Credal classification of uncertain data based on belief function theory

Zhun-Ga Liu

▶ **To cite this version:**

Zhun-Ga Liu. Credal classification of uncertain data based on belief function theory. Signal and Image processing. Télécom Bretagne; Université de Bretagne Occidentale, 2014. English. NNT : . tel-01212836

HAL Id: tel-01212836
https://hal.science/tel-01212836

Submitted on 7 Oct 2015

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

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

THÈSE / Télécom Bretagne
sous le sceau de l'Université européenne de Bretagne

pour obtenir le grade de Docteur de Télécom Bretagne
En accréditation conjointe avec l'Ecole Doctorale Sicma et en cotutelle avec Northwestern Polytechnical University (Chine)
Mention : Sciences et Technologies de l'Information et de la Communication

présentée par
Zhunga LIU
préparée dans le département Image et Traitement de l'Information
Laboratoire Labsticc/CID

Credal classification of uncertain data based on belief function theory

Thèse soutenue le 24 novembre 2014
Devant le jury composé de :

Yongmei Cheng
Professeur, Northwestern Polytechnical University – Chine / présidente

Arnaud Martin
Professeur, Irisa/Université de Rennes 1 / rapporteur

Hongbing Ji
Professeur, Xidian University – Chine / rapporteur

Jean Dezert
Senior research scientist, Onera - Palaiseau / examinateur

Quan Pan
Professeur, Northwestern Polytechnical University - Chine / co-directeur de thèse

Grégoire Mercier
Professeur, Télécom Bretagne / directeur de thèse

N° d'ordre : 2014 telb0333

Sous le sceau de l'Université européenne de Bretagne

Télécom Bretagne

En accréditation conjointe avec l'Ecole Doctorale SICMA

Co-tutelle avec Northwestern Polytechnical University

Credal classification of uncertain data based on belief function theory

---

Thèse de Doctorat

Mention : "Sciences et Technologies de l'Information et de la Communication"

Présentée par **Zhunga Liu**

Département : Image et Traitement de l'Information

Laboratoire : LabSTICC Pôle : CID

Directeur de thèse : Grégoire Mercier Quan Pan

Soutenue le 24 novembre 2014

Jury :

Présidente :	Prof.Yongmei Cheng, Northwestern Polytechnical University, China
Rapporteurs :	Prof. Arnaud Martin, Université de Rennes I, France
	Prof. Hongbing Ji, Xidian University, China
Examinateurs :	Dr. Jean Dezert, The French Aerospace Lab, France (co-supervisor)
	Prof. Grégoire Mercier, Telecom Bretagne, France
	Prof. Quan Pan, Northwestern Polytechnical University, China

Acknowledgement

This PhD thesis is co-supervised by Prof. Grégoire Mercier from Telecom Bretagne, Prof. Quan Pan from Northwestern Polytechnical University and Dr. Jean Dezert form the French Aerospace Lab (ONERA). I am very grateful to them for their important supervision and support during my Ph.D study.

I have been working with Prof. Grégoire Mercier since 2010, and I am deeply grateful to him for offering me an opportunity to study on the topic of belief functions and its application in Telecom Bretagne. Prof. Mercier has keen sense of technology developments, as well as wide knowledge in the field of image and information processing. I often discussed with him on the research work, and I always got important inspirations from his constructive comments, which is very useful for the improvement of my research work. I would like to sincerely thank him for his support, encouragement and advice over the period of my PhD studies, as well as the assistance with the preparation of this thesis.

I am very grateful to Prof. Quan Pan, who opens the door leading me to the interesting research work on belief functions. He has wide knowledge base on the information fusion and pattern recognition, and he has given me a lot of directions that I carefully followed during the Ph.D study. He has provided me great support for the development of research work, such as overseas study, attending of national and international conferences, and so on. His continuous support is also very important for my life. He always encouraged me to deal with the difficult problem encountered in the work, and go further and further on the way of research. His supervision is very useful and important for me to complete this thesis.

I would like to thank Dr. Jean Dezert very much. He is a very well known scientist in the field of information fusion, particularly for his current research topic of belief functions. I have met Dr. Dezert in 2009 during his first series of seminars in China, and since then I have been studying the belief function theory, and DSmT (Dezert-Smarandache Theory) as well thanks to him. Dr. Dezert is a very nice and hard-working scientist, and he usually returned his important comments on the work in a short time. He has influenced me a lot with his passion towards work. I have learned a great deal from tutorage of him specially in paper writing and publishing, and for developing a rigorous way of research.

I would like to thank my colleagues and friends, Dr. Yong Liu, Dr. Xiaoxu Wang, Dr. Kuang Zhou and Dr. Lianmeng Jiao for their interest in my study of belief functions and their great help on my thesis work. I am indebted to many other friends for their help.

I want to express my endless thanks to my parents, my wife and daughter for their long-standing support and care. Their love and encouragement give me confidence to complete the Ph.D study and go ahead on the way of research.

I am also very grateful to Prof. Arnaud Martin and Prof. Hongbing Ji for the constructive comments on my Ph.D thesis.

Abstract

How to well model and manage the uncertainty in the classification problem remains an important and interesting topic of research. In the classification of uncertain data, the available attribute information can be insufficient for the specific classification of object (pattern, sample), since several different classes may appear indistinguishable according to the used attribute data. In this case, it is hard to correctly commit one object into a particular class. Evidence theory also called belief function theory is appealing for dealing with such uncertain and imprecise information thanks to the belief functions. Credal classification of uncertain data based on belief function theory has been studied in this thesis, and it allows the object to belong not only to the single classes, but also to any set of classes (called meta-class) with different masses of belief. The credal classification is interesting to explore the imprecision of class, and it can also provide a deeper insight in the data structure. The object that is difficult to correctly classify can be reasonably assigned with a degree of belief to the proper meta-class defined by the disjunction of several single classes that the object is very likely to belong to, and this can also reduce errors.

The classification methods can be mainly identified by supervised, unsupervised and semi-supervised ones according to the availability of training information, and we focus on the supervised and unsupervised classifications in this thesis. When there are a lot of training samples available in the classification, two credal classifiers for uncertain data are proposed for dealing with different cases. Moreover, the missing attribute data is often encountered in classification problem. The different estimations of the missing values can lead to distinct classification results sometimes, and this yields high imprecision and uncertainty of classification due to the lack of information in the missing values. It is worth noting that the inherent nature of uncertainty in classification of the incomplete data and the uncertain (complete) data is the same, since both of them are caused by the insufficient knowledge (attribute information). So one credal classification method for classification of the incomplete data with missing values has been developed based on belief function theory to well characterize the uncertainty and imprecision. If the training information is not available, the data clustering (unsupervised) analysis can be applied, and the belief-structure-based fuzzy c-means clustering method has been proposed. The main content of this thesis are briefly introduced as follows.

A belief $c \times K$ neighbors (BCKN) classifier has been proposed based on belief function theory. In BCKN, the query object is classified according to its K nearest neighbors in each class, and $c \times K$ neighbors are involved in BCKN approach (c being the number of classes). $c \times K$ basic belief assignments (BBA's) are determined according to the distances between the object and these neighbors, and the global fusion of them is used for the credal classification of object. It allows to commit, with different masses of belief, an object not only to a specific class, but also to a set of classes (called meta-class), or eventually to the ignorant class characterizing the outlier. The objects that lie in the overlapping zone of different classes cannot be reasonably committed to a particular class, and that is why such objects will be assigned to the associated meta-class defined by the union of these different classes. Such approach allows to reduce the misclassification errors at the price of the detriment of the overall classification precision, which is usually preferable in some applications. The objects too far from the others will be naturally considered as outliers. The results of several experiments are given and analyzed to illustrate the potential of BCKN approach.

BCKN is able to deal with the general and complicate cases but with the big computation burden. When each class of data can be represented by the prototype vector, a simple credal

classification rule (CCR) has been developed using belief functions. Each specific class is characterized by a class center (i.e. prototype), and consists of all the objects that are sufficiently close to the center. The belief of the assignment of a given object to classify with a specific class is determined from the Mahalanobis distance between the object and the center of the corresponding class. The meta-classes are used to capture the imprecision in the classification of the objects when they are difficult to correctly classify because of the poor quality of available attributes. The selection of meta-classes depends on the application and the context, and a measure of the degree of indistinguishability between classes is introduced for the determination of mass of belief on meta-class. In CCR, the objects assigned to a meta-class should be close to the center of this meta-class having similar distances to all the involved specific classes' centers, and the objects too far from the others will be considered as outliers (noise). CCR provides robust credal classification results with a relatively low computational burden. Several experiments using both artificial and real data sets are presented to evaluate and compare the performances of this CCR method with respect to other classification methods.

It often happens that partial attribute values are missing in some applications. The missing values can bring high uncertainty in the classification, because the object (incomplete pattern) with different possible estimations of missing values may yield distinct classification results. A new prototype-based credal classification (PCC) method is proposed to deal with incomplete patterns under belief function framework. The class prototypes obtained by training samples are respectively used to estimate the missing values. Typically, in a c-class problem, one has to deal with c prototypes, which yield c estimations of the missing values. The different edited patterns based on each possible estimation are then classified by a standard classifier and we can get at most c distinct classification results with different weighting factors depending on the distances between the object and the corresponding prototypes. Because all these distinct classification results are potentially admissible, we propose to combine them altogether to obtain the final classification of the incomplete pattern. These classification results should be discounted using their weights before the fusion process. A new credal combination method is introduced for solving the classification problem, and it is able to characterize the inherent uncertainty due to the possible conflicting results delivered by different estimations of the missing values. The incomplete patterns that are very difficult to classify in a specific class will be reasonably and automatically committed to some proper meta-classes by PCC method in order to reduce errors. The effectiveness of PCC method has been tested through several experiments.

When the training samples are unavailable in the classification problem, a new credal c-means (CCM) clustering method working with credal partition is proposed to effectively deal with the uncertain data. In CCM, the object committed to one singleton cluster should be very close to the center of this cluster. One object can be simultaneously close to several clusters, and it is hard to be correctly classified into a particular cluster, since these several clusters seem not very distinguishable for this object. In such case, it will be cautiously committed to meta-cluster by CCM, which can well characterize the imprecision of the class of the object and can also reduce the misclassification errors. If one object is too far from all the clusters with respect to the given threshold, it will be naturally considered as outlier. So CCM is robust to the noisy data. The objective function of CCM is designed based this basic principle, and the clustering centers and the mass of belief for any object can be obtained by the linear optimization (minimization) of the proposed objective function. One tuning threshold for selecting of meta-cluster is introduced in the objective function to control the number of meta-class with big cardinality, and this is able to efficiently reduce the computation complexity. The credal partition can be simply reduced to fuzzy partition if necessary. The experimental evaluation over synthetic and real data demonstrates the effectiveness of CCM.

Acronyms

Artificial neural network	ANN
Basic belief assignment	BBA
Belief $C \times K$ neighbor	BCKN
Credal c-means	CCM
Classification and regression tree	CART
Credal classification rule	CCR
Dempster-Shafer Theory	DST
Dezert-Smarandache Theory	DSmT
Evidential c-means	ECM
Evidential K-nearest neighbor	EK-NN
Evidential neural network	ENN
Frame of discernment	FoD
Fuzzy c-means	FCM
K-nearest neighbor	K-NN
Prototype-based credal classification	PCC
Support vector machine	SVM
Transferable belief model	TBM

Contents

List of Tables

List of Figures

PREMIÈRE PARTIE : FRENCH ABSTRACT

TELECOM Bretagne

Résumé étendu

L'incertitude est une notion très importante mais difficile à intégrer proprement dans un processus de classification en raison du caractère aléatoire des données, de la connaissance insuffisante portée sur elles, voire de l'absence de certaines d'entre elles... Pour la classification des données incertaines, différentes classes peuvent être partiellement superposées selon l'utilisation (disponible) d'attributs, et les échantillons situés dans la zone de chevauchement sont assez difficiles à classer correctement, puisque ces classes associées apparaissent indiscernables. Les méthodes de classification se fondent la plupart du temps sur un cadre probabiliste, et les échantillons sont couramment affectés à la classe ayant une probabilité maximale. Néanmoins, le cadre probabiliste capte seulement l'aspect aléatoire des données. La théorie des fonctions de croyance également connue comme théorie de l'évidence ou théorie de Dempster-Shafer (DST) considérée comme la généralisation de la théorie des probabilités de Bayes permet de définir des fonctions de masse de croyance non seulement à des éléments uniques (correspondant à des classes simples), mais aussi à un ensemble d'éléments à l'aide d'une affectation des masses de croyances (BBA). La théorie des fonctions de croyance est un outil efficace pour modéliser et gérer l'information incertaine et imprécise, et elle a déjà été appliquée dans de nombreux domaines, tels que la classification des données, la segmentation et l'aide à la décision...

Les méthodes de classification peuvent être soit supervisées ou non supervisées. Lorsque des échantillons labellisés sont disponibles en nombre suffisant, une classification supervisée peut être appliquée, et le modèle de règle de décision est déterminé sur la base des données d'entraînement. Certains classifieurs ont été développés sur la base de la DST, et l'ignorance totale est caractérisée en utilisant une pondération des fonctions de masse. Néanmoins, l'information imprécise partielle n'est pas prise en compte dans ces méthodes. Dans les applications réelles, la classification est souvent partiellement (plutôt que totalement) imprécise entre un très petit nombre (par exemple, 2) de classes. Ainsi, la caractérisation appropriée de l'imprécision partielle est très importante. Dans cette thèse, nous avons étudié la classification crédibiliste de données incertaines sur la base des fonctions de croyance, et deux classificateurs crédibilistes ont été proposés pour faire face à différentes situations. La classification crédibiliste permet à un objet d'appartenir à des classes simples mais aussi à des méta-classes définies par l'union de plusieurs classes simples. Ces méta-classes sont introduites pour modéliser l'imprécision partielle de classification et pour réduire le taux d'erreur. Une méthode de classification crédibiliste appelée $c \times K$ plus proches voisins crédibilistes (*Belief* $c \times K$ *nearest neighbors*) a été introduite. Lorsque chaque classe peut être représentée par son centre de classe, nous avons également proposé une règle simple de classification crédibiliste (CCR), qui peut calculer directement la masse de croyance de l'échantillon appartenant à chaque classe et une méta-classe avec une faible complexité calculatoire.

En outre, la classification avec données manquantes est une problématique récurrente dans de nombreuses applications, et elle reste un sujet d'intérêt. Les méthodes classiques caractérisent généralement les données manquantes par une classe particulière de valeurs de probabilité maximale. Cependant, les différentes estimations des valeurs manquantes peuvent conduire à des résultats de classification distincts, et parfois avec une forte imprécision, et incertitude dans la prise de décision. Il est à noter que nous considérons que la classification des données incomplètes et celle des données incertaines (mais complètes) représentent deux problèmes de même nature, puisque les deux sont causées par une connaissance insuffisante sur les données. Une méthode de classification crédibiliste de données incomplètes a été également développée, et elle est capable de modéliser de telles informations incertaines et imprécises provenant de valeurs manquantes.

S'il n'y a pas de phase d'entraînement, l'analyse de la concentration de données aussi appelée classification non supervisée peut être utilisé, et les motifs seront automatiquement groupés en plusieurs groupes selon certaines mesures de (dis-)similarité. Les K-moyennes floues (*Fuzzy C-means*, FCM) est une méthode de classification très répandue, et une version crédibiliste de la FCM, K-Moyennes évidentielles (*Evidential C-Means*, ECM) a été introduite permettant de prendre en compte des fonctions de croyance. Néanmoins, lorsque les différents centres de classes (dans ce cas, les classes simples et les méta-classes) sont proches, ECM va produire des résultats très surprenants. Ainsi, nous proposons une nouvelle méthode de classification crédibiliste appelé K-moyennes floues et crédibiliste (*Credal C means*, CCM) comme une extension crédibiliste de la FCM permettant de contourner les limites de l'ECM.

