on System, Man and Cybernetics - Part B 34(1):782–787
564 Elouedi Z, Lefevre E, Mercier D (2010) Discounting of a belief function using a
565 confusion matrix. in Proceedings of the 22th IEEE International Conference
566 on Tools with Artificial Intelligence, ICTAI'2010, Arras, France pp 287–294
567 Frank A, Asuncion A (2010) UCI machine learning repository. URL
568 http://archive.ics.uci.edu/ml
569 George T, Pal NR (1996) Quantification of conflict in Dempster-Shafer frame-
570 work: A new approach. International Journal of General Systems 24(4):407–
571 423
572 Guo H, Shi W, Deng Y (2006) Evaluating sensor reliability in classification
573 problems based on evidencetheory. IEEE transactions on Systems, Man, and
574 Cybernetics, Part B 36(5):970–981
575 Jousselme AL, Maupin P (2012) Distance in evidence theory: Comprehensive
576 survey and generalizations. International Journal of Approximate Reasoning
577 53(2):118–145
578 Jousselme AL, Grenier D, Bossé E (2001) A new distance between two bodies
579 of evidence. Information Fusion 2:91–101
580 Klein J, Colot O (2010) Automatic discounting rate computation using a dis-
581 sent criterion. in Proceedings of the Workshop on the Theory of Belief Func-
582 tions, Brest, France pp 1–6
583 Klein J, Colot O (2011) Singular sources mining using evidential conflict anal-
584 ysis. International Journal of Approximate Reasoning 52(9):1433–1451
585 Liu W (2006) Analyzing the degree of conflict among belief functions. Artificial
586 Intelligence 170:909–924
587 Liu ZG, Pan Q, Cheng YM, Dezert J (2010) Sequential adaptative combination
588 of unreliable sources of evidence. in Proceedings of Workshop on the Theory
589 of Belief Function, Brest, France p Paper no 89
590 Martin A (2012) About conflict in the theory of belief functions. in Proceedings
591 of International Conference on Belief Functions, BELIEF'2012, Compiègne,
592 France pp 161–168
593 Martin A, Jousselme AL, Osswald C (2008) Conflict measure for the dis-
594 counting operation on belief functions. in Proceedings of 11th International
595 Conference on Information Fusion, Cologne, Germany pp 1003–1010
596 Samet A, Hammami I, Lefevre E, Hamouda A (2013) Generic discounting
597 evaluation approach for urban image classification. In Proceedings of 3rd
598 international symposium on Integrated Uncertainty in Knowledge Modelling
599 and Decision Making, IUKM'2013, Beijing, China pp 79–90
600 Schubert J (2011) Conflict management in Dempster-Shafer theory using the
601 degree of falsity. International Journal of Approximate Reasoning 52(3):449–
602 460

603 Sentz K, Ferson S (2002) Combination of evidence in Dempster-Shafer theory.
604 Tech. rep.
605 Shafer G (1976) A Mathematical Theory of Evidence. Princeton University
606 Press, Princeton
607 Smarandache F, Martin A, Osswald C (2011) Contradiction measures and
608 specificity degrees of basic belief assignments. in Proceedings of Interna-
609 tional Conference on Information Fusion, Chicago, Illinois pp 1–8
610 Smets P (1990) The combination of evidence in the Transferable Belief Model.
611 IEEE Transactions on Pattern Analysis and Machine Intelligence 12(5):447–
612 458
613 Smets P (2005) Decision making in the TBM : The necessity of the pignistic
614 transformation. International Journal of Approximate Reasoning 38:133–
615 147
616 Smets P (2007a) Analyzing the combination of conflicting belief functions.
617 Information Fusion 8(4):387–412
618 Smets P (2007b) Analyzing the combination of conflicting belief functions.
619 Information Fusion 8(4):387–412
620 Smets P, Kennes R (1994) The Transferable Belief Model. Artificial Intelli-
621 gence 66(2):191–234
622 Tessem B (1993) Approximations for efficient computation in the theory of
623 evidence. Artificial Intelligence pp 315–329
624 Vannoorenbergue P, Denoeux T (2001) Likelihood-based vs. distance-based
625 evidential classifiers. In Proceedings of International Conference on Fuzzy
626 Systems FUZZ-IEEE, Melbourne, Australia pp 320 – 323
627 Yager RR (1992) On considerations of credibility of evidence. International
628 Journal of Approximate Reasoning 7:45–72
629 Zadeh LA (1994) A simple view of the Dempster-Shafer theory of evidence and
630 its implicationfor the rule of combination. The AI Magazine 7, no. 2:85–90
631 Zouhal L, Denoeux T (1998) An evidence-theoretic K-NN rule with parameter
632 optimization. IEEE Transactions on Systems, Man and Cybernetics, Part
633 C 28(2):263–271

B.2 Evidential Database : a new generalization of databases ?

Cet article présente l'interêt de la fouille de données évidentielles. Dans ce papier, nous démontrons que la mesure de support précise est une généralisation des mesures de support binaire et probabiliste. En ce qui concerne la mesure de support flou, nous avons montré qu'elle peut être calculée dans le cadre évidentielle.

A. Samet, E. Lefevre, S. Ben Yahia : Evidential Database : a new generalization of databases ?. *In Proceedings of 3rd International Conference on Belief Functions, Belief 2014, London, UK*, pages 105-114, 2014

Evidential Database: a new generalization of databases?

Ahmed Samet[1,2], Eric Lefèvre[2], and Sadok Ben Yahia[1]

[1] Université Tunis El Manar, LIPAH, Faculty of Sciences of Tunis , Tunisia
{ahmed.samet, sadok.benyahia}@fst.rnu.tn
[2] Univ. Lille Nord de France UArtois, EA 3926 LGI2A, F-62400, Béthune, France
eric.lefevre@univ-artois.fr

Abstract. In this paper, we tackle the problem of data representation in several types of databases. A detailed survey of the different support measures in the major existing databases is described. The reminder of the paper aims to prove the importance of using evidential databases in case of handling imperfect information. The evidential database generalizes several ones by the use of specific Basic Belief Assignments. In addition, we show that the precise support, initially introduced on evidential database, generalizes several support measures.