Ces quatre méthodes, proposées dans cette thèse, ont été testées par de nombreuses expériences avec des données simulées et réelles. Nous proposons de donner dans ce résumé quelques exemples caractéristiques permettant d'illustrer l'utilisation de ces nouvelles méthodes.

F-1.1 Classification par K plus proches voisins crédibiliste (BCKN)

Dans le BCKN, chaque échantillon est classé selon son voisinage dans la topologie de l'espace de représentation, et les K voisins les plus proches (de KNN) dans chaque classe sont considérés. Un total de $c \times K$ voisins (c étant le nombre de classes) voisins est utilisé pour classer l'échantillon courant. $c \times K$ assignations des masses (BBA) sont réalisées en fonction de la distance entre l'échantillon courant et ses voisins. Une fusion globale de ces BBA est réalisée pour évaluer les fonctions de croyance de l'échantillon à chaque classe. Cela peut conduire à une prise de décision portant sur une classe simple, une méta-classe ou une classe aberrante.

Un échantillon qui est très proche d'une classe particulière sera associé à cette classe spécifique. Un échantillon trop loin de tous les autres échantillons sera naturellement considéré comme une valeur aberrante. Si l'échantillon est proche de plusieurs classes spécifiques, alors cet échantillon sera associé à une méta-classe définie par l'union de ces classes spécifiques. La méta-classe révèle l'imprécision dans la classification de cet échantillon, et permet également de réduire les erreurs de classification. Cette classification crédibiliste est très intéressante dans de nombreuses applications, en particulier celles liées à la défense et à la sécurité (comme dans la classification des cibles et la poursuite), car il est généralement préférable d'obtenir un résultat de classification plus robuste (et éventuellement partiellement imprécis) qui pourrait être raffiné plus tard avec d'autres techniques ou de ressources, plutôt que d'obtenir un résultat définitif avec un risque élevé de mauvaise classification.

Si certains échantillons se sont associés à des méta-classes, cela implique que l'information utilisée pour la classification des attributs est insuffisante pour obtenir la classification spécifique de ces échantillons. Ainsi, la sortie de BCKN peut être considérée comme une source d'information intéressante à fusionner avec d'autres sources d'information complémentaires disponibles (le cas échéant) pour obtenir des résultats de classification plus précis dans les systèmes d'informations multi-sources. D'autres techniques sophistiquées et coûteuses peuvent également être utilisées pour classer plus précisément les échantillons dans les méta-classes. L'utilisation de ces techniques sophistiquées supplémentaires dépend fortement de l'importance des conséquences de la décision à prendre. Les échantillons dans une méta-classe représentent généralement un petit sous-ensemble de l'ensemble des données. Donc le prix pour la classification spécifique de ces échantillons qui invoquent des techniques sophistiquées coûteuses ne peut être acceptable que pour un nombre limité d'échantillons, mais pas pour l'ensemble des échantillons, dès le début du processus de classification. Ainsi, la méthode de BCKN fournit un moyen de sélectionner les échantillons (en méta-classe) qui ont besoin d'une attention particulière qui doivent être traités avec prudence, dans la mesure

où des décisions importantes à prendre sont nécessaires.

F-1.1.1 Affectation du jeu de masse (BBA)

Considérons des échantillons $\mathbf{y}_s \in Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_h\}, s = 1, \ldots, h$ à classifier parmi un cadre de c classes $\Omega = \{w_1, \ldots, w_c\}$ avec un ensemble d'échantillons d'entraînement $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. w_0 représente une classe inconnue dans Ω pour respecter l'hypothèse d'exhaustivité (ou monde dit fermé). Les KNNs de $\mathbf{y}_s$ pour chaque classe doivent d'abord être trouvés, soit les $c \times K$ voisins sélectionnés. L'assignation des masses associées à $\mathbf{y}_s$ peut être effectuée à partir de la distance entre les $\mathbf{y}_s$ et les $c \times K$ voisins par l'équation suivante :

$$\begin{cases} m_{si}(w_g) = e^{-\gamma d_{si}} \\ m_{si}(\Omega) = 1 - e^{-\gamma d_{si}} \end{cases} \tag{F-1.1}$$

où d_{si} est la distance entre $\mathbf{y}_s$ et $\mathbf{x}_i$. $\gamma > 0$ de l'éq. (F-1.1) est un paramètre d'ajustement dans l'assignation des masses. $c \times K$ assignations de masses, correspondant aux $c \times K$ voisins sélectionnés de $\mathbf{y}_s$ dans chaque classe peuvent être effectuées en suivant cette procédure.

F-1.1.2 Fusion des assignements de masse

Les résultats de la fusion des $c \times K$ assignations de masses (BBA) sont utilisés pour le classement crédibiliste de l'échantillon courant. Les $c \times K$ BBA peuvent être classées en c groupes selon les étiquettes des voisins à partir desquelles les BBA ont été obtenues. Les BBA d'un même groupe sont toutes associées à la même classe, alors que les BBA des différents groupes correspondant aux différentes classes peuvent être en conflit. Donc dans ce cas, ces BBA sont fusionnées en suivant les deux étapes suivantes :

Etape 1 (sous-combinaison) : Nous combinons toutes les BBA appartenant aux mêmes classes, et ce sous-ensemble est appliqué à toutes les classes disponibles.

Etape 2 (fusion globale) : nous combinons les c BBA résultant de la sous-combinaison précédente.

Dans la première étape, il n'est pas approprié d'utiliser la règle de Dempster-Shafer (DS) ici pour la fusion des BBAs appartenant aux mêmes classes en raison de leur structure particulière. En effet, cela donnerait une convergence très rapide vers un singleton. Nous proposons d'utiliser une règle simple de fusion par la moyenne. Elle est définie pour $g = 1, \ldots, c$ par

$$\begin{cases} m_s^g(w_g) = \frac{1}{K} \sum_{i=1}^{K} m_{si}(w_g) \\ m_s^g(\Omega) = \frac{1}{K} \sum_{i=1}^{K} m_{si}(\Omega). \end{cases} \tag{F-1.2}$$

Les BBAs issues de l'étape 1 en fonction des différents groupes sont combinées lors de l'étape 2 pour la classification crédibiliste finale de l'échatillon $\mathbf{y}_s$. Dans ce processus de fusion globale, la croyance en conflit partiel produite par la conjonction de croyances des différentes classes spécifiques exhaustives reflète le degré d'ambiguïté (difficulté) de la classification des échantillons dans les classes spécifiques concernées. Par conséquent, les croyances contradictoires seront associées préférentiellement à la méta-classe correspondante. Si toutes les croyances contradictoires sont conservées et associées aux méta-classes correspondantes, alors trop d'échantillons seront affectés à ces méta-classes. Ce n'est pas une solution de classification des données très efficace, et nous

proposons de sélectionner les méta-classes disponibles en fonction de la masse de croyance des classes spécifiques dans le contexte actuel.

Les résultats de la sous-combinaison liés à $\mathbf{y}_s$ pour les différentes classes peuvent être fusionnés séquentiellement par

$$m_s^{1,g}(A) = \begin{cases} \sum\limits_{B_1,B_2\in 2^\Omega | B_1\cap B_2=A} m_s^{1,g-1}(B_1)m_s^g(B_2), & \text{pour } A \notin \Psi \\ \sum\limits_{B_1,B_2\in 2^\Omega | B_1\cup B_2=A} m_s^{1,g-1}(B_1)m_s^g(B_2), & \text{pour } A \in \Psi \end{cases} \tag{F-1.3}$$

où Ψ représente la méta-classe considérée. Le résultat de la fusion de l'éq. (F-1.3) n'est pas normalisé et il convient de le normaliser à la fin par

$$m_s(A) = \frac{m_s^{1,c}(A)}{\sum\limits_j m_s^{1,c}(B_j)}. \tag{F-1.4}$$

F-1.1.3 Applications

Avec le BCKN, les échantillons sont directement associés à la classe qui reçoit la masse de croyance maximale. Nous utilisons à la fois le taux d'erreur de classification, et un nouveau concept de taux d'imprécision pour évaluer la performance du BCKN. Pour un échantillon issu de w_i, s'il est classé dans A avec $w_i \cap A = \emptyset$, il sera considéré comme une erreur. Si $w_i \cap A \neq \emptyset$ et $A \neq w_i$, il sera considéré comme une prise de décision imprécise. Le taux d'erreur noté Re est calculé par $Re = N_e/T$, où N_e est le nombre d'échantillons mal classés, et T est le nombre total d'échantillons testés. Le taux d'imprécision noté Ri_j est calculé par $Ri_j = N_{I_j}/T$, où N_{I_j} est le nombre d'échantillons associés aux méta-classes ayant un cardinal j. Par commodité, nous avons noté $w_i \cup \cdots \cup w_j \triangleq w_{i,\ldots,j}$.

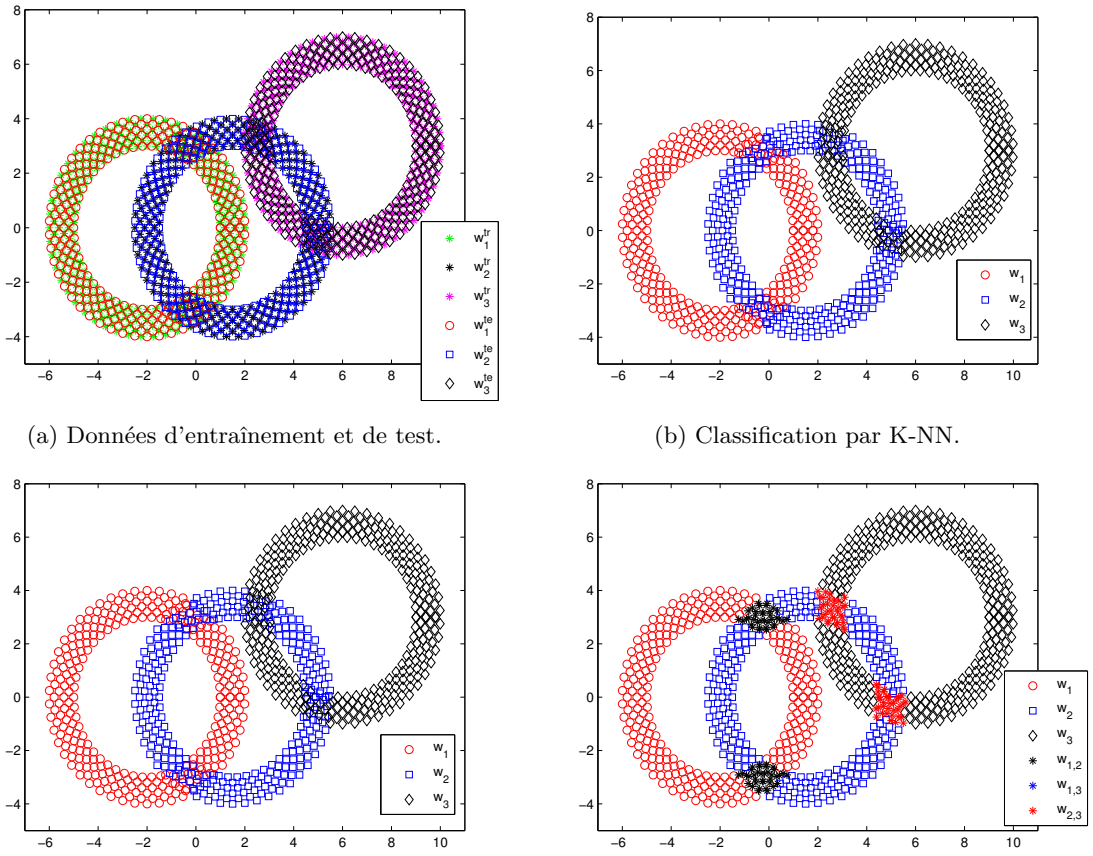
F-1.1.3.1 Application 1 (Données simulées)

Cette expérience montre comment fonctionne le BCKN et sa différence par rapport aux méthodes EK-NN (evidential K-nearest neighbor) et K-NN (K-nearest neighbor). Nous avons utilisé un jeu de données bi-dimensionnelles avec 3 classes, composé de trois anneaux comme représenté sur la figure.F-1.1-(a). Les résultats de la classification des données de test par K-NN, EK-NN et BCKN sont respectivement présentés sur la figure F-1.1-(b), (c), (d).

Nous pouvons voir que les trois anneaux se croisent, et les échantillons se trouvant dans la zone de recouvrement (intersection) sont impossibles à classer correctement. Dans les résultats de la classification de K-NN et EK-NN, tous ces échantillons sont associés à une classe particulière. K-NN et EK-NN génèrent tous deux 109 erreurs de classification. Dans BCKN, les échantillons dans les zones de recouvrement sont directement associés à des méta-classes, comme indiqué sur la figure. F-1.1-(d). Le BCKN ne produit que 4 erreurs de classification, mais il associe 141 échantillons dans les méta-classes. Cet exemple montre l'efficacité de BCKN pour traiter les données ambiguës dans une situation complexe.

F-1.1.3.2 Application 2 (Données réelles)

Quatre jeux de données bien connues et disponibles à partir de la base de l'UCI (vin, Iris, cancer du sein et données de levure) sont utilisés pour évaluer la performance du BCKN par rapport au ANN (artificial neural network), au CART (classification and regression tree) et au SVM (support



(a) Données d'entraînement et de test.

(b) Classification par K-NN.

(c) Classification par EK-NN.

(d) Classification par BCKN ($t = 0.002$).

FIGURE F-1.1 : Comparaison des résultats de classification à 3 classes par K-NN, EK-NN et BCKN .

vector machine). Le taux moyen d'erreur Re_a, le taux d'imprécision Ri_a (pour le BCKN) et le temps d'exécution (en secondes) avec un nombre de classes K variant de 5 à 15 pour les différentes méthodes sont donnés dans le tableau F-1.1.

On peut voir que le taux d'erreur de classification du BCKN est généralement plus faible que celui des autres méthodes puisque les échantillons difficiles à classer sont automatiquement associés aux méta-classes par le BCKN. Celui-ci requiert un peu plus de temps que K-NN et EK-NN comme le montre le tableau F-1.1. Les résultats du BCKN indiquent clairement que les attributs utilisés sont en fait insuffisants pour une bonne classification en dehors des méta-classes. Nous devrions traiter ces échantillons avec plus de prudence et/ou avec d'autres sources d'information pour obtenir des résultats plus spécifiques (si nécessaire). Nos tests et analyses illustrent l'intérêt et le potentiel de la méthode BCKN dans ce genre de problèmes de classification.

F-1.2 RÈGLE DE CLASSIFICATION CRÉDIBILISTE (CCR)

La classification BCKN peut ainsi être utilisée pour la classification crédibiliste de données incertaines en général, mais elle requiert une charge de calcul élevée, car les distances entre l'échantillon

TABLE F-1.1 : Résultats (en %) de classification pour différents jeux de données réelles.

		Levure	Cancer	Vin	Iris
K-NN	Re_a	35.97	3.16	30.45	2.79
	time	**0.0143**	**0.0098**	**0.0030**	**0.0024**
EK-NN	Re_a	35.24	2.99	30.25	3.15
	time	0.2261	0.1326	0.0228	0.0186
CART	Re	37.71	5.59	11.67	5.33
	time	0.8034	0.1934	0.1045	0.0811
ANN	Re	58.76	3.97	63.33	4.67
	time	6.7049	6.4584	3.3072	2.9905
SVM	Re	34.29	3.95	5.00	2.67
	time	2.1528	1.8861	0.2792	0.2308
BCKN	Re_a	**27.05**	**2.54**	**23.84**	**2.55**
	Ri_{a2}	17.59	1.34	16.01	0
	time	0.7484	0.2714	0.0211	0.0156

courant et tous les échantillons du voisinage étendu doivent être calculées. Donc, nous proposons ici une nouvelle solution simple, appelée règle de classification crédibiliste (Credal Classification rule, CCR), permettant de simplifier le calcul des masses dans les cas simples, c'est-à-dire où chaque classe peut être caractérisée par son centre calculé à partir de données d'entraînement.