Keywords: Evidential database, Binary database, Probabilistic database, Fuzzy database, Support

1 Introduction

Data mining is a technique that uses a variety of data analysis tools to discover, hidden but interesting patterns and relationships in data that may be used to make valid predictions. Thanks to its simple formulas, it associates performance and quality in its retrieved results. For this reason, it is used in various fields and attracted interest in different applications [9].
The first studies on data mining relies on a data model under which transactions captured doubtless facts about the items that are contained in each transaction. These *binary databases* have only two scenarios : 1 if an element exists, 0 otherwise. However, in many applications, the existence of an item in a transaction is better captured by likelihood measures. The obvious limits of the binary databases in handling such types of data led the data mining community to adopt imprecise frameworks in order to mine more pertinent knowledge.
In this paper, we present a non exhaustive review of existing data mining databases. The characteristics of binary, probabilistic, fuzzy and evidential databases are detailed. The support measures in the databases are presented. The aim of this paper is to demonstrate the pivotal role of the evidential database, which relies on the evidence theory [5, 12], in representing imprecision and uncertainty. The importance of using an evidential database rather than the other

ones is justified. Indeed, we prove that the precise support measure [10] in evidential databases is a generalization of that of the classical ones.
The remainder of the paper is organized as follows: in section 2, the key basic settings of the evidential database are recalled. In section 3, the binary database is studied and its relationship with the evidential database is highlighted. In section 4, probabilistic databases are scrutinized and the correlation between the precise support and the probabilistic support is highlighted. Section 5 stresses on the snugness connection between fuzzy databases with the evidential ones. Finally, we conclude and we describe issues for future work.

2 Evidential database and precise support

In this section, we detail the main concepts of evidential databases as well as as the notion of precise support.

2.1 Evidential Database concept

Introduced by Lee [8], the evidential database was aimed at modelling imperfect information. This type of database is supposed to handle imprecise and uncertain data. An evidential database is a triplet $\mathcal{EDB} = (\mathcal{A}_{\mathcal{EDB}}, \mathcal{O}, R_{\mathcal{EDB}})$. $\mathcal{A}_{\mathcal{EDB}}$ is a set of attributes and $\mathcal{O}$ is a set of d transactions (i.e., lines). Each column A_i $(1 \leq i \leq n)$ has a domain θ_{A_i} of discrete values. $R_{\mathcal{EDB}}$ expresses the relation between the j^{th} line (i.e., transaction T_j) and the i^{th} column (i.e., attribute A_i) by a normalized BBA as follows:

$$m_{ij} : 2^{\theta_{A_i}} \rightarrow [0,1] \quad with$$

$$\begin{cases} m_{ij}(\emptyset) = 0 \\ \sum\limits_{\omega \subseteq \theta_{A_i}} m_{ij}(\omega) = 1. \end{cases} \tag{1}$$

Table 1: Evidential transaction database $\mathcal{EDB}$

Transaction	Attribute A	Attribute B
T1	$m(A_1) = 0.7$	$m(B_1) = 0.4$
	$m(\theta_A) = 0.3$	$m(B_2) = 0.2$
		$m(\theta_B) = 0.4$
T2	$m(A_2) = 0.3$	$m(B_1) = 1$
	$m(\theta_A) = 0.7$	

Table 1 illustrates an example of an evidential database. An item corresponds to a focal element. An itemset corresponds to a conjunction of focal elements

having different domains. The inclusion operator is defined in [3] such that for two itemsets X and Y, we have:

$$X \subseteq Y \iff \forall x_i \in X, x_i \subseteq y_i.$$

where x_i and y_i are the i^{th} element of X and Y. For the same evidential itemsets X and Y, the intersection operator is defined as follows:

$$X \cap Y = Z \iff \forall z_i \in Z, z_i \subseteq x_i \text{ and } z_i \subseteq y_i.$$

An *evidential associative rule* R is a causal relationship between two itemsets that can be written in the following form $R : X \rightarrow Y$ such that $X \cap Y = \emptyset$.

Example 1. In Table 1, A_1 is an item and $\theta_A \times B_1$ is an itemset such that $A_1 \subset \theta_A \times B_1$ and $A_1 \cap \theta_A \times B_1 = A_1$. $A_1 \rightarrow B_1$ is an evidential associative rule.

In the following subsection, we consider the precise support and confidence measures.

2.2 Support and confidence in evidential database

Several definitions for the support's estimation have been proposed for the evidential itemsets such as [3, 6]. Those definitions assess the support based on the belief function $Bel()$. The based belief support is constructed from the Cartesian product applied to the evidential database. Interested readers may refer to [6]. The support is computed as follows:

$$Support_{\mathcal{EDB}}(X) = Bel_{\mathcal{EDB}}(X) \tag{2}$$

such that:

$$Bel : 2^{\theta} \rightarrow [0, 1] \tag{3}$$

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B). \tag{4}$$

In a previous work [10], we introduced a new metric for support estimation. The latter has been shown to provide more accuracy and to overcome several drawbacks of using the belief function. This measure is called Precise support Pr and it is defined by:

$$Pr : 2^{\theta_i} \rightarrow [0, 1] \tag{5}$$

$$Pr(x_i) = \sum_{x \subseteq \theta_i} \frac{|x_i \cap x|}{|x|} \times m_{ij}(x) \qquad \forall x_i \in 2^{\theta_i}. \tag{6}$$

The evidential support of an itemset $X = \prod_{i \in [1 \ldots n]} x_i$ in the transaction T_j (i.e., Pr_{T_j}) is then equal to:

$$Pr_{T_j}(X) = \prod_{x_i \in \theta_i, i \in [1 \ldots n]} Pr(x_i). \tag{7}$$

Thus, the evidential support $Support_{\mathcal{EDB}}$ of the itemset X becomes:

$$Support_{\mathcal{EDB}}(X) = \frac{1}{d} \sum_{j=1}^{d} Pr_{T_j}(X). \tag{8}$$

Additionally, in [11], we introduced a new measure of confidence for evidential associative rules that we called the *precise confidence measure*. Let us assume an evidential association rule such as $R : R_a \rightarrow R_c$, where R_c and R_a respectively, denote the conclusion and the antecedent (premise) part of the rule R. The precise confidence measure can be written as follows:

$$Confidence(R : R_a \rightarrow R_c) = \frac{\sum_{j=1}^{d} Pr_{T_j}(R_a) \times Pr_{T_j}(R_c)}{\sum_{j=1}^{d} Pr_{T_j}(R_a)}. \tag{9}$$

In the following sections, we highlight the relationships between evidential databases and the main other ones. The link between existing measures and the evidential precise one is also demonstrated.

3 Binary data mining

The first database variants studied from a data mining view are the binary ones. A binary database can be represented by a triplet $\mathcal{BDB} = (\mathcal{A}, \mathcal{O}, R_{\mathcal{BDB}})$. $\mathcal{A}$ represents the set of n binary attributes (i.e., columns). $R_{\mathcal{BDB}}$ is the relation that reflects the existence of an item in a transaction by only the values 0 and 1. $R_{\mathcal{BDB}}(A_i, T_j) = 1$ means that the item A_i exists in the transaction T_j and $R_{\mathcal{BDB}}(A_i, T_j)$ is set equal to 0 otherwise.

Since the inception of the Apriori algorithm [2], several other approaches have been introduced to reduce the computational complexity of mining these "frequent" binary itemsets. The support of an item A_i in a transaction T_j is defined as follows:

$$Support_{T_j}(A_i) = R_{\mathcal{BDB}}(A_i, T_j). \tag{10}$$

The support of an item A_i in those binary databases is still computed with the same manner:

$$Support(A_i) = \sum_{j=1}^{d} R_{\mathcal{BDB}}(A_i, T_j) = count(A_i). \tag{11}$$

The same goes for an itemset $A \cup B$ (or $A \times B$ if we keep the product notation):

$$Support(A \times B) = count(A \cup B). \tag{12}$$

Thus, the support is computed by counting the number of transactions having both A and B. From the support, the confidence measure of a rule $R : R_a \rightarrow R_c$ is computed as follows:

$$confidence(R : R_a \rightarrow R_c) = \frac{count(R_a \cup R_c)}{count(R_a)}. \tag{13}$$

A binary database can be constructed by redefining the $R_{\mathcal{EDB}}$ as a precise BBA. Indeed, each item $A_i \in \mathcal{A}$ can be redefined as an evidential item having the following frame of discernment $\theta_{A_i} = \{\exists, \nexists\}$. $\exists$ and $\nexists$ denote respectively the existence and absence of the attribute A_i in the considered transaction. Such a BBA can be written as follows:

$$\begin{cases} m_{ij}(\{\exists\}) = R_{\mathcal{BDB}}(A_i, T_j) \\ m_{ij}(\{\nexists\}) = 1 - R_{\mathcal{BDB}}(A_i, T_j) \end{cases} \tag{14}$$

where m_{ij} is equivalent to a certain BBA. In that case, the support measure proposed in [10] is equivalent to the binary support equation defined in Equation (10). To demonstrate that equivalence, let us consider a binary database $\mathcal{D}$ and the evidential database $\mathcal{EDB}$ constructed as in the described procedure. Suppose that $R_{\mathcal{BDB}}(A_i, T_j) = 1$ such that $A_i \in \mathcal{A}$, then the corresponding evidential attribute is an $A_i \in \mathcal{A}_{\mathcal{EDB}}$ with $\theta_{A_i} = \{\exists, \nexists\}$:

$$Pr_{T_j}(\exists) = \frac{|\exists \cap \exists|}{|\exists|} m_{ij}(\{\exists\}) + \frac{|\nexists \cap \exists|}{|\nexists|} m_{ij}(\{\nexists\}) = m_{ij}(\{\exists\}) = R_{\mathcal{BDB}}(A_i, T_j). \tag{15}$$

From this point, we deduce that the evidential precise support is a generalization of the binary one. The same goes for the precise confidence given in Equation (9) that generalizes binary confidence since they both rely on the same support fraction.

Example 2. In this example, Table 2 shows how to create an evidential database from a binary one.

The equivalency of the support measure is shown for the itemset $B \times C$.

The support of the itemset $B \times C$ from the transactions of Table 2.a is $Support(B \times C) = \frac{2}{3}$. In the evidential database, it is computed as follows:

$$Support_{\mathcal{EDB}}(B \times C) = \frac{1}{3} \sum_{j=1}^{3} Pr_{T_j}(A) \times Pr_{T_j}(B)$$
$$Support_{\mathcal{EDB}}(B \times C) = \frac{1}{3}(m_{21}(\{\exists\}) \times m_{31}(\{\exists\}) + m_{22}(\{\exists\}) \times m_{32}(\{\exists\}) + m_{23}(\{\exists\}) \times m_{33}(\{\exists\})) = \frac{2}{3}$$

Thus, the support retrieved from the binary database is the same as the precise support computed from the evidential database.

Table 2: The evidential transformation of $\mathcal{BDB}$ (Table (a)) to $\mathcal{EDB}$ (Table (b))

	A	B	C
T_1		X	X
T_2	X	X	
T_3		X	X

(a)

	A	B	C
T_1	$m_{11}(\{\exists\}) = 0$	$m_{21}(\{\exists\}) = 1$	$m_{31}(\{\exists\}) = 1$
	$m_{11}(\{\nexists\}) = 1$	$m_{21}(\{\nexists\}) = 0$	$m_{31}(\{\nexists\}) = 0$
T_2	$m_{12}(\{\exists\}) = 1$	$m_{22}(\{\exists\}) = 1$	$m_{32}(\{\exists\}) = 0$
	$m_{12}(\{\nexists\}) = 0$	$m_{22}(\{\nexists\}) = 0$	$m_{32}(\{\nexists\}) = 1$
T_3	$m_{13}(\{\exists\}) = 0$	$m_{23}(\{\exists\}) = 1$	$m_{33}(\{\exists\}) = 1$
	$m_{13}(\{\nexists\}) = 1$	$m_{23}(\{\nexists\}) = 0$	$m_{33}(\{\nexists\}) = 0$

(b)

In the following section, we review the basics of the probabilistic support. A transformation method from a probabilitic database to evidential one is introduced. The equivalency between the probabilistic support and the precise one is studied.

4 Probabilistic data mining

Probabilistic data mining [1] was introduced to represent imperfect information thanks to the probability support. It can be represented by a triplet $\mathcal{PDB} = (\mathcal{A}_{\mathcal{PDB}}, \mathcal{O}, R_{\mathcal{PDB}})$. The degree of existence of the item A_i in the transaction T_j is measured through the probability function $p(A_i, T_j) \in [0, 1]$. The support of an itemset $X \in \mathcal{A}_{\mathcal{PDB}}$ in such type of database is defined as follows [4]:

$$p(X, T_j) = \prod_{i \in X} p(i, T_j). \tag{16}$$

Thus, the support of an itemset X in a database is the sum of its expected probability in the transaction:

$$Support_{\mathcal{PDB}}(X) = \sum_{j=1}^{d} p(X, T_j). \tag{17}$$

An equivalent evidential database can be constructed through using Bayesian BBA[3]. The BBA can be modeled on a two-member-based frame of discernment $\theta_i = \{\exists, \nexists\}$ where $\exists$ indicates that A_i belongs to the considered transaction, whereas $\nexists$ performs the opposite. Such a BBA can be constructed as follows:

$$\begin{cases} m_{ij}(\{\exists\}) = p(i, T_j) \\ m_{ij}(\{\nexists\}) = 1 - p(i, T_j). \end{cases} \tag{18}$$

With this construction, the probabilistic support defined in Equation (17) is equivalent to the proposed precise support. Indeed, the assertion can be verified

[3] A BBA is called Bayesian only if all its focal sets are singletons.

i.e.:

$$Pr_{T_j}(\exists) = \frac{|\exists \cap \exists|}{|\exists|} m_{ij}(\{\exists\}) + \frac{|\not\exists \cap \exists|}{\not\exists} m_{ij}(\{\not\exists\}) = m_{ij}(\{\exists\}) = p(i, T_j). \quad (19)$$

As is the case for a binary database, the Evidential Data mining Algorithm (EDMA) generalizes the probabilistic version of Apriori: i.e., U-Apriori [4].

Example 3. Table 3 shows how to create an evidential database from a probabilistic one.