Dans l'approche CCR, le centre de chaque classe peut être obtenu simplement par calcul de barycentre des échantillons d'entraînement d'une même classe. Le centre d'une méta-classe est calculé à partir des centres des classes simples incluses dans la méta-classe. Dans le problème de classification multi-classe, il y a habituellement quelques classes (pas toutes) qui se chevauchent en partie, et la plupart des classes qui sont en fait loin les unes des autres peuvent être séparées facilement. Les méta-classes, définies par l'union des classes qui sont éloignées les unes des autres, ne sont pas utiles dans des applications réelles. Afin de réduire la complexité de calcul, nous avons juste besoin de sélectionner les méta-classes utiles en fonction du contexte applicatif. L'attribution des masses de croyance de l'échantillon courant est déterminé en fonction de la distance de Mahalanobis qui le sépare des centres de classe simple. Le rapport de la distance maximale de l'échantillon dans les centres des classes simples sur la distance minimale, est introduit pour mesurer le degré de séparation de ces classes. Ainsi, la masse d'une méta-classe est déterminée à partir de la distance entre l'échantillon et le centre de la méta-classe et de la valeur du ratio correspondant. L'approche CCR fournit des résultats de classification crédibliliste avec une faible complexité calculatoire. Le CCR se décompose en deux étapes principales : 1) la détermination des centres des classes simples et méta-classes, et 2) la phase de calcul des BBA est basée sur les distances entre l'échantillon et chaque centre de classe.

F-1.2.1 Détermination des centres de classe

Considérons des données $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ qui doivent être classées dans un ensemble de classes $\Omega = \{w_1, \ldots, w_h\}$ en utilisant les échantillons d'entraînement $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_g\}$. Le centre de chaque classe est simplement défini par la valeur moyenne des données d'apprentissage $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_g\}$ appartenant à la classe correspondante. On suppose que les centres de classes $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_h\}$ sont donnés, et correspondent aux centres des classes simples $\{w_1, \ldots, w_h\}$.

Nous considérons que les centres des classes simples impliquées dans une méta-classe sont indiscernables pour le centre de la méta-classe selon la mesure de distance. Dans ce travail, nous proposons que le centre d'une méta-classe doit être situé à la même distance de Mahalanobis de

tous les centres des classes spécifiques incluses dans la méta-classe. Donc, pour le centre $\mathbf{c}_U$ de la méta-classe U, les conditions suivantes doivent être respectées

$$d(\mathbf{c}_U, \mathbf{c}_i) = d(\mathbf{c}_U, \mathbf{c}_j), \forall w_i, w_j \in U, i \neq j. \tag{F-1.5}$$

Comme on peut obtenir un ensemble de $|U| - 1$ contraintes indépendantes à partir de l'équation (F-1.5), il y aura une seule solution de $\mathbf{c}_U$ lorsque le nombre des attributs de données disponibles est égal à $|U| - 1$. Si le nombre d'attributs est plus grand que $|U| - 1$, il existe de nombreuses solutions pour $\mathbf{c}_U$, et nous choisirons la solution qui est la plus proche de tous les centres des classes simples incluses dans U. Cette solution est donnée par $\mathbf{c}_U = \arg[\min_{\mathbf{c}} \sum_{w_j \in U} (d(\mathbf{c}, \mathbf{c}_j))]$. Si la dimension de $\mathbf{c}_U$ est inférieure à $|U|-1$, on doit résoudre un problème d'optimisation pour chercher la solution $\mathbf{c}_U$ qui satisfasse toutes les contraintes, autant que possible, c'est-à-dire

$$\forall w_i, w_j \in U, i \neq j, \quad \arg[\min_{\mathbf{c}_U} \sum_{w_i, w_j \in U} (d(\mathbf{c}_U, \mathbf{c}_i) - d(\mathbf{c}_U, \mathbf{c}_j))^2].$$

Ceci peut être réalisé en utilisant n'importe quel procédé d'optimisation non linéaire classique. En outre, le centre de méta-classe doit être plus près des centres de ces classes spécifiques impliqués que d'autres centres de classes incompatibles telles que $\max_{w_i \in U} d_{Ui} < \min_{w_k \notin U} d_{Uj}$. Sinon, cette méta-classe ne peut pas être incluse dans les résultats de la classification crédibiliste.

F-1.2.2 Assignation des masses

La masse de croyance de la classe simple (par exemple w_i) est basée sur la distance de Mahalanobis entre l'échantillon et le centre de la classe correspondante, et il peut être défini par

$$\tilde{m}(w_i) = e^{-d(\mathbf{y}_s, \mathbf{c}_i)}. \tag{F-1.6}$$

Dans la détermination de la masse sur la méta-classe, le rapport $\gamma = d_{\max}/d_{\min}$ de la distance maximale $d_{\max}$ de l'échantillon aux centres des classes simples incluses dans U sur la distance minimale $d_{\min}$ est introduite afin de mesurer le degré de distinction entre les classes de U. Un ratio de faible valeur indique un degré de distinction faible parmi les classes de U de l'échantillon. Ainsi, la valeur du rapport γ servira à mettre plus ou moins de masse de croyance à la méta-classe U. La masse de l'échantillon $\mathbf{y}_s$ avec la méta-classe U est mathématiquement définie par

$$\tilde{m}(U) = e^{-\lambda_U \gamma_U d(\mathbf{y}_s, \mathbf{c}_U)}, \quad \text{for } |U| \geq 2 \tag{F-1.7}$$

où

$$d(\mathbf{y}_s, \mathbf{c}_U) = \frac{1}{|U|} \sum_{w_i \in U} \sqrt{\sum_{k=1}^{N} \frac{(\mathbf{y}_s(k) - \mathbf{c}_U(k))^2}{\delta_i(k)^2}} \tag{F-1.8}$$

$$\gamma_U = \frac{\max_{w_i \in U} d(\mathbf{y}_s, \mathbf{c}_i)}{\min_{w_i \in U} d(\mathbf{y}_s, \mathbf{c}_i)} \tag{F-1.9}$$

avec $\lambda_U = \eta |U|^\alpha$. La quantité $|U|^\alpha$ est une pondération de pénalité pour les méta-classes ayant une grande cardinalité. η est un paramètre de réglage utilisé pour gérer le nombre d'échantillons déterminés pour les méta-classes. L'échantillon sera considéré comme aberrant s'il apparaît loin de toutes les autres classes selon un seuil d'aberration t. La masse de l'échantillon dans la classe des valeurs aberrantes w_0 est définie par :

$$\tilde{m}(w_0) = e^{-t} \tag{F-1.10}$$

$\forall A \subseteq \Omega$, la masse non normalisée précédente de croyance $\tilde{m}(\cdot)$ peut être simplement normalisée pour la classification crédibiliste de l'échantillon $\mathbf{y}_s$.

F-1.2.3 Evaluation du CCR sur des données simulées de grande dimension

La performance du CCR a été évaluée sur plusieurs ensembles de données simulées et réelles. Nous donnons ici un exemple simple pour illustrer l'utilisation du CCR dans le traitement à grande échelle. Les simulations sont générées avec quatre classes w_1, w_2, w_3 et w_4 à l'aide de 4 gaussiennes en dimension 30 ayant des moyennes et matrices de covariance comme suit (selon des instruction **matlab**®) : $\mu_1 = \mathbf{zeros}(\mathbf{1}, \mathbf{30}), \mathbf{\Sigma}_1 = 10 \cdot \mathbf{I}; \mu_2 = 5 \cdot \mathbf{ones}(\mathbf{1}, \mathbf{30}), \mathbf{\Sigma}_2 = 10 \cdot \mathbf{I}; \mu_3 = 20 \cdot \mathbf{ones}(\mathbf{1}, \mathbf{30}), \mathbf{\Sigma}_3 = 15 \cdot \mathbf{I}; \mu_4 = 30 \cdot \mathbf{ones}(\mathbf{1}, \mathbf{30}), \mathbf{\Sigma}_4 = 15 \cdot \mathbf{I}$.

Dans chaque classe, nous utilisons le même nombre (*i.e.* n) d'échantillons d'apprentissage et d'échantillons de test. Donc, il y a en tout $N = 4 \times n$ échantillons d'apprentissage et $N = 4 \times n$ échantillons de test, avec $N = 8000, 40000, 200000, 1000000$. Le taux d'erreur Re, le taux d'imprécision Ri_j, et le temps de calcul t (en secondes) sont moyennés sur 10 simulations de Monte Carlo, et sont donnés dans le tableau F-1.2. "NA" (not available) signifie "Sans échantillon".

TABLE F-1.2 : Résultats (en %) de classification avec des données de grande dimension.

	ANN (Re, time)	CART (Re, time)	EK-NN (Re, time)	CCR (Re, Ri_2, time)
N=8000	(33.09, 15.6313)	(29.59, 1.2168)	(8.46, 47.5023)	(**5.26**, 5.84, **0.2340**)
N=40000	(35.04, 58.9684)	(26.66, 6.4428)	(8.25, 1669.1)	(**5.15**, 6.41, **1.1544**)
N=200000	(33.93, 241.7703)	(24.34, 35.1470)	NA	(**5.11**, 6.24, **5.8032**)
N=1000000	NA	(22.25, 200.3053)	NA	(**5.14**, 6.16, **29.0162**)

Nous pouvons voir que la classification CCR produit le taux d'erreur le plus bas avec quelques imprécisions partielles, dues à la difficulté d'associer certains échantillons à la bonne méta-classe. Cependant, le CCR est beaucoup plus rapide que les autres méthodes. EK-NN peut obtenir des résultats de classification raisonnables, mais il requiert plus de temps d'execution. ANN et CART induisent des taux d'erreur beaucoup plus élevés que le CCR et EK-NN, et ils sont aussi beaucoup plus lents que CCR. ANN n'est pas applicable pour les grands ensembles de données (surtout pour $N = 1000000$) en raison de sa haute charge de calcul. Ainsi, il apparaît que le CCR est une alternative intéressante permettant de faire face aux grands ensembles de données à grande dimension grâce à sa faible charge de calcul et sa faible complexité.

F-1.3 CLASSIFICATION CÉDIBILISTE PAR PROTOTYPE (PCC) POUR LES DONNÉES INCOMPLÈTES

Dans de nombreuses applications, la qualité des données peut souffrir du fait que certains échantillons soient incomplets avec des composantes manquantes ou inconnues. Les différentes estimations de ces valeurs manquantes peuvent conduire à des résultats de classification différents, et il est difficile de classer correctement l'échantillon dans la bonne classe parce que ses composantes disponibles peuvent être insuffisantes pour une classification simple. Une nouvelle méthode de classification crédibiliste basée sur des prototypes (PCC) pour les modèles incomplets a donc été développée. Un échantillon ne pouvant être classé correctement en raison de l'imprécision causée par des valeurs manquantes de ses composantes sera raisonnablement associé à une méta-classe appropriée définie par l'union (disjonction) de plusieurs classes simples dans lesquelles cet échantillon peut appartenir. Cette approche nous permet de réduire le taux d'erreur de classification et de révéler l'imprécision de la classification. Dans PCC, les prototypes de toutes les classes obtenues par les données d'entraînement (dans un modèle complet) sont utilisées pour estimer les

valeurs manquantes du modèle incomplet. Ainsi, dans un problème avec c-classes, on a affaire à c estimations des valeurs manquantes. L'échantillon, avec c valeurs estimées, est classé selon un classificateur standard, et le PCC va produire c résultats de classifications partielles représentées par des assignations de masse (BBA). Ces c résultats partiels ont différents facteurs de pondération (déterminés par les distances entre l'échantillon et les prototypes) qui servent à l'affaiblissement des masses. La fusion globale des c résultats actualisés fournit la classification finale crédibiliste de l'échantillon. Dans ce processus de fusion, les méta-classes seront conditionnellement conservées en fonction du degré d'incertitude des échantillons qui sont difficiles à classer correctement. Les croyances contradictoires peuvent bien représenter l'imprécision (ambiguïté) du degré de classification, et elles sont transférées aux méta-classes correspondantes en fonction du contexte.

F-1.3.1 Classification de données incomplètes avec c estimations

Considérons un jeu de données de test $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ et un jeu d'entraînement $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_H\}$ avec un ensemble de classes $\Omega = \{\omega_1, \ldots, \omega_c\}$. On suppose que les échantillons de test (qui sont des vecteurs) sont tous incomplets avec des composantes manquantes uniques ou multiples, alors que les données d'apprentissage Y sont complètes.

Le prototype de chaque classe *i.e.* $\{\mathbf{o}_1, \ldots, \mathbf{o}_c\}$ est calculé par simple barycentre des données d'entraînement de chaque classe, et $\mathbf{o}_g$ correspond au centre de la classe ω_g. Une fois les prototypes de classe estimés, nous utilisons le barycentre du prototype pour combler les valeurs manquantes de l'échantillon à la composante correspondante. Comme on a considéré c classes possibles, on obtient c valeurs estimées différentes. Pour chacune d'entre-elles $\mathbf{y}_i^g$, $g = 1, 2 \ldots, c$, nous pouvons obtenir une classification en utilisant n'importe quel classificateur standard noté $\Gamma(\cdot)$. Les résultats des c classifications des $\mathbf{y}_i$ sont donnés pour $g = 1, \ldots, c$ par

$$\mathbf{P}_i^g = \Gamma(\mathbf{y}_i^g | Y). \tag{F-1.11}$$

Dans le PCC, nous proposons de combiner ces c classifications partielles pour obtenir une classification crédibiliste de l'échantillon incomplet. Le facteur de pondération (pour l'affaiblissement) de chaque classification peut être déterminé par la distance entre l'échantillon et le centre de la classe simple correspondante, c'est-à-dire par

$$w_i^g = e^{-d_{ig}} \tag{F-1.12}$$

où $d_{ig} = \sqrt{\frac{1}{p}\sum_{s=1}^{p}\left(\frac{y_{is}-o_{gs}}{\delta_{gs}}\right)^2}$. p est le nombre de composantes connues de $\mathbf{y}_i$. δ_{gs} est la distance moyenne de toutes les données d'entraînement appartenant à la classe ω_g restreinte à la composante s du centre o_g. T_g est le nombre d'échantillons d'apprentissage dans la classe ω_g.

À partir de ces facteurs de pondération w_i^g pour $g = 1, \ldots, c$, on définit les facteurs de fiabilité α_i^g par $\alpha_i^g = \frac{w_i^g}{w_i^{\max}}$ avec $w_i^{\max} = \max(w_i^1, \ldots, w_i^c)$. La méthode d'affaiblissement des sources est appliquée ici, et les masses réduites sont obtenues pour $g = 1, \ldots, c$ par

$$\begin{cases} m_i^g(A) = \alpha_i^s P_i^g(A), \quad A \subset \Omega \\ m_i^g(\Omega) = 1 - \alpha_i^g + \alpha_i^g P_i^g(\Omega). \end{cases} \tag{F-1.13}$$

F-1.3.2 Fusion globale des c classificateurs affaiblis

Les c classifications peuvent être divisées en plusieurs groupes distincts $G_1, G_2, \ldots, G_r$ selon les classes qui sont supportées. Les résultats de la classification dans un même groupe sont combinés

entre eux, puis fusionnés dans la classification crédibiliste. Les résultats de la classification d'un même groupe ne sont généralement pas en conflit sévère. Par conséquent, on propose d'appliquer la règle DS. Pour $G_s = \{\mathbf{m}_i^j, \ldots, \mathbf{m}_i^k\}$, les résultats de la fusion des BBA du groupe G_s en utilisant la règle DS sont notés par $\mathbf{m}_i^{\omega_s}(\cdot)$.