Table 3: The evidential transformation of $\mathcal{PDB}$ (Table (a)) to $\mathcal{EDB}$ (Table (b))

	A	B	C
T_1	0.0	0.7	0.8
T_2	0.9	0.7	0.1
T_3	0	0.8	0.7

(a)

	A	B	C
T_1	$m_{11}(\{\exists\}) = 0$	$m_{21}(\{\exists\}) = 0.7$	$m_{31}(\{\exists\}) = 0.8$
	$m_{11}(\{\not\exists\}) = 1$	$m_{21}(\{\not\exists\}) = 0.3$	$m_{31}(\{\not\exists\}) = 0.2$
T_2	$m_{12}(\{\exists\}) = 0.9$	$m_{22}(\{\exists\}) = 0.7$	$m_{32}(\{\exists\}) = 0.1$
	$m_{12}(\{\not\exists\}) = 0.1$	$m_{22}(\{\not\exists\}) = 0.3$	$m_{32}(\{\not\exists\}) = 0.9$
T_3	$m_{13}(\{\exists\}) = 0$	$m_{23}(\{\exists\}) = 0.8$	$m_{33}(\{\exists\}) = 0.7$
	$m_{13}(\{\not\exists\}) = 1$	$m_{23}(\{\not\exists\}) = 0.2$	$m_{33}(\{\not\exists\}) = 0.3$

(b)

The equivalency of the support measure is shown for the itemset $B \times C$. The support of the itemset $B \times C$ from the transactions of the Table 3.a is $Support(B \times C) = \frac{(0.7\times0.8)+(0.7\times0.1)+(0.8\times0.7)}{3} = 0.4$. In the evidential database, it is computed as follows:

$$Support_{\mathcal{EDB}}(B \times C) = \frac{1}{3} \sum_{j=1}^{3} Pr_{T_j}(A) \times Pr_{T_j}(B)$$
$$Support_{\mathcal{EDB}}(B \times C) = \frac{1}{3}(m_{21}(\{\exists\}) \times m_{31}(\{\exists\}) + m_{22}(\{\exists\}) \times m_{32}(\{\exists\}) + m_{23}(\{\exists\}) \times m_{33}(\{\exists\})) = \frac{1.2}{3} = 0.4$$

Thus, the support retrieved from the probabilistic database is the same as the precise support computed from the evidential database.

In the following section, we review the basics of fuzzy data mining and we study its relation with the evidential one.

5 Fuzzy Data mining

Let us assume the triplet $\mathcal{FDB} = (\mathcal{A}_{\mathcal{FDB}}, \mathcal{O}, R_{\mathcal{FDB}})$ that denotes a fuzzy database. $R_{\mathcal{FDB}}$ denotes the fuzzy relationship between an item and a transaction expressed through a membership function. The membership function $\mu_{T_j}(i) = \alpha$ ($\alpha \in [0, 1]$) rates the degree of membership of the considered item to the transaction T_j. The support computation in such databases is done by the use of the

count() function in the following manner [7]:

$$count(i) = \sum_{j=1}^{d} \mu_{T_j}(i). \tag{20}$$

The support of item i in the fuzzy database is found as follows:

$$Support(i) = \frac{count(i)}{d}. \tag{21}$$

Thus, for an itemset X of size q such that $x_i \in X$ and $i \in [1, q]$, the support becomes:

$$support(X) = \frac{\sum_{j=1}^{d} min\{\mu_{T_j}(x_i), i = 1 \dots q\}}{d}. \tag{22}$$

The numerator of the support could be seen as the Gödel t-norm (minimum t-norm).
Assuming a fuzzy database is available, it is possible to construct an evidential database. In addition, the precise support sustains fuzzy support in its formulation. Indeed, as can be seen in Equation (8), the precise support is also equal to the sum of the transactional support divided by the database size.
In the following, we show how to obtain analogous evidential support of the fuzzy support. Assuming an attribute $A_i \in \mathcal{A}_{\mathcal{EDB}}$ having a frame of discernment θ_{A_i} such that $\omega_1 \subset \dots \subset \omega_n \subseteq \theta_{A_i}$, the corresponding BBA for a fuzzy relation $R_{\mathcal{FDB}}(\omega_1, T_j) = \mu_{T_j}(\omega_1)$ is constructed in this form:

$$\begin{cases} m_{ij}(\omega_1) = \mu_{T_j}(\omega_1) \\ \sum m(\cup_k \omega_k) = 1 - \mu_{T_j}(\omega_1). \end{cases} \tag{23}$$

We can obviously remark that:

$$T(\mu(A_i), \mu(A_j)) = min(Bel(A_i), Bel(A_j)) \tag{24}$$

where T is a minimum t-norm. Thus, the fuzzy support can be retrieved in an evidential database as follows:

$$Support_{\mathcal{FDB}}(X) = \frac{\sum_{T_j \in \mathcal{O}} min\{Bel(x_i), x_i \in X\}}{d}. \tag{25}$$

Interestingly enough, an equivalent to fuzzy database support in evidential database does exists.

Example 4. Table 4 shows how to create an evidential database from a fuzzy one.

Table 4: The evidential transformation of $\mathcal{FDB}$ (Table (a)) to $\mathcal{EDB}$ (Table (b))

	A		B	
	ω_1	ω_2	ω_1	ω_2
T_1	0.3	0.7	0.1	0.8
T_2	0.5	0.2	0.3	0.8
T_3	0.8	0.1	1.0	0.2

(a)

	A		B	
	ω_1	ω_2	ω_1	ω_2
T_1	$m_{11}(\omega_1) = 0.3$	$m_{21}(\omega_2) = 0.7$	$m_{31}(\omega_1) = 0.1$	$m_{41}(\omega_2) = 0.8$
	$m_{11}(\Omega) = 0.7$	$m_{21}(\Omega) = 0.3$	$m_{31}(\Omega) = 0.9$	$m_{41}(\Omega) = 0.2$
T_2	$m_{12}(\omega_1) = 0.5$	$m_{22}(\omega_2) = 0.2$	$m_{32}(\omega_1) = 0.3$	$m_{42}(\omega_2) = 0.8$
	$m_{12}(\Omega) = 0.5$	$m_{22}(\Omega) = 0.8$	$m_{32}(\Omega) = 0.7$	$m_{42}(\Omega) = 0.2$
T_3	$m_{11}(\omega_1) = 0.8$	$m_{21}(\omega_2) = 0.1$	$m_{31}(\omega_1) = 1.0$	$m_{41}(\omega_2) = 0.2$
	$m_{11}(\Omega) = 0.2$	$m_{21}(\Omega) = 0.9$	$m_{31}(\Omega) = 0$	$m_{41}(\Omega) = 0.8$

(b)

The equivalency of the support measure is shown for the itemset $B \times C$. The support of the itemset $A_{\omega_1} \times B_{\omega_2}$ from the Table 4.a is $Support(A_{\omega_1} \times B_{\omega_2}) = \frac{0.3+0.5+0.2}{3} = 1.0$. In the evidential database, Table 4.b, it is computed as follows:

$$
\begin{aligned}
&Support_{\mathcal{EDB}}(A_{\omega_1} \times B_{\omega_2}) = \tfrac{1}{3} \sum_{j=1}^{3} min(Bel(A_{\omega_1}), Bel(A_{\omega_2})) \\
&Support_{\mathcal{EDB}}(A_{\omega_1} \times B_{\omega_2}) = \tfrac{1}{3}(Bel_{T_1}(A_{\omega_1}) + Bel_{T_2}(A_{\omega_1}) + Bel_{T_2}(B_{\omega_2})) \\
&Support_{\mathcal{EDB}}(A_{\omega_1} \times B_{\omega_2}) = 1.0
\end{aligned}
$$

Despite the fact that the precise support is not equivalent to the fuzzy support, it is still possible to recover the same value with the use of the Equation (25).

6 Conclusion

In this paper, we detailed the data mining measures such as the support and the confidence on the several databases such as binary, probabilistic, fuzzy databases. We have proven the generalization relation between precise measures in evidential databases and measures used in other databases. In future works, we aim to study the evidential transformation of other imperfect databases such as fuzzy-possibilistic database [13].

References

1. Aggarwal, C.C.: Managing and Mining Uncertain Data. Springer Publishing Company, Incorporated (2009)

2. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules. In Proceedings of international conference on Very Large DataBases, VLDB, Santiago de Chile, Chile pp. 487–499 (1994)
3. Bach Tobji, M.A., Ben Yaghlane, B., Mellouli, K.: Incremental maintenance of frequent itemsets in evidential databases. In Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Verona, Italy pp. 457–468 (2009)
4. Chui, C.K., Kao, B., Hung, E.: Mining frequent itemsets from uncertain data. In: Zhou, Z.H., Li, H., Yang, Q. (eds.) Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, vol. 4426, pp. 47–58. Springer Berlin Heidelberg (2007)
5. Dempster, A.: Upper and lower probabilities induced by multivalued mapping. AMS-38 (1967)
6. Hewawasam, K.K.R., Premaratne, K., Shyu, M.L.: Rule mining and classification in a situation assessment application: A belief-theoretic approach for handling data imperfections. Trans. Sys. Man Cyber. Part B 37(6), 1446–1459 (2007)
7. Hong, T.P., Kuo, C.S., Wang, S.L.: A fuzzy AprioriTid mining algorithm with reduced computational time. Applied Soft Computing 5(1), 1–10 (2004)
8. Lee, S.: Imprecise and uncertain information in databases: an evidential approach. In Proceedings of Eighth International Conference on Data Engineering, Tempe, AZ pp. 614–621 (1992)
9. Liao, S.H., Chu, P.H., Hsiao, P.Y.: Data mining techniques and applications a decade review from 2000 to 2011. Expert Systems with Applications 39(12), 11303–11311 (2012)
10. Samet, A., Lefevre, E., Ben Yahia, S.: Mining frequent itemsets in evidential database. In Proceedings of the fifth International Conference on Knowledge and Systems Engeneering, Hanoi, Vietnam pp. 377–388 (2013)
11. Samet, A., Lefèvre, E., Ben Yahia, S.: Classification with evidential associative rules. In Proccedings of 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU, Montpellier, France , to appear (2014)
12. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
13. Weng, C., Chen, Y.: Mining fuzzy association rules from uncertain data, springer-verlag new york, inc. new york, ny, usa issn: 0219-1377 doi. knowledge and information systems 23, 129–152 (2010)

Annexe C

Bibliographie de l'auteur

Publication dans des revues internationales

A. Samet, E. Lefevre et S. Ben Yahia : Integration of extra-information for belief function theory conflict management problem through generic association rules. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 22(04) :531-551, 2014.

Publication dans des conférences internationales avec avec comité de lecture

A. Samet, Z. Ben Dhiaf, A. Hamouda et E. Lefevre : Classification of high-resolution remote sensing image by adapting the distancebelief function estimation model. *In Proceedings of International Conference on Communications, Computing and Control Applications, CCCA'2011, Hammamet, Tunisia*, pages 1-6, 2011.

A. Samet, I. Hammami, E. Lefevre et A. Hamouda : Generic discounting evaluation approach for urban image classification. *In Proceedings of 3rd international symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, IUKM'2013, Beijing, China*, pages 79-90, 2013.

A. Samet, E. Lefevre et S. Ben Yahia : Mining frequent itemsets in evidential database. *In Proceedings of the fifth International Conference on Knowledge and Systems Engeneering, Hanoi, Vietnam*, pages 377-388, 2013.

A. Samet, E. Lefevre et S. Ben Yahia : Reliability estimation with extrinsic and intrinsic measure in belief function theory. *In Proceedings of 5th International Conference on Modeling, Simulation and Applied Optimization, ICMSAO'13, Hammamet, Tunisia*, pages 1-6, 2013.

A. Samet, E. Lefèvre et S. Ben Yahia : Classification with evidential associative rules. *In Proceedings of 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Montpellier, France*, pages 25-35, 2014.

A. Samet, E. Lefevre et S. Ben Yahia : Evidential database : a new generalization of databases ? *In Proceedings of 3rd International Conference on Belief Functions, Belief 2014, Oxford, UK*, pages 105-114, 2014.

Bibliographie

[1] C.C. Aggarwal : *Managing and Mining Uncertain Data : 3, A.*, volume 3. Springer, 2010.

[2] R. Agrawal, T. Imieliński et A. Swami : Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, 1993.

[3] R. Agrawal et R. Srikant : Fast algorithm for mining association rules. *In Proceedings of international conference on Very Large DataBases, VLDB, Santiago de Chile, Chile*, pages 487–499, 1994.

[4] A. Appriou : Probabilités et incertitude en fusion de données multisenseurs. *Revue Scientifique et Technique de la Défense*, 11:27–40, 1991.

[5] A. Appriou : Multisensor signal processing in the framework of the theory of evidence. *Application of Mathematical Signal Processing Techniques to Mission Systems*, pages 5–1, 1999.

[6] M. A Bach Tobji, B. Ben Yaghlane et K. Mellouli : Frequent itemset mining from databases including one evidential attribute. *In Proceedings of second international conference on Scalable Uncertainty Management, Napoli, Italy*, 5291:19–32, 2008.

[7] M. A Bach Tobji, B. Ben Yaghlane et K. Mellouli : Incremental maintenance of frequent itemsets in evidential databases. *In Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Verona, Italy*, pages 457–468, 2009.

[8] Y. Bastide, N. Pasquier, R. Taouil, L. Lakhal et G. Stumme : Mining minimal non-redundant association rules using frequent closed itemsets. *In Proceedings of the International Conference DOOD2000, LNAI Springer-Verlag, London, UK*, volume 1861:972–986, 2000.

[9] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme et L. Lakhal : Pascal : un algorithme d'extraction des motifs fréquents. *Techniques et Sciences Informatiques*, 21(1):65–95, 2002.

[10] R.J. Bayardo : Efficiently mining long patterns from databases. *In Proceedings of 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD'98), Seattle*, pages 85–93, June 1998.