Dans le processus de fusion globale, la combinaison des résultats des différents groupes de résultats peuvent être en conflit sévère à cause des classes différentes qu'ils soutiennent fortement. Nous proposons de sélectionner les croyances conflictuelles qui doivent être transférés aux méta-classes correspondantes. L'ensemble des classes auxquelles l'échantillon appartient probablement est donné par $\Lambda_i = \{\omega_s | m_i^{\omega_{\max}}(\omega_{\max}) - m_i^{\omega_s}(\omega_s) < \epsilon\}$. $\epsilon \in [0,1]$ est un seuil choisi, et ω_{max} est la classe ayant la plus grande masse pour l'échantillon considéré. Nous proposons de garder tous les sous-ensembles de Λ_i dans le processus de fusion et nous traitons les méta-classes associées.

La règle de fusion globale est alors définie par :

$$\tilde{m}_i(A) = \begin{cases} \sum\limits_{\bigcap_{g=1}^{r} B_g = A} m_i^{\omega_1}(B_1) \cdots m_i^{\omega_r}(B_r), & \text{pour } A \in \Omega \text{ avec } |A| = 1, \text{ou } A = \Omega \\ \sum\limits_{\bigcap_{i=1}^{|A|} B_i = \emptyset \; \bigcup_{i=1}^{|A|} B_i = A} \left[m_i^{\omega_1}(B_1) \cdots m_i^{\omega_s}(B_s) \prod\limits_{g=|A|+1}^{r} m_i^{\omega_g}(\Omega) \right] & \text{pour } A \subseteq \Lambda_i, \text{avec } |A| \geq 2. \end{cases} \tag{F-1.14}$$

Dans l'équation (F-1.14), r est le nombre des groupes des c classifications. Comme toutes les masses en conflit partiel ne sont pas transférées dans les méta-classes à travers la formule de fusion globale (F-1.14), la fonction de masse finale est normalisée avant la prise de décision.

F-1.3.3 Application de la PCC

Dans cette expérience, on utilise quatre jeux de données réelles pour évaluer la performance du PCC par rapport à d'autres méthodes, telles que l'EK-NN (SC :A) et l'ENN (SC :B) qui ont été choisies ici comme classificateurs standard. Une simple validation croisée deux-par-deux a été réalisée sur les quatre jeux de données avec les différentes méthodes de classification. Chaque échantillon de test a n valeurs manquantes (ou inconnues), et cela est fait par tirage aléatoire selon les différentes composantes. Le taux d'erreur moyen Re_a des différentes méthodes classiques et le taux d'imprécision Ri_a (pour le PCC) sont donnés dans le tableau F-1.3.

Les résultats du tableau F-1.3 montrent clairement que la méthode PCC produit généralement un taux d'erreur inférieur à celui des méthodes de classification MI, KNNI et FCMI, mais parallèlement, il donne une certaine imprécision dans le résultat de la classification due à l'introduction des méta-classes, dont la présence indique que certains échantillons incomplets reste très difficiles à classer. L'augmentation du nombre (n) de valeurs manquantes dans chaque échantillon de test provoque généralement l'augmentation du taux d'erreur dans les classificateurs. Le taux de l'imprécision devient plus important dans le PCC, étant donné que les valeurs manquantes conduisent à une plus grande imprécision (incertitude). Ainsi, la classification crédibiliste est très utile et efficace pour représenter le degré de l'imprécision et il peut aussi aider à diminuer le taux d'erreur.

TABLE F-1.3 : Résultat (en %) de classification des différents jeux de données.

data set	(n, SC)	MI	KNNI	FCMI	PCC
		Re	Re	Re	$\{Re, Ri_2\}$
Cancer du Sein	(3, A)	4.71	6.10	3.95	{4.10, 3.38}
	(3, B)	4.25	3.95	3.81	{3.81, 2.34}
	(5, A)	8.20	8.15	5.07	{4.38, 4.69}
	(5, B)	6.44	5.76	5.27	{3.81, 6.00}
	(7, A)	38.33	14.35	13.00	{7.91, 8.05}
	(7, B)	14.64	11.54	11.42	{6.88, 12.44}
Levure	(1, A)	37.59	38.13	38.54	{34.36, 6.95}
	(1, B)	37.71	36.70	36.19	{32.67, 6.19}
	(3, A)	45.08	44.29	45.95	{34.71, 18.00}
	(3, B)	42.10	40.90	41.33	{34.19, 14.95}
	(5, A)	51.16	50.95	51.11	{33.46, 31.01}
	(5, B)	49.33	49.22	46.00	{32.29, 27.62}
Graines	(3, A)	21.03	9.68	12.46	{7.14, 3.72}
	(3, B)	21.43	11.19	13.33	{9.05, 2.86}
	(5, A)	33.49	12.54	20.08	{9.67, 6.70}
	(5, B)	31.43	12.14	20.00	{9.52, 9.05}
	(6, A)	40.71	25.87	21.75	{16.79, 12.77}
	(6, B)	39.52	25.71	20.95	{16.19, 14.76}
Vin	(3, A)	30.71	26.59	30.15	{26.05, 1.05}
	(3, B)	29.78	26.97	26.97	{26.97, 1.69}
	(6, A)	34.93	25.84	32.12	{26.62, 0.84}
	(6, B)	33.71	28.09	32.02	{27.53, 1.12}
	(10, A)	39.23	30.90	32.30	{25.84, 3.86}
	(10, B)	37.64	31.18	31.46	{27.53, 3.93}

F-1.4 K-MOYENNES FLOUES ET ÉVIDENTIELLES (CCM)

Le partitionnement crédibiliste a été proposé récemment pour le regroupement (clustering) de données en utilisant des fonctions de croyance. Il permet aux échantillons d'appartenir non seulement aux hypothèses simples, mais aussi à un ensemble de classes (c'est-à-dire des méta-classes). Nous proposons une nouvelle version crédibiliste de la FCM appelée K-moyennes floues et évidentielles (CCM) qui travaille à partir d'un partitionnement crédibiliste. Dans la CCM, la notion de méta-classe est considérée comme une sorte de groupe de transition entre les différentes classes simples. Si un échantillon est considéré dans une méta-classe, il doit être à la fois proche des classes simples incluses dans la méta-classe, et son appartenance dépend principalement des distances relatives entre l'échantillon et les centres des classes simples. Cependant, ces classes simples doivent être indiscernables pour l'échantillon courant, ce qui indique qu'il est difficile à associer correctement à une des classes simples. Cela dépend principalement de la distance au centre de la méta-classe. Ainsi, dans la détermination de la croyance sur la méta-classe, nous devons prendre en compte non seulement la distance au centre de la méta-classe, mais aussi les distances aux centres des classes simples concernées.

Une fonction de coût intervient dans la méthode CCM. Elle est définie par le principe suivant. Les centres de classe et la croyance de chaque classe pour les échantillons peuvent être obtenus par la minimisation de cette fonction de coût. Pour des données avec c-classes, le partitionnement crédibiliste produit 2^c classes, et sa complexité de calcul est très élevée lorsque c est grand. Dans les applications réelles, la classification des données imprécises est généralement distribuée parmi

plusieurs (un très petit nombre, par exemple, deux ou trois) classes simples, et il y a très peu de données appartenant à des méta-classes ayant une grande cardinalité. Donc, un seuil T_c est introduit dans le CCM pour éliminer les méta-classes de grande cardinalité, ce qui réduit de façon significative le coût calculatoire.

F-1.4.1 Fonction de coût du CCM

Considérons un ensemble de $n > 1$ échantillons à classifier dans $\Omega = \{w_1, w_2, \ldots, w_c\}$ avec les centroïdes correspondants $\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_c}\}$. La partition crédibiliste est définie dans le cadre de discernement 2^Ω. En fait, nous n'avons pas à tenir compte de toutes les méta-classes de 2^Ω, et nous pouvons éliminer certaines d'entre elles qui ont une cardinalité importante en fonction d'un seuil $T_c \in [2, 2^{|\Omega|}]$. Dans le CCM, l'ensemble des clusters disponibles est donné par $S^\Omega = \{A_i / |A_i| \leq T_c\} \subseteq 2^\Omega$. T_c est généralement petit (limité à deux, ou trois). Cela peut réduire significativement la complexité de calcul de la segmentation.

La fonction de coût du CCM est donnée par :

$$J_{CCM}(M, V) = \sum_{i=1}^{n} \sum_{j/A_j \in S^\Omega} m_{ij}^\beta D_{ij}^2 \tag{F-1.15}$$

avec

$$D_{ij}^2 = \begin{cases} \delta^2; & \text{pour } A_j = \emptyset \\ d_{ij}^2; & \text{lorsque } |A_j| = 1, \\ \dfrac{\sum\limits_{A_k \in A_j} d_{ik}^2 + \gamma d_{ij}^2}{|A_j| + \gamma}; & \text{si } |A_j| > 1, \end{cases} \tag{F-1.16}$$

où $M = (m_1, \cdots, m_n) \in \mathbb{R}^{n \times 2^{|\Omega|}}$ est la matrice regroupant les masses de tous les échantillons et $V_{c \times p}$ est la matrice de centres de classe. J_{CCM} doit satisfaire la contrainte suivante :

$$\sum_{j | A_j \in S^\Omega} m_{ij} = 1, \tag{F-1.17}$$

d_{ij} étant la distance entre les échantillons $\mathbf{x}_i$ et les centroïdes A_j. Si A_j est le centroïde d'une classe simple, alors il coïncide avec $\mathbf{v_j}$. Sinon, c'est le barycentre des centres des classes simples incluses dans A_j.

La justification de la fonction de coût de la CCM est la suivante. La croyance d'un échantillon sur la classe aberrante, représentée par $\emptyset$, est principalement déterminée par la valeur du seuil δ. La croyance d'un échantillon sur une classe simple est proportionnelle à la distance entre l'échantillon et le centre de la classe simple. La croyance d'un échantillon sur une méta-classe est proportionnelle à la distance moyenne des centres de classes simples impliquées, et aussi à la distance au centre de la méta-classe avec un facteur de pondération γ.

Dans le CCM, la matrice des masses de croyance $M = (\mathbf{m}_1, \ldots, \mathbf{m}_n)$ et la matrice des centres de regroupement $V_{c \times p}$ peuvent être obtenues par la minimisation de la fonction de coût J_{CCM}. La masse de croyance associée à différents éléments focaux est donnée par

$$\sum(D) = \sum_{A_j = \emptyset} \delta^{\frac{-2}{\beta - 1}} + \sum_{|A_j| = 1} d_{ij}^{\frac{-2}{\beta - 1}} + \sum_{|A_j| > 1} \left(\frac{\sum\limits_{A_k \in A_j} d_{ik}^2 + \gamma d_{ij}^2}{|A_j| + \gamma} \right)^{\frac{-1}{\beta - 1}} \tag{F-1.18}$$

$$m_{ij} = \begin{cases} \frac{\delta^{\frac{-2}{\beta-1}}}{\sum(D)}; & \text{lorsque } A_j = \emptyset \\ \frac{d_{ij}^{\frac{-2}{\beta-1}}}{\sum(D)}; & \text{lorsque } |A_j| = 1 \\ \frac{(\frac{\sum_{A_k \in A_j} d_{ik}^2 + \gamma d_{ij}^2}{|A_j|+\gamma})^{\frac{-1}{\beta-1}}}{\sum(D)}; & \text{lorsque } |A_j| > 1. \end{cases} \tag{F-1.19}$$

Les centres de classe V sont définis par

$$B_{c\times n} X_{n\times p} = H_{c\times c} V_{c\times p} \tag{F-1.20}$$

où

$$B_{li} \triangleq m_{il}^{\beta} + \sum_{A_l \in A_j} m_{ij}^{\beta} \frac{1 + \frac{\gamma}{|A_j|}}{|A_j| + \gamma} \tag{F-1.21}$$

$$H_{ll} \triangleq \sum_{i=1}^{n} m_{il}^{\beta} + \sum_{i=1}^{n} \sum_{A_l \in A_j} m_{ij}^{\beta} \frac{1 + \frac{\gamma}{|A_j|^2}}{|A_j| + \gamma} \tag{F-1.22}$$

$$H_{lq} \triangleq \sum_{i=1}^{n} \sum_{A_l \in A_k, A_q \in A_k} m_{ik}^{\beta} \frac{\gamma}{|A_k|^2(|A_k| + \gamma)}, l \neq q \tag{F-1.23}$$

V est donc solution du système linéaire de l'équation (F-1.20) à travers :

$$V_{c\times p} = H_{c\times c}^{-1} B_{c\times n} X_{n\times p}. \tag{F-1.24}$$

Le pseudo-code de l'algorithme CCM est donné à la table F-1.4.

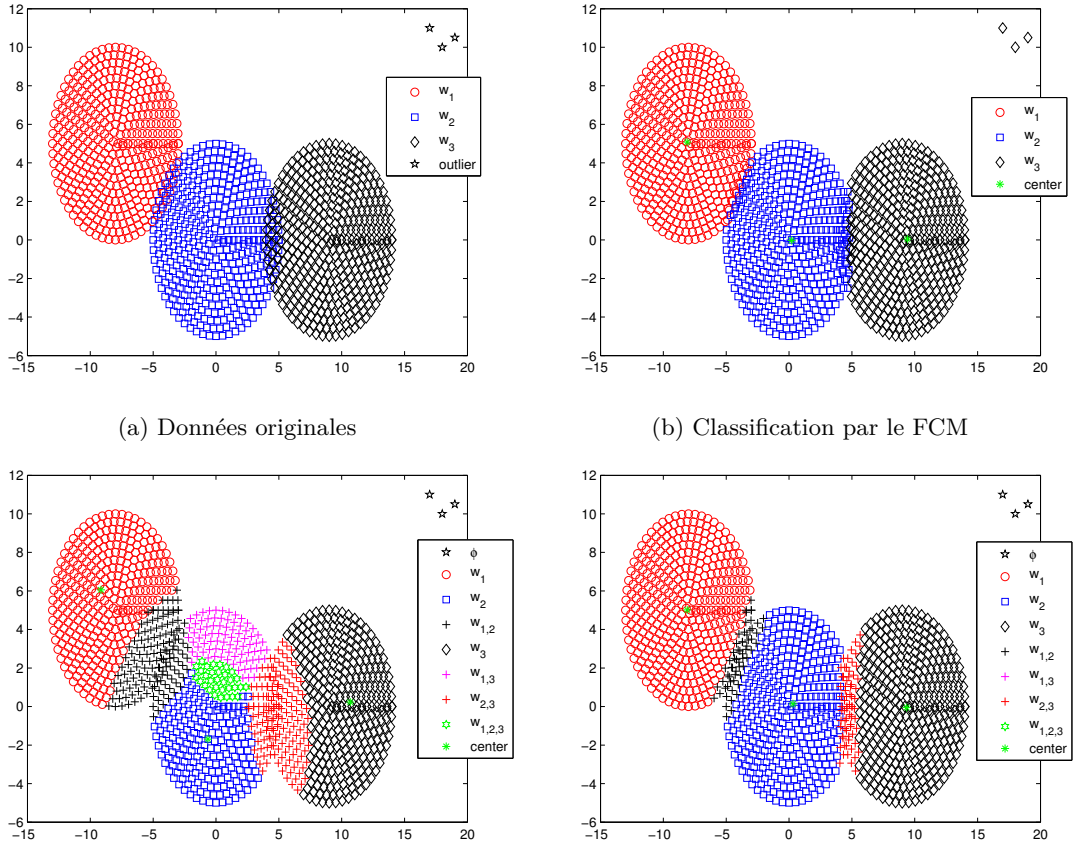
TABLE F-1.4 : Algorithme CCM.