[11] S. Ben Yahia et E. Mephu Nguifo : Approches d'extraction de règles d'association basées sur la correspondance de galois. *Ingénierie des Systèmes d'Information*, 9(3-4):23–55, 2004.

[12] S. BenYahia et E. Mephu Nguifo : A survey of Galois connection semantics-based approaches for mining association rules. Technical report, IUT-Lens, Centre de Recherche en Informatique de Lens (CRIL), Lens, France, January 2004.

[13] J.C. Bezdek : *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[14] P.P. Bonissone et R.M. Tong : Reasoning with uncertainty in expert systems. *International journal of man-machine studies*, 22(3):241–250, 1985.

[15] I. Bouzouita, S. Elloumi et S. Ben Yahia : GARC : A new associative classification approach. *in Proceedings of 8th International Conference on Data Warehousing and Knowledge Discovery, DaWaK, Krakow, Poland*, pages 554–565, 2006.

[16] B. Bringmann, S. Nijssen et A. Zimmermann : Pattern-based classification : A unifying perspective. *CoRR*, abs/1111.6191, 2011.

[17] YL. Chen et CH. Weng : Mining association rules from imprecise ordinal data. *Fuzzy Set Syst*, 159(4):460–474, 2008.

[18] L. Cholvy : Applying theory of evidence in multisensor data fusion : a logical interpretation. *In Proceedings of the Third International Conference on Information Fusion, 2000, FUSION 2000, Paris, France*, 1:17–24, July 2000.

[19] C-K Chui, B. Kao et E. Hung : Mining frequent itemsets from uncertain data. *in Proceedings of the 11th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Nanjing, China*, pages 47–58, 2007.

[20] K-L Chung : *A course in probability theory.* Academic press, 2001.

[21] R. T. Cox : Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, 14:1–14, 1946.

[22] F. Cuzzolin : A geometric approach to the theory of evidence. *IEEE transaction, Man, and Cybernetics-Part C : Application and reviews*, 38(4):522–534, 2008.

[23] M. Daniel : Conflicts within and between belief functions. *In International Conference on Information Processing and Management of Uncertainty, IPMU'10*, pages 696–705, 2010.

[24] B.A Davey et H.A. Priestley : *Introduction to Lattices and Order.* Cambridge University Press, 2002.

[25] A.P. Dempster : *Upper and lower probabilities induced by multivalued mapping.* AMS-38, 1967.

[26] A.P. Dempster : A generalization of bayesian inference. *In* R.R. Yager et L. Liu, éditeurs : *Classic Works of the Dempster-Shafer Theory of Belief Functions*, volume 219 de *Studies in Fuzziness and Soft Computing*, pages 73–104. Springer Berlin Heidelberg, 2008.

[27] Y. Deng, W.K. Shi, Z.F. Zhu et Q. Liu : Combining belief functions based on distance of evidence. *Decision Support Systems*, 38(3):489–493, 2004.

[28] T. Denœux : K-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(5):804–813, 1995.

[29] T. Denœux : Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recogn.*, 30(7):1095–1107, juillet 1997.

[30] P. Diaconis : Review of glenn shafer's "a mathematical theory of evidence". *Journal of the American Statistical Association*, 73(363):677–678, 1978.

[31] Y. Djouadi, S. Redaoui et K. Amroun : Mining association rules under imprecision and vagueness : towards a possibilistic approach. *in Proceedings of IEEE International Fuzzy Systems Conference, FUZZ-IEEE 2007, Imperial College, London, UK*, pages 1–6, 2007.

[32] D. Dubois et H. Prade : On the unicity of Dempster rule of combination. *International Journal of Intelligent Systems*, 1(2):133–142, 1986.

[33] D. Dubois et H. Prade : *Possibility Theory : An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, 1988.

[34] D. Dubois et H. Prade : Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4:244–264, 1988.

[35] D. Dubois et H. Prade : The three semantics of fuzzy sets. *Fuzzy Sets and Systems*, (90):141–150, 1997.

[36] Z. Elouedi, E. Lefevre et D. Mercier : Discounting of a belief function using a confusion matrix. *in Proceedings of 22th IEEE International Conference on Tools with Artificial Intelligence, ICTAI'2010, Arras, France*, pages 287–294, 2010.

[37] A. Ennaceur, Z. Elouedi et E. Lefevre : Reasoning under uncertainty in the ahp method using the belief function theory. *In Proceedings of 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2012, Catania, Italy*, 4:373–382, 2012.

[38] R. Fagin et J.Y. Halpern : A new approach to updating beliefs. *Uncertainty in Artificial Intelligence*, pages 347–374, 1991.

[39] E. Fix et J.L. Hodges : Discriminatory analysis, nonparametric discrimination : Consistency properties. Rapport technique, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[40] A. Frank et A. Asuncion : UCI machine learning repository (2010). *URL : http ://archive.ics.uci.edu/ml.*

[41] B. Ganter et R. Wille : *Formal Concept Analysis*. Springer-Verlag, Heidelberg, 1999.

[42] G. Gasmi, S. Ben Yahia, E. Mephu Nguifo et Y. Slimani : Igb : A new informative generic base of association rules. *Proceedings of the Intl. Ninth Pacific-Asia Conference on Knowledge Data Discovery (PAKDD05), LNAI 3518, Hanoi, Vietnam, Springler-Verlag*, pages 81–90, 2005.

[43] T. George et N.R. Pal : Quantification of conflict in Dempster-Shafer framework : A New Approach. *International Journal of General Systems*, 24(4):407–423, 1996.

[44] H. Guo, W. Shi et Y. Deng : Evaluating sensor reliability in classification problems based on evidencetheory. *IEEE transactions on Systems, Man, and Cybernetics, Part B*, 36(5):970–981, 2006.

[45] J. Han, J. Pei et Y. Yin : Mining frequent patterns without candidate generation. *In Proceedings of the ACM-SIGMOD Intl. Conference on Management of Data (SIGMOD'00), Dallas, Texas*, pages 1–12, May 2000.

[46] K.K. Rohitha Hewawasam, K. Premaratne et M-L Shyu : Rule mining and classification in a situation assessment application : A belief-theoretic approach for handling data imperfections. *Trans. Sys. Man Cyber. Part B*, 37(6):1446–1459, 2007.

[47] K.K. Rohitha Hewawasam, K. Premaratne, M-L Shyu et S.P. Subasingha : Rule mining and classification in the presence of feature level and class label ambiguities. *In SPIE 5803, Intelligent Computing : Theory and Applications III, 98*, 2005.

[48] T-P. Hong, C-S. Kuo et S-L. Wang : A fuzzy AprioriTid mining algorithm with reduced computational time. *Applied Soft Computing*, 5(1):1–10, 2004.

[49] T. Inagaki : Interdependence between safety-control policy and multiple-sensor schemes via Dempster-Shafer theory. *IEEE Transaction on reliability*, 40(2):182–188, 1991.

[50] S. Jabbour, M. Khiari, L. Sais, Y. Salhi et K. Tabia : Symmetry-based pruning in itemset mining. *in Proceedings of IEEE 25th International Conference on Tools with Artificial Intelligence, ICTAI'13, Washington DC, USA*, pages 483–490, 2013.

[51] J. Jacod et P. E. Protter : *Probability essentials*. Springer, 2003.

[52] A.-L. Jousselme, D. Grenier et E. Bossé : A new distance between two bodies of evidence. *Information Fusion*, 2:91–101, 2001.

[53] A-L. Jousselme et P. Maupin : Distance in evidence theory : Comprehensive survey and generalizations. *International Journal of Approximate Reasoning*, 53(2):118–145, 2012.

[54] F. Klawonn et E. Schwecke : On the axiomatic justification of Dempster's rule of combination. *International Journal of Intelligent Systems*, 7(5):469–478, 1992.

[55] F. Klawonn et P. Smets : The dynamic of belief in the transferable belief model and specialization-generalization matrices. *in Proceedings of the Eighth international conference on Uncertainty in artificial intelligence*, pages 130–137, 1992.

[56] J. Klein et O. Colot : Automatic discounting rate computation using a dissent criterion. *In Workshop on the Theory of Belief Functions*, 2010.

[57] J. Klein et O. Colot : Singular sources mining using evidential conflict analysis. *International Journal of Approximate Reasoning*, 52(9):1433–1451, 2011.

[58] J. Kohlas et P-A. Monney : *A Mathematical Theory of Hints. An Approach to the Dempster-Shafer Theory of Evidence*, volume 425. Lecture Notes in Economics and Mathematical Systems, 1995.

[59] M. Kryszkiewicz : Concise representations of association rules. *in Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery, London, UK*, pages 92–109, 2002.

[60] S.K. Lee : An extended relational database model for uncertain and imprecise information. *In Proceedings of the 18th International Conference on Very Large Data Bases*, pages 211–220, 1992.

[61] S.K. Lee : Imprecise and uncertain information in databases : an evidential approach. *In Proceedings of Eighth International Conference on Data Engineering, Tempe, AZ*, pages 614–621, 1992.

[62] W. Lee et S. J. Stolfo : Data mining approaches for intrusion detection. *in Proceedings of the 7th Conference on USENIX Security Symposium, San Antonio, Texas, USA*, 7:1–6, 1998.

[63] E. Lefevre : *Fusion Adaptée D'informations Conflictuelles Dans Le Cadre de la Théorie de L'évidence*. Thèse de doctorat, Institut National des Sciences Appliquées de Rouen, 2001.

[64] E. Lefevre, O. Colot et P. Vannoorenberghe : Belief functions combination and conflict management. *Information Fusion*, 3(2):149–162, 2002.

[65] C.K.-S. Leung, C.L. Carmichael et B. Hao : Efficient mining of frequent patterns from uncertain data. *In Proceedings of Seventh IEEE International Conference on Data Mining Workshops, ICDM Workshops, Omaha, Nebraska, USA*, pages 489–494, Oct 2007.

[66] W. Li, J. Han et J. Pei : CMAR : accurate and efficient classification based on multiple class-association rules. *In Proceedings of IEEE International Conference on Data Mining, San Jose, California, USA*, pages 369–376, 2001.

[67] D. Lin et Z.M. Kedem : Pincer-Search : A new algorithm for discovering the maximum frequent sets. *LNCS vol 1377*, pages 105–119, March 1998.

[68] B. Liu, W. Hsu et Y. Ma : Integrating classification and association rule mining. *in Proceedings of the Fourth International Conference onKnowledge Discovery and Data Mining, New York City, New York, USA*, pages 80–86, 1998.

[69] R. Manjusha et R. Ramachandran : Web mining framework for security in e-commerce. *In Proceedings of International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, India*, pages 1043 –1048, 2011.

[70] A. Martin, A.-L. Jousselme et C. Osswald : Conflict measure for the discounting operation on belief functions. *In International Conference on Information Fusion, Brest, France*, pages 1003–1010, 2008.

[71] F. Masseglia : Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel. *HDR, Université de Versailles St-Quentin en Yvelines*, 2002.

[72] M-H. Masson : *Apports de la théorie des possibilités et des fonctions de croyance à lanalyse de données imprécises.* Thèse de doctorat, Université de Technologie de Compiègne, 2005.

[73] M-H Masson et T. Denœux : ECM : An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384–1397, 2008.

[74] D. Mercier, T. Denoeux et M.-H. Masson : *Belief function correction mechanisms*, volume 249, chapitre Studies in Fuzziness and Soft Computing, pages 203–222. january 2010.

[75] D. Mercier, B. Quost et T. Denœux : Contextual discounting of belief functions. *8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Proceedings of ECSQARU, Barcelona, Spain*, pages 552–562, 2005.

[76] D. Mercier, B. Quost et T. Denœux : Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion*, 9(2):246–258, 2008.

[77] C. Ordonez et E. Omiecinski : Discovering association rules based on image content. *In Proceedings of the IEEE Advances in Digital Libraries Conference (ADL'99), Baltimore, MD*, pages 38–49, 1999.

[78] N. Pasquier : *Data Mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données.* Doctorat d'université, Université de Clermont-Ferrand II, France, 2000.

[79] N. Pasquier, Y. Bastide, R.Taouil et L. Lakhal : Efficient mining of association rules using closed itemset lattices. *Journal of Information Systems*, 24:25–46, 1999.

[80] N. Pasquier, Y. Bastide, R. Touil et L. Lakhal : Pruning closed itemset lattices for association rules. *in Proceedings of 14th International Conference Bases de Données Avancées, Hammamet, Tunisia*, pages 177–196, 1998.

[81] J. Pearl : *Probabilistic reasoning in intelligent systems : networks of plausible inference.* Morgan Kaufmann, 1988.

[82] J. Pei, J. Han et R. Mao : CLOSET : An efficient algorithm for mining frequent closed itemsets. *ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, pages 21–30, 2000.

[83] A. Salleb : *Recherche de motifs fréquents pour l'extraction de règles d'association et de caractérisation.* 2003.

[84] A. Samet, Z. Ben Dhiaf, A. Hamouda et E. Lefevre : Classification of high-resolution remote sensing image by adapting the distancebelief function estimation model. *In Proceedings of International Conference on Communications, Computing and Control Applications, CCCA'2011, Hammamet, Tunisia*, pages 1–6, 2011.

[85] A. Samet, I. Hammami, E. Lefevre et A. Hamouda : Generic discounting evaluation approach for urban image classification. *In Proceedings of 3rd in-*

ternational symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, IUKM'2013, Beijing, China, pages 79–90, 2013.

[86] A. Samet, E. Lefevre et S. Ben Yahia : Mining frequent itemsets in evidential database. *In Proceedings of the fifth International Conference on Knowledge and Systems Engeneering, Hanoi, Vietnam*, pages 377–388, 2013.