Entrée :	Données : $X = \{\mathbf{x_1}, \cdots, \mathbf{x_n}\}$ dans $\mathbb{R}^p$
Paramètres :	c : nombre de classes, $2 \leq c < n$ t_c : seuil des meta-classes (par défaut $t_c = 3$) $\delta > 0$: seuil des données aberrantes $\gamma > 0$ pondération de la distance (par défaut $\gamma = 1$) $\epsilon > 0$: seuil de fin (par défaut $\epsilon = 0.001$)
Initialisation :	Sélection aléatoire des masses initiales M_0
	t $\leftarrow 0$ Répéter $t \leftarrow t + 1$ Calcul de B_t et H_t avec (F-1.21)-(F-1.23) ; Calcul de V_t par résolution de (F-1.24) ; Calcul de M_t avec (F-1.19) ; jusqu'à $\|V_t - V_{t-1}\| < \epsilon$

F-1.4.2 Evaluation des performances de l'algorithme CCM

F-1.4.2.1 Classification de données simulées à 3 classes

Considérons un ensemble de données avec 3 classes ayant des formes rondes comme indiqué sur la figure F-1.2-(a). Cet ensemble de données est constitué de 1245 points, dont 3 sont aberrantes. Les méthodes FCM, l'ECM et le CCM sont appliquées pour classifier ce jeu de données. Les résultats de classification obtenus avec le FCM, l'ECM et le CCM sont présentés sur la figure F-1.2-(b)-(d).



(a) Données originales

(b) Classification par le FCM

(c) Classification par l'ECM avec $\alpha = 1$.

(d) Classification par le CCM avec $\gamma = 1$.

FIGURE F-1.2 : Comparaison de 3 méthodes de classification évidentielles sur ce jeu de données à 3 classes.

Avec le FCM, les points de la zone de chevauchement sont tous associés à une classe simple, ce qui est susceptible de provoquer des erreurs. w_1 et w_3 ne sont pas proches l'un de l'autre et même totalement séparés, mais il y a beaucoup de points de w_2 qui sont abusivement associés à la méta-classe $w_1 \cup w_3$ par l'ECM. D'ailleurs, de nombreux points de w_2 sont même considérés dans la classe de totale ignorance $w_1 \cup w_2 \cup w_3$. Avec le CCM, aucun point n'est affecté à $w_1 \cup w_3$ ni à $w_1 \cup w_2 \cup w_3$. Certains points à l'intersection de w_1 et w_2 ou w_2 et w_3 sont respectivement associés à $w_1 \cup w_2$ et $w_2 \cup w_3$ comme le montre la figure F-1.2-(d), puisque ces points sont vraiment difficiles à classer correctement dans une classe simple. La méta-classe $w_1 \cup w_2$ peut être interprétée comme

la classe de transition entre w_1 et w_2 (de même pour $w_2 \cup w_3$). Cela peut effectivement réduire les erreurs de classification en utilisant ces méta-classes. Il apparaît finalement que cette approche CCM a de bien meilleures performances que l'ECM et le FCM pour traiter les méta-classes.

F-1.4.2.2 Classification avec des données réelles

Nous donnons ici un résultat de classification sur des données réelles en utilisant les données de la base Iris. Ce résultat illustre la performance du CCM par rapport à l'ECM et au FCM.

TABLE F-1.5 : Résultats (en %) de classification de la base Iris à l'aide des différentes méthodes.

		Re	Ri_2	Ri_3
FCM		10.67	NA	NA
ECM	α=2.0	8.00	**4.67**	0
	α=1.5	10.00	8.67	0.67
	α=1.0	10.00	15.33	6.00
CCM ($t_c = 3$)	γ=1.0	5.33	8.00	0
	γ=1.5	4.67	10.67	0
	γ=2.0	4.00	12.00	0
CCM ($t_c = 2$)	γ=1.0	5.33	8.00	NA
	γ=1.5	4.67	10.00	NA
	γ=2.0	**3.33**	12.00	NA

Dans cette application, le CCM fournit le plus petit nombre d'erreurs par rapport aux autres méthodes, et le nombre d'échantillons dans les méta-classes est plus faible que celui obtenu avec l'ECM. L'augmentation du paramètre γ provoque la diminution de l'erreur mais l'augmentation du degré d'imprécision. Il nous faut donc trouver un compromis entre l'erreur et d'imprécision, et cela dépend aussi du taux d'imprécision que l'utilisateur est prêt à accepter. Avec le CCM, si le seuil T_c de la cardinalité de la méta-classe passe de $T_c = 3$ à $T_c = 2$, cela signifie que la méta-classe dont la cardinalité est trois sera éliminé, et la complexité de calcul se verra diminuée. Cependant, nous constatons que les résultats de cette classification sont presque identiques malgré ces différentes valeurs de T_c. Ainsi, cela montre que l'on peut choisir une petite valeur de T_c dans les applications réelles. Par ailleurs, le CCM peut fournir de bons résultats de classification tout en préservant une charge de calcul acceptable.

PART II : CONTRIBUTIONS

TELECOM Bretagne

1 Introduction

In the classification problem, the uncertainty is a big challenge we often encounter in many applications. The uncertainty can be caused by various reasons, such the randomness, insufficient knowledge, missing values, etc. The traditional classification methods usually work with probabilistic framework, and the probability measures are used to characterize the uncertainty. The object is commonly assigned to the class with the maximum probability. Nevertheless, the probabilistic framework captures only the randomness aspect of the data, but not the fuzziness, nor imprecision which is another inherent aspect of information content [1, 2]. Belief function theory [3–7] also known as evidence theory and Dempster-Shafer theory is often considered as the general extension of Bayes probability theory, and it allows to commit the masses of belief not only to singleton elements (as done in probability framework) but also to any set of elements using basic belief assignments (BBA's). Meanwhile, there have been many methods [4,8–11] emerged for the fusion of multiple BBA's. Belief function theory is an efficient tool to well model and manage the uncertain and imprecise information, and the classification of uncertain data based on belief function theory will de deeply studied in this work. In fact, belief functions has already been more or less successfully applied in many fields, such as the data classification [12–22], clustering [23–28], decision making support [29–31], and image processing [32–37], etc.

The classification methods can be either supervised or unsupervised according to the training information one has or not. When many training samples with known class labels are available, the supervised classification can be applied, and the class of patterns is determined based on the training information. Some classifiers [12, 13, 38] have been developed based on belief functions, which allow to model the total ignorance. Nevertheless, the partial imprecise information has barely been taken into account in the literature so far. In the real applications, the classification is often partially (rather than totally) imprecise among a very small number (e.g. two or three) of classes. The efficient characterization of the partial imprecision is an important topic in the classification problem. In this thesis, we have studied the credal classification of uncertain and imprecise data based on belief functions, and two credal classifiers have been proposed to deal with the different cases. The proposed credal classification allows the object to belong to not only single classes but also meta-class defined by the union of several (any number) single classes with different masses of belief, and the meta-class is introduced to model the partial imprecision of classification and to reduce the error rate. A credal classification method called belief $c \times K$ nearest neighbors has been introduced using the nearest neighbors in each class to deal with the general and more complicate cases. When each class can be represented by one prototype, we have proposed a simple credal classification rule (CCR) which can directly compute the mass of belief of object belonging to each specific class and meta-class with low computation burden. Moreover, the missing attribute data is often encountered in many applications, and the classification of incomplete patterns remains an interesting and important topic. The classical classification methods generally commit the incomplete pattern to a particular class with the maximum probability measure. However, the different estimations of the missing values can produce distinct classification results sometimes, and this causes high imprecision and uncertainty of classification. One credal classification method

for classification of the incomplete data with missing values has been developed based on belief function theory, and it is able to well characterize such uncertain and imprecise information thanks to the meta-class. If there is no training information available, the data clustering analysis (also called unsupervised classification) will be used, and the patterns will be automatically grouped into several clusters according to some (dis-)similarity measures. Fuzzy c-means is (FCM) [39] a very well known clustering method, and an evidential version of FCM (ECM) [23] has been developed for working with belief functions. Nevertheless, when the different clustering centers (i.e. singleton clusters and meta-clusters) are close, ECM usually produces very unreasonable results. Thus, we propose a new credal clustering method called Credal c means (CCM) as an alternative evidential extension of FCM to overcome the limitations of ECM. The main contents of this thesis are organized as follows.

The overview of related literatures is given in Chapter 2. We briefly recall the belief function theory, which proposes a rigorous way to model the uncertainty and imprecision of classification. The well known supervised classifiers, incomplete pattern classification methods and data clustering analysis methods have been simply introduced. Particularly, the evidential K-nearest neighbor (EK-NN) classifier [12] and evidential c-means (ECM) clustering method [23] have been discussed in detail and their limitations are clearly pointed, since we get important inspiration from the two methods.

In Chapter 3, a new belief $c \times K$ neighbors (BCKN) classifier working with credal classification is proposed to deal with the uncertain data. The K nearest neighbors in each class are used for the credal classification, and there are total $c \times K$ neighbors involved (c being the number of classes). The corresponding $c \times K$ BBA's are constructed according to the distances between the object and each neighbor, and the classification result of the object depends on the global fusion of the $c \times K$ bba's. In the fusion process, the bba's obtained from the neighbors in the same class are combined by the simple average method, and these sub-combination results are globally fused by a new proposed combination rule. The conflicting beliefs are conditionally kept according to the context for the selected meta-classes to reveal the imprecision degree of the classification. In BCKN, the object can belong to either the specific classes, or the meta-classes (i.e. the set of several specific classes) for the object hard to correctly classify, or eventually the ignorant class for noisy data. If the object is committed to the meta-classes, it indicates that the specific classes included in the meta-class cannot be clearly distinguished using the available information, and hard classification of the object in a particular class will very likely lead to misclassification. The BCKN credal classification is able to efficiently reduce the error rate thanks to the use of meta-class, which also characterizes the partial imprecision of the classification. BCKN can also warn the user that some other techniques or complementary information sources are necessary to obtain the specific classification of the samples of the meta-class. The performance of BCKN method with respect to other classical methods has been analyzed through several experiments.

BCKN can well deal with the general and complicate situation in the classification problem, but unfortunately it has big computation burden. So in Chapter 4, a new simple and effective credal classification rule (CCR) based on the belief functions has been presented, and it is to manage the case where each class can be well characterized by the prototype vector. In CCR, each specific class corresponds to a center (i.e. prototype) obtained by the mean value of the training data in the same class, and the center of meta-class is considered with the equal Mahalanobis distances to all the centers of the involved specific classes. The acceptable meta-classes are selected according to the current context and distance ratios, and all the unacceptable meta-classes are removed to reduce the number of focal elements and the computational complexity. CCR provides a direct way to compute the mass of belief of the object belonging to different classes (i.e. specific classes and meta-classes) based on the distances between the object and the class centers, and the computational complexity of CCR is quite low. A tuning parameter has been introduced in CCR to control the imprecision rate of classification due to the meta-classes. The experiments using both the artificial

and real data sets will be presented in chapter 4 to evaluate the performance of CCR with respect to other methods.

There exist many industrial and research data sets with missing attribute values caused by various reasons, such as failure of observation, equipment errors and incorrect measurements. The classification of incomplete pattern with missing values is still an important topic, and it is also a very challenging problem, because the different estimations of the missing values can lead to different classification results. Such uncertainty is mainly due to the lack of information of the missing data. In Chapter 5, a prototype-based credal classification (PCC) method is developed for the incomplete patterns based on belief function theory. In PCC, each class prototype vector obtained from training data are respectively used to estimate the missing values in the pattern. For a c-class problem, the object with each of the c estimations can be classified by the normal classifier (for complete pattern), and it produces c pieces of classification results with different weighting factors determined by the distances between the object and the prototypes. These results are discounted according to their relative weights. The global fusion of these discounted results is adopted for making credal classification of the object. In the fusion process, if a high conflict occurs, it means that the class of the object is quite uncertain and imprecise only based on the known attributes information, and conflicting beliefs will be committed conditionally to the selected meta-class. So PCC method also allows the incomplete pattern to belong not only to specific classes, but also to meta-classes with different masses of belief. The meta-class is introduced to characterize the imprecision of classification due to the missing values, and it can also reduce errors. Once an object is committed to a meta-class, it means that the specific classes included in the meta-class seem undistinguishable for this object based on the known attributes. If one wants to get more precise result, some other (possibly costly) techniques or information sources must be developed and used. Some results of experiments with artificial and real data sets are also given in Chapter 5 to illustrate the effectiveness of PCC.

When the training information is not available in the classification, a clustering technique must be applied. Fuzzy C-means (FCM) remains the most popular clustering method. In Chapter 6, FCM is extended under belief functions framework to well characterize the uncertainty and imprecision of information, and the new clustering method is called credal c-means (CCM). CCM working with credal partition can produce three kinds of clusters: singleton clusters, meta-clusters and outlier cluster. The belief of specific class is proportional to the distance of the object to the center of this center, and the smaller distance leads to the bigger belief degree. The belief of meta-class depends on both the distance of object to the meta-class center defined by the simple mean value of the centers of the involved specific classes, and the distances to centers of all the involved specific classes, since the object in the meta-class should be also simultaneously close to these classes included in the meta-class. The object too far from the others according to the given outlier threshold will be naturally considered as outliers. The objective function is defined according to this basic principle. Moreover, a meta-cluster threshold is introduced in CCM to eliminate the meta-clusters with big cardinality, and to reduce the computational burden. The clustering centers and bba's matrix can be derived by the optimization (minimization) of the objective function. If one object is considered in the meta-cluster, it means that the singleton clusters included in the meta-cluster appear indistinguishable for the object. This indicates the used information is not sufficient for making the specific classification of these objects in meta-clusters, and these objects should be treated more cautiously. CCM can effectively reduce the misclassification errors thanks to the meta-clusters, and it is also robust to the outliers. The credal partition can be easily approximated to a fuzzy partition if necessary, and the transformation rule is also given in Chapter 6. The use and potential of CCM is demonstrated through different experiments with real data sets.

A summary of the four new methods developed in this thesis is given in Table1.1 to show their purposes and working conditions for convenience.

Table 1.1 : The summary of the four proposed methods

method	purpose	conditions
BCKN (chap. 3)	Credal classification of uncertain data using the K-nearest neighbors in each class	Different classes are overlapped in complicate cases, and the data set cannot be too big.
CCR (chap. 4)	Direct computation of BBA's of the object belonging to different classes with low computational burden	Each class can be represented by one prototype, and CCR can deal with the large scale data set.
PCC (chap. 5)	Credal classification of incomplete data with estimation of missing values using prototypes of classes	The pattern to classify contains missing values that cannot be exactly determined.
CCM (chap. 6)	Credal clustering of uncertain data with admissible computational burden controlled by a given parameter	No training data is available, and different clusters can be partly overlapped.

2 Literature overview

2.1 Introduction

The credal classification of uncertain data based on belief function theory is studied in this thesis. In this chapter, we propose an overview of the main related research works available in the literature. At first, the belief function theory, which is an important and efficient tool we used for classification, is briefly introduced, with the basic definitions, the combination rules, and so on. Then, the traditional data classification methods for dealing with different cases (i.e. supervised classification, incomplete pattern classification, clustering analysis) are recalled. When the training information is available for classification, the supervised classifier can be applied, and a brief survey of supervised methods is presented. Particularly, an evidential K-nearest neighbor (EK-NN) [12] classifier is reviewed with some comments on its limitations. Our new BCKN method, inspired by EK-NN, to overcome its limitations will be presented in next chapter. The missing attribute data is often encountered in classification problem, and the normal methods for classifying the incomplete patterns are also summarized here. If the training samples are unavailable for making the classification, the data clustering analysis must be done. That is why the well known clustering methods like Fuzzy c-means [39] and the evidential c-means (ECM) clustering method [23] are also recalled in this chapter. The latter is closely related to our proposed method CCM (Credal c-means) that will be presented in details in Chapter 6.