[87] A Samet, E. Lefevre et S. Ben Yahia : Reliability estimation with extrinsic and intrinsic measure in belief function theory. *In Proceedings of 5th International Conference on Modeling, Simulation and Applied Optimization, ICMSAO'13, Hammamet, Tunisia*, pages 1–6, 2013.

[88] A. Samet, E. Lefevre et S. Ben Yahia : Classification with evidential associative rules. *In Proceedings of 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Montpellier, France*, pages 25–35, 2014.

[89] A. Samet, E. Lefevre et S. Ben Yahia : Evidential database : a new generalization of databases ? *In Proceedings of 3rd International Conference on Belief Functions, Belief 2014, Oxford, UK*, pages 105–114, 2014.

[90] A. Samet, E. Lefevre et S. Ben Yahia : Integration of extra-information for belief function theory conflict management problem through generic association rules. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 22(04):531–551, 2014.

[91] A. Savasere, E. Omiecinski et S. B. Navathe : An efficient algorithm for mining association rules in large databases. *in Proceedings of the 21th International Conference on Very Large Data Bases, Zurich, Switzerland*, pages 432–444, 1995.

[92] K. Sentz et S. Ferson : Combination of evidence in Dempster-Shafer theory. Rapport technique, 2002.

[93] G. Shafer : *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[94] M-L. Shyu, C. Haruechaiyasak, S-C. Chen et K. Premaratne : Mining association rules with uncertain item relationships. *In Proceedings of 6th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2002), Orlando, Florida, USA, July 14-18*, pages 435–440, 2002.

[95] F. Smarandache, A. Martin et C. Osswald : Contradiction measures and specificity degrees of basic belief assignments. *In International Conference on Information Fusion*, pages 1–8, 2011.

[96] P. Smets : Belief functions. *in Non Standard Logics for Automated Reasoning , P. Smets, A. Mamdani, D. Dubois, and H. Prade, Eds. London,U.K : Academic*, pages 253–286, 1988.

[97] P. Smets : The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.

[98] P. Smets : The Transferable Belief Model and other interpretations of Dempster-Shafer's Model. *In Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI'90, MIT, Cambridge, MA*, pages 375–383, 1990.

[99] P. Smets : Beliefs functions : The Disjunctive Rule of Combination and the Generalized Bayesian Theorem. *International Journal of Approximate Reasoning*, 9(1):1–35, 1993.

[100] P. Smets : Quantifying beliefs by belief functions : An axiomatic justification. *in Proceedings of the 13th International Joint Conference on Artifical Intelligence - Volume 1, IJCAI'93, Chambery, France*, pages 598–603, 1993.

[101] P. Smets : What is Dempster-Shafers model. *Advances in the Dempster-Shafer theory of evidence*, pages 5–34, 1994.

[102] P. Smets : Imperfect information : Imprecision and uncertainty. *In Uncertainty Management in Information Systems*, pages 225–254. Springer, 1997.

[103] P. Smets : Decision making in the TBM : The necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38:133–147, 2005.

[104] P. Smets : Analyzing the combination of conflicting belief functions. *Information Fusion*, 8(4):387–412, 2007.

[105] P. Smets et R. Kennes : The Transferable Belief Model. *Artificial Intelligence*, 66(2):191–234, 1994.

[106] P. Vannoorenberghe : Un état de lart sur les fonctions de croyance appliquées au traitement de linformation. *Revue I3*, 3(1):9–45, 2003.

[107] P. Walley : *Statistical reasoning with imprecise probabilities*. Chapman and Hall London, 1991.

[108] C-H Weng et Y-L Chen : Mining fuzzy association rules from uncertain data. *Knowledge and Information Systems*, 23(2):129–152, 2010.

[109] R.R. Yager : On the Dempster-Shafer framework and new combination rule. *Information Sciences*, 41:93–138., 1987.

[110] R.R. Yager : On considerations of credibility of evidence. *International Journal of Approximate Reasoning*, 7:45–72, 1992.

[111] X. Yin et J. Han : Cpar : Classification based on predictive association rules. *In Proceedings of 2003 SIAM International Conference on Data Mining (SDM03), San Francisco, CA*, 2003.

[112] L.A. Zadeh : Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[113] L.A. Zadeh : Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, (1):3–28, 1978.

[114] L.A. Zadeh : Review of Shafer's mathematical theory of evidence. *AI Magazine 5*, pages 81–83, 1984.

[115] L.A. Zadeh : A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *The AI Magazine*, 7, no. 2:85–90, 1994.

[116] M. Zaffalon et G. Cooman : Editorial : Imprecise probability perspectives on artificial intelligence. *Annals of Mathematics and Artificial Intelligence*, 45(1-2):1–4, 2005.

[117] M.J. Zaki : Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390, May 2000.

[118] M.J. Zaki et C-J. Hsiao : CHARM : An efficient algorithm for closed itemset mining. *In SDM*, volume 2, pages 457–473. SIAM, 2002.

[119] L.M. Zouhal et T. Denoeux : An evidence-theoretic K-NN rule with parameter optimization. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, 28(2):263–271, 1998.

Théorie des fonctions de croyance : application des outils de data mining pour le traitement de données imparfaites

Notre travail s'inscrit dans l'intersection de deux disciplines qui sont la théorie des fonctions de croyance (TFC) et la fouille de données. L'interaction pouvant exister entre la TFC et la fouille de données est étudiée sous deux volets. La première interaction souligne l'apport des règles associatives génériques au sein de la TFC. Nous nous sommes intéressés au problème de fusion de sources non fiables dont la principale conséquence est l'apparition de conflit lors de la combinaison. Une approche de gestion de conflit reposant sur les règles d'association génériques appelée ACM a été proposée.

La deuxième interaction s'intéresse aux bases de données imparfaites en particulier les bases de données évidentielles. Les informations, représentées par des fonctions de masse, sont étudiées afin d'extraire des connaissances cachées par le biais des outils de fouille de données. L'extraction des informations pertinentes et cachées de la base se fait grâce à la redéfinition de la mesure du support et de la confiance. Ces mesures introduites ont été les fondements d'un nouveau classifieur associatif que nous avons appelé EDMA.

Mots-clés : fouille de données, Théorie des fonctions de croyance, Gestion de conflit, Classification associative.

Belief function theory : application of data mining tools for imperfect data treatment

This thesis explores the relation between two domains which are the belief function theory BFT and data mining. Two main interactions between those domains have been pointed out. The first interaction studies the contribution of the generic associative rules in the BFT. We were interested in managing conflict in case of fusing conflictual information sources. A new approach for conflict management based on generic association rules has been proposed called ACM. The second interaction studies imperfect databases such as evidential databases. Those kind of databases where information is represented by belief functions, are studied in order to extract hidden knowledge using data mining tools. The extraction of those knowledges was possible thanks to a new definition of the support and the confidence measures. Those measures were integrated into a new evidential associative classifier called EDMA.

Key-words : Data mining, Belief function theory, conflict management, Associative classification.