2.2 Brief introduction of belief function theory

2.2.1 Basics of belief function theory

The belief functions have been introduced in 1976 by Shafer in his mathematical theory of evidence, known also as belief function theory or Dempster-Shafer theory (DST) [3–7] because Shafer uses Dempster's fusion rule for combining belief basic assignments. We consider a finite discrete set $\Omega = \{w_1, w_2, \ldots, w_c\}$. Ω of $c > 1$ mutually exclusive and exhaustive hypotheses, which is called the *frame of discernment* (FoD) of the problem under consideration. The power-set of Ω denoted by 2^Ω contains all the subsets of Ω. The singleton elements in 2^Ω represent the specific classes, and the set of elements (i.e. disjunction of elements) can be considered as meta-class. For example, if $\Omega = \{w_1, w_2, w_3\}$, then $2^\Omega = \{\emptyset, w_1, w_2, w_3, w_1 \cup w_2, w_1 \cup w_3, w_2 \cup w_3, \Omega\}$. The union $w_i \cup w_j = \{w_i, w_j\}$ is interpreted as the proposition "the truth value of unknown solution of the problem under concern is either in w_i or in w_j". So that Ω represents the full ignorance (uncertainty).

Glenn Shafer [3] considers the subsets as propositions in the case we are concerned with the true value of some quantity ω taking its possible values in Ω. Then the propositions $\mathcal{P}_w(A)$ of interest are those of the form[1]: $\mathcal{P}_w(A) \triangleq$ *The true value of* w *is in a subset* A *of* Ω. Any proposition $\mathcal{P}_w(A)$

[1] We use the symbol $\triangleq$ to mean *equals by definition*; the right-hand side of the equation is the definition of the

is thus in one-to-one correspondence with the subset A of Ω. Such correspondence is very useful since it translates the logical notions of conjunction $\wedge$, disjunction $\vee$, implication $\Rightarrow$ and negation $\neg$ into the set-theoretic notions of intersection $\cap$, union $\cup$, inclusion $\subset$ and complementation $c(.)$. Indeed, if $\mathcal{P}_w(A)$ and $\mathcal{P}_w(B)$ are two propositions corresponding to subsets A and B of Ω, then the conjunction $\mathcal{P}_w(A) \wedge \mathcal{P}_w(B)$ corresponds to the intersection $A \cap B$ and the disjunction $\mathcal{P}_w(A) \vee \mathcal{P}_w(B)$ corresponds to the union $A \cup B$. A is a subset of B if and only if $\mathcal{P}_w(A) \Rightarrow \mathcal{P}_w(B)$ and A is the set-theoretic complement of B with respect to Ω (written $A = c_w(B)$) if and only if $\mathcal{P}_w(A) = \neg\mathcal{P}_w(B)$. In other words, the following equivalences are then used between the operations on the subsets and on the propositions: (intersection$\equiv$conjunction), (union$\equiv$ disjunction), (inclusion$\equiv$ implication) and (complementation$\equiv$ negation).

A basic belief assignment (BBA) is a function $m(.)$ from 2^Ω to $[0, 1]$ satisfying

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

The subsets A of Ω such that $m(A) > 0$ are called the *focal elements* of $m(.)$, and the set of all its focal elements is called the *core* of $m(.)$. If A is a singleton element corresponding to specific class, the quantity $m(A)$ can be interpreted as the exact belief committed to the class A. $m(A \cup B)$ reflects the imprecision (non-specificity or ambiguity) degree between the class A and B for the classification of the object.

A *Credal partition* [23, 24] for a data clustering over the frame Ω is defined as n-tuple $M = (\mathbf{m_1}, \ldots, \mathbf{m_n})$, where $\mathbf{m}_i$ is the basic belief assignment of the sample $\mathbf{x}_i \in X$, $i = 1, \ldots, n$ associated with the different elements of the power-set 2^Ω. So the

The belief function $Bel(.)$ and the plausibility function $Pl(.)$ [3], are usually interpreted as lower and upper probabilities of the hypothesis. They are mathematically defined for any $A \in 2^\Omega$ from a given BBA $m(.)$ by

$$Bel(A) = \sum_{B \in 2^\Omega | B \subseteq A} m(B) \tag{2.2}$$

$$Pl(A) = \sum_{B \in 2^\Omega | A \cap B \neq \emptyset} m(B) \tag{2.3}$$

Shafer [3] has proposed to use Dempster's fusion rule to combine several distinct bodies of evidence characterized by different BBA's in the development of DST. This rule will be denoted DS (Dempster-Shafer) rule for short in the sequel. Mathematically, the DS rule of combination of two BBA's $m_1(.)$ and $m_2(.)$ defined on 2^Ω is defined by $m_{DS}(\emptyset) = 0$ and for $A \neq \emptyset \in 2^\Omega$ by

$$m_{DS}(A) = \frac{\sum_{B,C \in 2^\Omega | B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B,C \in 2^\Omega | B \cap C = \emptyset} m_1(B) m_2(C)} = \frac{\sum_{B,C \in 2^\Omega | B \cap C = A} m_1(B) m_2(C)}{\sum_{B,C \in 2^\Omega | B \cap C \neq \emptyset} m_1(B) m_2(C)} \tag{2.4}$$

In DS rule, the total conflicting belief mass $\sum_{B,C \in 2^\Omega | B \cap C = \emptyset} m_1(B) m_2(C)$ is redistributed back to all the focal elements through the normalization. In the combination of high conflicting sources of evidence, DS rule will produce very unreasonable results, and it is better to not use it in such cases [40–44]. It has also been proved recently that DS rule suffers also of a very serious flaw even in low conflicting cases for very specific belief structures [45, 46].

In DS rule, each source of evidence is treated equally. When the reliability of each source of evidence is different, then the corresponding BBA's should be discounted before entering the fusion

left-hand side.

process. The classical discounting method has been introduced by Shafer in [3], and it is given by

$$\begin{cases} m'(A) = \alpha \cdot m(A), \quad A \neq \Omega \\ m'(\Omega) = 1 - \sum_{\substack{A \in 2^\Omega \\ A \neq \Omega}} m'(A) \end{cases} \tag{2.5}$$

where $\alpha \in [0, 1]$ is discounting factor of $m(\cdot)$. One can see that the discounted information is all assigned to the ignorance element Ω. If the evidence is totally reliable, i.e. $\alpha = 1$, then the discounting factor has no influence on the BBA's at all. Whereas, if the evidence is totally unreliable, i.e. $\alpha = 0$, the discounted BBA will become $m(\Omega) = 1$, and it indicates this evidence plays a neutral role in the fusion process.

One alternative discounting method called contextual discounting operation [47] has been also developed by Mercier, et al., and it allows the user to quantify the confidence in the reliability of the source, conditionally on different hypotheses regarding the variable on interest. The discounting procedure is not controlled by a single discount rate as done in previous classical discounting method, and the sources of evidence are discounted by a vector denoting the expected reliability of the source in different contexts. This discounting approach is more precise but also much more complicate, which is not convenient for the real applications.

2.2.2 Several alternative combination rules

Many experts and users of belief functions considers that the problem of DS rule mainly lies in the redistribution of conflicting beliefs [4, 8–10], which cannot be done according to the normalization procedure used in this rule. So many alternative rules of combination [4, 8–11, 48–50] working with Shafer's model have been proposed by modifying the distribution ways in order to palliate the drawbacks of DS rule. Several well known rules will be briefly recalled here.

- **Yager's rule** [8]:

In Yager's rule, the conflicting information is considered totally uncertain, and it should be committed to the ignorance element denoted by Ω.

$\forall A \neq \emptyset, B, C \in 2^\Omega$

$$\begin{cases} m_Y(\emptyset) = 0 \\ m_Y(A) = \sum\limits_{B \cap C = A} m_1(B) m_2(C) \\ m_Y(\Omega) = m_1(\Omega) m_2(\Omega) + \sum\limits_{B \cap C = \emptyset} m_1(B) m_2(C) \end{cases} \tag{2.6}$$

Yager's rule is a quite prudent and even pessimistic rule, since the mass of belief on ignorance element will increase with the combination of more sources of evidence, which makes the fusion results become more and more ignorant.

- **Dubois &Prade (DP) rule** [9]:

Dubois and Prade have proposed another rule, denoted DP rule in [9]. When there is no conflict between two sources of evidence, both evidence are considered as reliable. If some conflict occurs, one of the sources is considered as reliable whereas the other one is considered as unreliable. In other words, if the hypothesis A is considered true in one source, whereas the other source states

that B is true, then the intersected element $A \cap B \neq \emptyset$ will be identified as truth. However, if it holds that $A \cap B = \emptyset$, the truth will be considered in the set $A \cup B$. So DP rule transfers each partial conflicting mass to the union of the elements involved in the partial conflict. Mathematically, DP rule is defined by $m_{DP}(\emptyset) = 0$ and for $A \in 2^\Omega \setminus \{\emptyset\}$ by

$$\begin{cases} m_{DP}(\emptyset) = 0 \\ m_{DP}(A) = \sum\limits_{B,C \in 2^\Omega | B \cap C = A} m_1(B) m_2(C) + \sum\limits_{\substack{B,C \in 2^\Omega \\ B \cap C = \emptyset \\ B \cup C = A}} m_1(B) m_2(C) \end{cases} \tag{2.7}$$

- **Smets combination rule** [5, 10]:

In TBM, Smets allows to assign the mass of belief to empty set $\emptyset$, and thus he extends the close world in Shafer's model to an open world. In fact, Smets rule is the un-normalized version of DS rule, and it commits all the conflicting beliefs to the empty set $\emptyset$ contrariwise to what is done in all other rules of combination. Smets rule is defined as: $\forall A, B, C \in 2^\Omega$,

$$\begin{cases} m_S(\emptyset) = k_{12} = \sum\limits_{B \cap C = \emptyset} m_1(B) m_2(C) \\ m_S(A) = \sum\limits_{B \cap C = A} m_1(B) m_2(C), A \neq \emptyset \end{cases} \tag{2.8}$$

These combination rules work for the the fusion of the cognitively independent sources of evidence,and they can be easily extended for the fusion of more than two sources if necessary. A combination method for combining non-distinct sources of evidence has also been introduced in [51].

2.2.3 A brief review of DSmT

The purpose of Dezert-Smarandache Theory (DSmT) [4, 52, 53] is to overcome the limitations of DST [3] mainly by proposing new underlying models for the frames of discernment in order to fit better with the nature of real problems, and proposing new efficient combination and conditioning rules. In DSmT framework, the elements ω_i, $i = 1, 2, \ldots, n$ of a given frame Ω are not necessarily exclusive, and there is no restriction on ω_i but their exclusivity. The hyper-power set D^Ω in DSmT [54] is defined as the set of all composite propositions built from elements of Ω with operators $\cup$ and $\cap$. For instance, if $\Omega = \{\omega_1, \omega_2\}$, then $D^\Omega = \{\emptyset, \omega_1, \omega_2, \omega_1 \cap \omega_2, \omega_1 \cup \omega_2\}$. A (generalized) basic belief assignment (BBA for short) is defined as the mapping $m : D^\Omega \to [0, 1]$. The generalized belief and plausibility functions are defined in almost the same manner as in DST.

Two models[2] (the free model and hybrid model) in DSmT can be used to define the BBA's to combine. In the free DSm model, the sources of evidence are combined without taking into account integrity constraints, and its combination rule is given by

$$m_f(A) = \sum_{\substack{B,C \in D^\Omega \\ B \cap C = A}} m_1(B) m_2(C), \forall A \in D^\Omega \tag{2.9}$$

The intersected elements are kept since the elements are not necessarily exclusive here.

[2]Actually, Shafer's model, considering all elements of the frame as truly exclusive, can be viewed as a special case of DSmT hybrid model.

When the free DSm model does not hold because the true nature of the fusion problem under consideration, we can take into account some known integrity constraints and define BBA's to combine using the proper hybrid DSm model. Many combination rules have been developed in hybrid DSm model, such as DSmH, PCR1-PCR6, and these rules are all fit for the Shafer's model. Particularly, PCR6 is considered with the most precise distribution way of conflicting beliefs, and it is defined by:

$$m_{PCR6}(A) = \sum_{\substack{B,C\in D^\Theta \\ B\cap C=A}} m_1(B)m_2(C) + \sum_{\substack{X,Y\in D^\Theta \\ X\cap Y=\emptyset}} [\frac{m_1(A)^2 m_2(X)}{m_1(A)+m_2(X)} + \frac{m_2(A)^2 m_1(Y)}{m_2(A)+m_1(Y)}], \forall A \in D^\Theta \tag{2.10}$$

The formula of the combination rule for multiple (more than two) sources of evidence by PCR6 is also given in [55, 56].

The complexity of DSmT is quite high especially when the number of elements in the frame of discernment is big. For example, if one has $|\Omega| = n$, then the number of the elements in D^Ω follows the Dedkind's sequence. So the proper rule working with Shafer's model is usually chosen for convenience when the elements in FoD are exclusive in the applications.

2.2.4 Decision making support

Let us consider that $\mathbf{m}$ is a BBA obtained by the fusion of multi-source of evidence for one object belonging to different focal elements (classes). The decision of class can be directly made according to the criterion that the class of the object should receive the biggest mass of belief, and this is called hard credal partition [23]. Then the object can be committed to either a specific (singleton) class, or the imprecise meta-class (i.e. the union of several classes). Several alternative decision rules that also allows the object to belong to imprecise class have been introduced in [57].

The probability transformation, which can approximate a BBA to the probabilistic (fuzzy) measure, is often used for making the hard classification decision. If so, the masses of belief on the meta-classes must be redistributed to the other specific classes by a chosen method. There exist many methods to transform the BBA to the probabilistic measures, such as the pignistic transformation method $BetP(.)$ [5–7], the plausibility transformation method [58], or more sophisticate method [53]. Particularly, the well known pignistic probability transformation $BetP(.)$ introduced by Smets in his transferable belief model will be briefly presented here, since it is often used in many applications.

$BetP(A)$ is defined for $A \in 2^\Omega \setminus \{\emptyset\}$ by

$$BetP(A) = \sum_{B\in 2^\Omega, A\subseteq B} \frac{|A\cap B|}{|B|} m(B) \tag{2.11}$$

where $|X|$ is the cardinality of the element X (i.e. the number of the singleton elements in X, for example $|w_1 \cup w_2| = 2$ where w_1 and w_2 are singleton classes.). Other transformations exist to approximate a BBA into a probability measure but they are more complex to implement and so we suggest to use $BetP(.)$ for decision-making support if the computational burden is a strong constraint (like in real-time military classification applications).

2.3 Supervised classification of data

For the supervised classification, the classifiers broadly belong to two families: model-based classifiers and case-based classifiers [13]. Bayes theorem is usually applied in the model based classifiers

to calculate the posterior probability estimates according to the class-conditional densities and prior probabilities. However, the complete statistical knowledge of the conditional density functions of each class is hard to obtain in many cases, and this precludes the application of model-based classification procedure. When the model-based classifiers become unavailable, one can choose the case-based classifier computing the class probabilities using the correctly labeled training data set without considering the density estimation.

There exist many well known case-based classifiers, typically the support vector machine (SVM) [59, 60], the artificial neuron network (ANN) [61], decision trees [62] and the K-nearest neighbor (K-NN) classifier [63], etc. Support vector machine allows to construct a hyperplane or set of hyperplanes in a high-dimensional space using the training data set for classification, regression, or other tasks. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), because the larger the margin generally leads to the lower error rate of the classifier. Artificial neuron network (ANN) is defined by a set of input neurons which may be activated by the input data. After being weighted and transformed by a function determined by the user, the activations of these neurons are then passed on to other neurons. This process is repeated until an output neuron is activated. Decision trees are used as a predictive model mapping observations about an item to conclusions about target value, and they consist of two main types: classification and regression tree (CART) as an umbrella term. In the tree structure, each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome, and each leaf node represents a class label. A tree can be learned by splitting the source set into subsets based on an attribute value test, and the process is repeated on each derived subset in a recursive manner until the splitting cannot add value to the predictions any more. K-nearest neighbor (K-NN) rule is a simple but efficient non-parametric procedure, and it remains an interesting topic of research. In K-NN, an unclassified sample is classified into the class which the majority of its K nearest neighbors (KNNs)[3] in the training set belong to.

2.3.1 Evidential classification

These traditional classifiers generally work with the probability measure, like ANN, and the imprecise (ignorance) information is not considered in the classification process. The belief function theory has been already applied for the data classification [12–17, 19–23, 38, 64–70], data clustering [23,24,71,72], and imprecise decision-making support [18,30,57,73–75] to model the uncertainty and imprecision.

Some data classifiers have already been developed based on belief functions in the past. For instance, Smets [76] and Appriou [77] have proposed the model-based classifier based on the Generalized Bayes Theorem (GBT) [76] which is an extension of Bayes theorem in Smets transferable belief model (TBM) [5–7]. There are some other case-based evidential classifiers based on neural network [38], K-nearest neighbors [12], decision trees [78], and SVM [64]. Particularly, the evidential version of K-nearest neighbors method (EK-NN) has been proposed in [12] based on DST, for working only with the specific classes and the extra ignorant class defined by the union of all the specific classes. In EK-NN, K BBA's can be determined according to the distance between the object and the selected K nearest neighbors, and each BBA has two focal elements: singleton class (i.e. the label of the corresponding neighbor) and ignorance class (i.e. the frame of discernment). The combination of the K BBA's by DS rule is used for the classification of the object. A fuzzy version of EK-NN, denoted FEK-NN, has been also developed in [79]. An ensemble technique for the combination of evidential K-NN classifier based on DST has been proposed in [80] to improve the accuracy. A neural network classifier has also been developed in [38] under the belief functions

[3] In this thesis, K-NN denotes the K-nearest neighbor classifier, whereas KNNs represents the K nearest neighbors.

framework that allows one extra ignorant class as possible output of this classifier, and it can reduce the computation burden with respect to EK-NN but it requires more complicate training process. The relationship between the case-based and model-based approaches working with belief functions have also been studied, and it shows that both methods actually proceed from the same underlying GBT principle, and that they essentially differ by the nature of the assumed available information. This is helpful for us to choose the most appropriate method for the particular application depending on the nature of the information.

2.3.2 Brief recall and comments on EK-NN

Let us consider a group of objects $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ to classify over the frame of the classes $\Omega = \{w_1, \cdots, w_h\}$. In EK-NN [12], the imperfect information is modeled by the ignorant focal element $w_1 \cup \ldots \cup w_i \cup \ldots \cup w_h$ denoted also by Ω, and the BBA of the object $\mathbf{x}_i$ associated to its close neighbor $\mathbf{x}_j$ labeled by $w_s \in \Omega$ is defined as:

$$m_j^{\mathbf{x}_i}(w_s) = \alpha e^{-\gamma_s d_{ij}^{\beta}} \tag{2.12}$$

$$m_j^{\mathbf{x}_i}(\Omega) = 1 - \alpha e^{-\gamma_s d_{ij}^{\beta}} \tag{2.13}$$

As recommended in [12], the default value of the discounting factor α can be taken to 0.95, and the parameter γ_s must be chosen inversely proportional to the mean distance between two training data belonging to class w_s. The parameter β usually takes a small value because it has in fact a very little influence on the performance of EK-NN, one takes $\beta = 1$ as its default value. Generally, d_{ij} corresponds to the Euclidean distance between the object $\mathbf{x}_i$ and the training data $\mathbf{x}_j$. K BBA's corresponding to the K nearest neighbors of the object are constructed according to Eqs. (2.12)-(2.13). From Eq. (2.12), one can see that the bigger distance d_{ij} will yield smaller masses of belief on the corresponding class w_s. It means that if the object $\mathbf{x}_i$ is far from the neighbor labeled by class w_s, this neighbor can provide little support to $\mathbf{x}_i$ belonging to the corresponding class w_s.

From Eq. (2.12) and (2.13), one sees that only the two focal elements w_s and Ω are involved in a given BBA $m_j^{x_i}(.)$, $i = 1, \ldots, n$. The classification result of each object $\mathbf{x}_i$ is obtained by the fusion of the K BBA's using DS rule given in Eq. (2.4). Because of the very particular structure of the BBA's, DS rule produces as focal elements of the fusion only a specific class w_s and the total ignorance class Ω. Therefore, no partial ignorance (meta-classes), say $w_i \cup \cdots \cup w_j \equiv \{w_i, \cdots, w_j\}$, can become a focal element in EK-NN method.

It is worth to note that DS rule for the fusion of the K bba's with such particular structure is not very effective for the outlier detection. Indeed, for any outlier very far from its KNNs having same class label, the most mass of belief will be committed to the specific class after the fusion process using DS rule (when K is big enough). This behavior is abnormal since the ignorance class Ω should normally take a larger mass of belief when the object corresponds to a true outlier. Because of this DS behavior, this object will be incorrectly considered as belonging to a specific class rather than to the outlier class. This behavior is not satisfactory in some real applications, like in target tracking in cluttered environments. This behavior is clearly illustrated in the following simple example.

Example 2.1: Let's assume that an object $\mathbf{x}$ is located very far from all the training data (so that x must reasonably be considered as a true outlier), and assume that all its KNNs belong to the class w. The biggest distance between $\mathbf{x}$ and the K neighbors is d, and the corresponding BBA's are $m_k(w) = \delta$ and $m_k(\Omega) = 1 - \delta$ with $\delta \in (0, 1)$ for $k = 1, 2, \ldots, K$. In such case, the combination of these K BBA's with DS rule gives $m(w) \geq 1 - (1 - \delta)^K$ and $m(\Omega) \leq (1 - \delta)^K$. This

result indicates clearly that the belief on the specific class $m(w)$ increases when the number K of nearest neighbors increases. So even if the value $\delta > 0$ is very small, the belief committed to w can become very large when K is big enough. It can be easily verified that when $K > \frac{\log 0.5}{\log(1-\delta)}$, one has $m(w) > 0.5 > m(\Omega)$. For example, if one takes $\delta = 0.1$, and $K > 6$, then $m(w) > 0.5 > m(\Omega)$. This inappropriate result is due to the DS combination of such particular structure of BBA's.

Moreover, the EK-NN method appears not very effective to reveal the imprecision degree of the objects that belong to different classes, especially for the objects lying in the overlapped zone of different classes, as shown in the following example.

Example 2.2: Let us consider that an object $\mathbf{x}$ lies in the partially overlapped zone of two classes w_1 and w_2. We select $K = 2p$ nearest neighbors for the classification, and we assume that p neighbors labeled by w_1 are at the same distance d_1 of $\mathbf{x}$, and the other p neighbors labeled by w_2 are also at the same distance d_2 of $\mathbf{x}$. If d_1 is quite close to d_2, the object very likely belongs to w_1 and w_2 with similar degrees of belief, and the corresponding bba's are given by: $m_i(w_1) = \alpha, m_i(\Omega) = 1-\alpha, i = 1, \dots p$ and $m_j(w_2) = \beta, m_j(\Omega) = 1-\beta, j = p+1, \dots 2p$ with $\alpha = \beta + \epsilon$ (ϵ being a very small positive value). The fusion results of the $2p$ BBA's using DS rule are obtained by :

$$m(w_1) = \frac{[1-(1-\alpha)^p](1-\beta)^p}{(1-\alpha)^p + (1-\beta)^p - (1-\alpha)^p(1-\beta)^p}$$
$$m(w_2) = \frac{[1-(1-\beta)^p](1-\alpha)^p}{(1-\alpha)^p + (1-\beta)^p - (1-\alpha)^p(1-\beta)^p}$$
$$m(\Omega) = \frac{(1-\alpha)^p(1-\beta)^p}{(1-\alpha)^p + (1-\beta)^p - (1-\alpha)^p(1-\beta)^p}$$

and therefore,

$$m(w_1) - m(w_2) = \frac{(1-\beta)^p - (1-\alpha)^p}{(1-\alpha)^p + (1-\beta)^p - (1-\alpha)^p(1-\beta)^p}$$

One clearly sees that $m(w_1)$ can be much bigger than $m(w_2)$ when p is sufficiently large, although the object has the same number of neighbors in each class, and it also has the very close masses α and β of belief committed to each. For instance, if one takes $p = 7$, $\alpha = 0.96$ and $\beta = 0.95$, which indicates that the class of $\mathbf{x}$ is quite uncertain and imprecise between class w_1 and w_2, then $m(w_1) = 0.8266$ and $m(w_2) = 0.1734$ with $m(w_1) - m(w_2) = 0.6532$. If the value of p increases to $p = 8$, one gets $m(w_1) = 0.8563$ and $m(w_2) = 0.1437$ with $m(w_1) - m(w_2) = 0.7126$. Hence, the object $\mathbf{x}$ will be associated to class w_1 with a strong belief according to the fusion results, and $m(w_1)$ becomes bigger when p increases. This result is abnormal because we know that this object should better belong to w_1 or w_2 because its class remains very imprecise based only on the available original BBA's. This inappropriate DS fusion result cannot reveal the imprecision of the class of this object, and will yield misclassification errors. This behavior is due to DS rule used for the fusion of such particular structure of BBA's.

The limitations of EK-NN classifier have been shown through the two examples, and the new credal classifiers will be studied based on belief function theory in this thesis to better deal with the uncertain and imprecise data.

2.4 Data clustering

Data clustering is a kind of unsupervised classification, and its purpose is to group a set of objects in such a way that objects in the same group (i.e. cluster) are more similar to each other than to those in other groups (i.e. clusters). It can deal with object data and relational data. The object

data is represented by vector composed by the numeric attributes of the object, whereas, relational data denotes the pairwise similarity or dissimilarity measurements of the objects. The clustering techniques mainly include hierarchical clustering [81–83] where objects belong to a child cluster and also belong to the parent cluster, hard clustering [84] where each object belongs to exactly one cluster and soft (fuzzy) clustering [39, 85–87] that allows the object to belong to different clusters with different probabilities (fuzzy membership). K-means clustering [84] is a very popular approach for cluster analysis in data mining, and it aims to partition the objects into K clusters in which each object belongs to the cluster with the nearest mean, serving as a prototype of the cluster. In K-means clustering, the object belongs completely to only one cluster, and it is a hard clustering method, whereas fuzzy c-means clustering [39] can be considered as the soft version of K-means, and it allows that each object has a fuzzy degree of belonging to each cluster. Fuzzy c-means (FCM) remains a very well known clustering method and it is an important tool for pattern recognition and data mining.

FCM [39] seeks for c clustering centers of the data set, and minimizes the sum of weighted distances between the object and the centers. It is based on minimization of the following objective function:

$$J_{FCM}(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{\beta} d_{ij}^{2} \tag{2.14}$$

with subject to the constraint:

$$\begin{cases} \sum\limits_{k=1}^{c} u_{ik} = 1, i = 1, \ldots, n. \\ \sum\limits_{i=1}^{n} u_{ik} > 0, k = 1, \ldots, c. \end{cases} \tag{2.15}$$

where β is any real number greater than 1, u_{ij} is the degree of membership of the object $\mathbf{x}_i$ in the cluster j and d_{ij} is the distance between the object $\mathbf{x}_i$ and the center $\mathbf{v}_j$.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers $\mathbf{v}_j$ defined by:

$$\mathbf{v}_k = \frac{\sum\limits_{i=1}^{n} u_{ik}^{\beta}\mathbf{x}_i}{\sum\limits_{i=1}^{n} u_{ik}^{\beta}}, \forall k = 1, \ldots, c. \tag{2.16}$$

$$\mathbf{u}_{ij} = \frac{d_{ij}^{-2/(\beta-1)}}{\sum\limits_{k=1}^{c} d_{ik}^{-2/(\beta-1)}}, \forall i = 1, \ldots, n; \forall j = 1, \ldots, c. \tag{2.17}$$

The algorithm is composed of the following steps.

This iteration will stop when $||U(s+1) - U(s)|| < \epsilon$, and this procedure converges to a local minimum or a saddle point of $J_{FCM}(U,V)$.

Table 2.1 : Fuzzy c-means procedure

Input:	Training data: $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ in $\mathbb{R}^p$
Parameters:	c: number of clusters
	ϵ: termination criterion
Initialization:	U=$[u_{ij}]$ matrix, U(0)
	for s=1 to k
	Calculate the centers vectors $\mathbf{v}_k$ with U(s) using eq. (2.16);
	Update U(s+1)using eq. (2.17)
	If $\|\|U(s+1) - U(s)\|\| < \epsilon$ then STOP.
	end

2.4.1 Credal partition using belief functions

A concept of partition named credal partition [25, 26] has been recently proposed by Denœux and Masson for data clustering under belief functions framework to well model the uncertain and imprecise information. Credal partition can be considered as an extension of the existing concepts of hard [84], fuzzy [39] and possibilistic partition [85], since it allows that the objects belong to not only the singleton clusters in $\Omega = \{w_1, \dots w_c\}$ but also any subsets of Ω (i.e. meta-clusters) with different masses of belief. It has been reported in [23] that this additional flexibility of credal partition is able to gain a deeper insight in the data and to improve robustness with respect to outliers. An EVidential CLUStering (EVCLUS) [24] algorithm working with credal partition has been developed for relational data. In EVCLUS, each object is assigned a BBA over a given set of classes, and the degree of conflict between two BBA's is used to reflect the dissimilarity between the corresponding objects. The bigger dissimilarity corresponds to the higher conflict degree. A recent constrained clustering method called CEVCLUS [72] based on the EVCLUS algorithm has been proposed in the theoretical framework of belief functions, and it is designed for dissimilarity data for taking into account the background knowledge in form of pairwise constraints. An evidential EM algorithm [71] has been recently developed for the parameter estimation in statistical models when the uncertainty on the data can be modeled by belief functions. Evidential C-Means (ECM) [23] clustering method inspired from FCM [39] and Noise-Clustering (NC) algorithm [88–90] were also proposed for the credal partition of object data.

2.4.2 Brief review of Evidential C-Means (ECM)

ECM [23] working with credal partition can produce three kinds of cluster: singleton (specific) clusters (e.g. w_i), meta-clusters (e.g. $w_j \cup \cdots \cup w_k$) defined by disjunction of several singleton clusters and outlier cluster represented by $\emptyset$. Each cluster (e.g. $w_i, i = 1, \dots, c$) corresponds to one clustering center (prototype) (e.g. $\mathbf{v}_i, i = 1, \dots, c$), and the meta-cluster's center is obtained by the arithmetic mean value of the prototype vectors of the singleton clusters included in the meta-cluster. The belief on each cluster mainly depends on the distance between the object and the corresponding clustering center taking into account the cardinality of the cluster.

Let us consider a finite and discrete set of objects $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ to be clustered over a given frame of discernment $\Omega = \{w_1, w_2, \dots, w_c\}$ with $|\Omega| = c$, and each data point is a p-dimension vector as $\mathbf{x}_i = (x_{i_1}, \cdots, x_{i_p})$. In ECM, the class membership of an object $\mathbf{x}_i$ is represented by a BBA $m_i(.)$ over the power-set 2^Ω, and 2^Ω contains $2^{|\Omega|}$ elements (clusters). This representation is able to model all situations ranging from complete ignorance to full certainty concerning the class of $\mathbf{x}_i$.

ECM [23] is a direct extension of FCM based on credal partition. The mass of belief for

associating the object $\mathbf{x}_i$ with an element A_j of 2^{Ω} denoted by $m_{ij} \triangleq m_{\mathbf{x}_i}(A_j)$, is determined from the distance d_{ij} between $\mathbf{x}_i$ and the prototype vector $\bar{\mathbf{v}}_j$. Note that A_j can either be a singleton cluster, or a meta-cluster. The prototype vector (center) $\bar{\mathbf{v}}_j$ of A_j, is defined as the mean value of the singleton clusters included in A_j. $\bar{\mathbf{v}}_j$ is defined mathematically by

$$\bar{\mathbf{v}}_j = \frac{1}{|A_j|}\sum_{k=1}^{c} s_{kj}\mathbf{v}_k \quad \text{with} \quad s_{kj} = \begin{cases} 1, & \text{if } w_k \in A_j \\ 0, & \text{otherwise} \end{cases} \tag{2.18}$$

where $\mathbf{v}_k$ is center of the singleton cluster w_k, and $|A_j|$ denotes the cardinality of A_j, and d_{ij} denotes the Euclidean distance between $\mathbf{x}_i$ and $\bar{\mathbf{v}}_j$. In the clustering analysis methods (e.g. FCM, ECM), each singleton cluster (class) is considered with a prototype vector (clustering center) which has the same dimension (i.e. p) of the samples to be clustered. The center of meta-cluster in ECM is calculated by the mean value of the centers of the singleton clusters included in this meta-cluster. For instance, let us consider that a 3-class data set that has to be clustered over the frame of discernment $\Omega = \{w_1, w_2, w_3\}$, where the center of each singleton cluster (i.e. w_1, w_2, w_3) is respectively denoted by a vector $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_3$. The center of the meta-cluster $A \triangleq w_i \cup w_k; i = 1,2,3; k = 1,2,3; i \neq k$ is given by $\bar{\mathbf{v}}_A = \frac{\mathbf{v}_i+\mathbf{v}_k}{|A|} = \frac{\mathbf{v}_i+\mathbf{v}_k}{2}$, and the center of the bigger meta-cluster $B \triangleq w_1 \cup w_2 \cup w_3$ can be obtained by $\bar{\mathbf{v}}_B = \frac{\mathbf{v}_1+\mathbf{v}_2+\mathbf{v}_3}{|B|} = \frac{\mathbf{v}_1+\mathbf{v}_2+\mathbf{v}_3}{3}$.

In ECM, the mass of belief of the object belonging to each cluster (i.e. singleton cluster or meta-cluster) is considered proportional to the distance between the object and the corresponding clustering center, and the bigger distance leads to the smaller mass of belief, and the objective function is defined by

$$J_{ECM} = \sum_{i=1}^{n}\sum_{A_j \subseteq \Omega, A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta} \tag{2.19}$$

Subject to

$$\sum_{A_j \subseteq \Omega, A_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1 \tag{2.20}$$

The mass of belief $m_{ij} \triangleq m_{\mathbf{x}_i}(A_j)$ is obtained by the minimization of this objective function. In fact, the objective function is the sum of weighted distances between the objects and each cluster center. The minimization of eq. (2.19) under the constraint eq. (2.20) ensures that the object is as close as possible to the center of cluster it belongs to with highest mass of belief (weight), and the mass of belief on the cluster far from the object is small. This is similar to what is done in the FCM approach. Moreover, the mass of belief of the outlier cluster should be small when the chosen outlier threshold is big with respect to the distances of the object to the other centers.

The solution of the minimization of (2.19) under the constraint (2.20) has been established by Masson and Denœux in [23] and it is given for each object $\mathbf{x}_i$, $(i = 1, 2, \ldots, n)$ by:

- For all $A_j \subseteq \Omega$ and $A_j \neq \emptyset$,

$$m_{ij} = \frac{|A_j|^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{A_k \neq \emptyset} |A_k|^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}} \tag{2.21}$$

  where α is a tuning parameter allowing to control the degree of penalization; β is a weighting exponent (its suggested default value in [23] is $\beta = 2$); δ is a given threshold tuning parameter for the filtering of the outliers.

- For $A_j = \emptyset$,

$$m_{i\emptyset} \triangleq m_{\mathbf{x}_i}(\emptyset) = 1 - \sum_{A_j \neq \emptyset} m_{ij}. \tag{2.22}$$

The centers of the class are given by the rows of the matrix $V_{c\times p}$

$$V_{c\times p} = H_{c\times c}^{-1}.B_{c\times p} \tag{2.23}$$

where the elements B_{lq} of $B_{c\times p}$ matrix for $l = 1, 2, \ldots, c$, $q = 1, 2, \ldots p$, and the elements H_{lk} of $H_{c\times c}$ matrix for $l, k = 1, 2, \ldots, c$ are given by:

$$B_{lq} = \sum_{i=1}^{n} x_{i_q} \sum_{w_l \in A_j} |A_j|^{\alpha-1} m_{ij}^{\beta} \tag{2.24}$$

$$H_{lk} = \sum_{i=1}^{n} \sum_{\{w_k, w_l\} \subseteq A_j} |A_j|^{\alpha-2} m_{ij}^{\beta} \tag{2.25}$$

In ECM, the clustering center of meta-cluster is calculated by the arithmetic mean value of the prototype vectors of the singleton clusters included in the meta-cluster. Because of this, some distinct cluster centers can be very close and even overlapped. For example, the center of meta-cluster (e.g. $w_i \cup \cdots \cup w_j$) can be very close to the center of an incompatible singleton cluster (e.g. w_k), and the centers of incompatible meta-clusters (e.g. $w_g \cup \cdots \cup w_h$ and $w_p \cup \cdots \cup w_q$) can also be very close (even overlapped). Moreover, the mass of belief on any cluster (i.e. singleton cluster or meta-cluster) is determined only by the distance between the object and the corresponding center with a tuning parameter α. Therefore, some objects originating from w_k may be wrongly classified into $w_i \cup \cdots \cup w_j$, even if the singleton clusters (i.e. $w_i, \ldots, w_j$) included in $w_i \cup \cdots \cup w_j$ are quite far from each other and clearly separated. If the distinct clusters $w_g \cup \cdots \cup w_h$ and $w_p \cup \cdots \cup w_q$ contain the same number of singleton clusters, they will be considered even undistinguishable in the clustering results because of their close centers, which brings big trouble to associate the objects with these clusters.

This unexpected behavior is illustrated in the following example. Let's consider a simple 4-classes data set and the frame $\Omega = \{w_1, w_2, w_3, w_4\}$ with the corresponding centers $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, $\mathbf{v}_4$. We denote by $\mathbf{v}_{i,j}$ the center of the meta-cluster $w_i \cup w_j$. It is possible that one has $\mathbf{v}_2 \approx (\mathbf{v}_1 + \mathbf{v}_3)/2 = \mathbf{v}_{1,3} \approx (\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3)/3 = \mathbf{v}_{1,2,3}$, but the clusters of w_1 and w_3 are far from each other, which means w_1 and w_3 are fully well separate. Then with ECM, some objects belonging to w_2 are likely to be considered in the incompatible clusters $w_1 \cup w_3$ or the ignorant cluster $w_1 \cup w_2 \cup w_3$ when they are very close to the clustering centers $\mathbf{v}_{1,2}$ or $\mathbf{v}_{1,2,3}$. Since it is still possible that $\mathbf{v}_{2,3} = (\mathbf{v}_2 + \mathbf{v}_3)/2 \approx (\mathbf{v}_1 + \mathbf{v}_4)/2 = \mathbf{v}_{1,4}$, then the meta-clusters $w_1 \cup w_4$ and $w_2 \cup w_3$ cannot be clearly distinguished for the objects in the clustering, although they are incompatible. Some objects belonging to w_1 or w_4 (w_2 or w_3) may be wrongly committed to $w_2 \cup w_3$ ($w_1 \cup w_4$). Such behavior seems very unreasonable and counterintuitive. In fact, there is an infinity of such cases that ECM cannot well deal with.

The relational version of ECM (RECM) [27] is also derived for dealing with relational data. RECM and EVCLUS are compared in [27], and it is pointed that RECM provides similar results to those given by EVCLUS, but the optimization procedure of RECM is computationally much more efficient than the gradient-based procedure of EVCLUS. The constrained ECM (CECM) [28] method has also been recently proposed for taking into account the pairwise constraints information. In our preliminary research work, we developed a method called belief c-means (BCM) [91] to deal with the close clustering centers by introducing an alternative interpretation of the meta-class.

In BCM, meta-class was considered consisting of the objects far from the specific classes included in the meta-class, but much farther to the other class, and this interpretation is quite different from that in ECM. Hence, BCM mainly focused on outliers detection, but the uncertainty of clustering for the objects lying in the overlapped zones of different clusters cannot be well captured in BCM. In this thesis, a new evidential version of fuzzy c-means clustering method will be proposed to well model the uncertainty and imprecision in clustering problem.

2.5 Classification of incomplete data with missing values

In many data classification problems, the quality of the data can suffer from a common drawback that some samples are incomplete feature vectors with missing or unknown attribute values [92–94]. For example, some results can be missing in an industrial experiment due to the mechanical/electronic failures during the data acquisition process. In medical diagnosis, some tests cannot be done when the hospital lacks the necessary medical equipment. In a social survey, the results can be incomplete since respondents may refuse to respond to some questions. Moreover, UCI repository [95] is one very well known data set collection for benchmarking machine learning procedures, but 45% of data sets in the UCI repository contain missing values.

It is important to get a knowledge about how data attributes were missing before one selects the appropriate way to handle incomplete data. The missing data mechanism can be mainly identified by three types [93]:

- Missing completely at random (MCAR): the missing variable is independent of the variable itself and any other external influences,
- Missing at random (MAR): the missing value is independent of the missing variables but may depend on the known values, and it is predictable using other known values in the data set.
- Not missing at random (NMAR): the missing value depends on the missing variable itself, and the missing values cannot be predicted only using the available information in the database.

If the data is missing as in the case of MCAR or MAR, the reasons for missing data in the analysis of the data can be ignored, and it makes the methods used for missing data analysis become simple. Thus, most current research works for dealing with missing data mainly focus on MAR or MCAR cases [92].

There have been a number of methods emerged to classify the incomplete data. The classification of incomplete data with missing values generally concerns two problems including handling missing values and classification procedure. The existing methods can be broadly divided into four groups according to their solutions [92, 93] as follows:

- The incomplete pattern is simply discarded [96]. This is the most simple method, but incomplete pattern with missing values cannot be classified since it will be eliminated. This process usually works just when the incomplete patterns take a small rate (i.e. 5%) in the whole data set.
- Model-based procedures are applied, and the data distributions can be estimated by some methods. For example, the maximum likelihood procedures (e.g. Expectation-Maximization algorithm [97]) can be used to estimate the model parameters. Then, the patterns can be classified based on Bayes theory.

- The missing data is estimated at first and then the incomplete patterns with estimated values are classified. In such case, the handling missing value and pattern classification are treated separately, and this is also the most used procedure. Many imputation methods have been developed, for instance, K-NN imputation [98, 99], mean imputation [100, 101], hot dock imputation [101], regression inputation [96], multiple imputation [102], and SOM imputation [103], etc.
- The missing data is incorporated to the classifier using the machine learning procedures without the previous estimation of missing data, and there exist many methods directly designed for incomplete pattern classification, like neural network ensembles [104,105], decision trees [106], fuzzy approaches [107–109] and support vector machines [110], etc.

The common methodology for pattern classification dealing with missing data consists at first to estimate the missing values using some techniques in order to complete the data, and then to apply a chosen standard pattern classification algorithm. The estimation step plays a crucial role in the procedure, and it can be done by statistical analysis based methods or machine learning methods.

We briefly recall the principles of the three main classical statistical imputation methods:

- Mean imputation [100, 101]: The missing components are replaced by the average value of that component in all the observed cases in the unconditional mean imputation, and they can also be estimated by the average value from the complete cases with the same class label as the incomplete pattern in the class-conditional mean imputation. So the missing data in the same attribute (of the same class) will be all the same. This method is very simple, but it ignores the available attributes information of the pattern, and cannot capture the influence of the known attribute values on the missing data.
- Multiple imputation [102]: The missing values are imputed M times to produce M complete data sets with estimated values using an appropriate model that incorporates random variation, and it can be used for handling of missing data in multivariate analysis. Thus, each missing component is replaced by M plausible values rather than a single one, and this reflects the uncertainty of estimation. Nevertheless, the appropriate model is difficult to obtain in many applications.
- Hot dock imputation [101]: The imputation of the missing component value is based on a similar complete pattern, and the missing value will be filled by corresponding components of the most similar complete vector. However, this imputation mainly depends on a single complete vector in the data set, and the global properties of data set are not taken into account.

Two machine learning based imputation methods: K-NN imputation [98, 99] and SOM imputation method [103] will be briefly introduced here.

- K-NN imputation [98, 99]: The imputation of the missing value is based on the K-nearest neighbors selected from the training patterns with known values in the attributes to be imputed (according to a chosen distance metric). Each of the K neighbors can have different weight in the estimation of the missing values, and the smaller distance to the incomplete pattern leads to bigger weight. The distance between the incomplete pattern and the training samples are calculated using the available attribute data of the incomplete pattern and the corresponding components in the training samples, and the missing values are ignored. If one sets $K = 1$, K-NN imputation will become a particular hot dock imputation. K-NN imputation generally can produce good performance in many applications, but the high computation burden is its main drawback for processing big data set.

- SOM (Self-Organizing Map)[4] imputation [103]: In SOM, the image-node of the incomplete pattern is chosen only measuring the distances with the known attributes, and an activation group composed of image-node's neighbors is selected. Each estimated value of the missing attribute is computed according to the weights of the activation group of nodes in the corresponding dimensions. This approach has been compared with hot-deck and standard multi-layer perceptron (MLP)[5] [112] based imputation, and the result of the comparative analysis indicates that SOM outperforms the other two methods. Particularly, SOM-based method requires less learning observations than other models.

The existing classification methods generally commit the incomplete pattern into one specific class with biggest probability. However, the missing values can play crucial role in classification, and the classification result can be distinct with the different estimations of the missing values. In such case, it is quite likely to cause misclassification error once the object is classified into a particular class, which cannot well reflect the imprecision and uncertainty of classification due to the missing attributes. Belief functions will be introduced to capture such imprecise and uncertain information in the classification of incomplete pattern.

2.6 Conclusion

The credal classification of uncertain data based on belief function theory is studied in this thesis. Belief function theory is the main tool used here, and its basic definitions, several often used combination rules and the decision making support have been briefly introduced. In the classification problem, we mainly consider the supervised and unsupervised conditions. Some supervised classifiers based on belief functions have been recalled. Particularly, the well known EK-NN classifier is commented and reviewed through two examples to show its limitation. Moreover, the classification methods for incomplete pattern with missing values that is often encountered have been also briefly surveyed. For the unsupervised classification, we mainly focus on the c-means clustering methods, and the evidential c-means method is criticized by pointing its weakness. We have proposed four new credal methods for dealing these different cases, and they will be presented in details.

[4] A self-organizing map (SOM) [111] is a type of artificial neural network (ANN), and it is trained by unsupervised learning to produce a low-dimensional representation of the input space of the training samples. SOM uses a neighborhood function to preserve the topological properties of the input space.

[5] The MLP imputation approach commonly consists of training MLP based on only the complete cases as regression model: each incomplete attribute is learned as output by means of the remaining complete attributes given as inputs.

3 Credal classification of uncertain data using close neighbors

3.1 Introduction

In classification problems, the case-based classifier can be a good solution to classify the new input sample (the query object under test) using the collection of labeled (training) samples when the complete statistical knowledge regarding the conditional density functions is not available. In the classification of uncertain and imprecise data, the given attribute information can be insufficient for making a correct specific classification of the objects. For example, the attribute data from different classes can be partly overlapped sometimes. Such objects lying in the overlapping zone are in fact very difficult to classify correctly in a specific class, since the (partly) overlapped classes become undistinguishable. Moreover, some outliers (noisy data) can also be present in some applications.

The well known case-based classification methods, like K-nearest neighbor (K-NN) [113–115], decision trees [62], support vector machine (SVM) [59], artificial neural networks (ANN) [61], have been developed essentially based on probability measure, or fuzzy number for dealing with the uncertain data. The samples are allowed to belong to different specific classes with different memberships, and the class with the biggest membership is usually chosen as final assignment of the object to a class (i.e. the decision-making). However, the probabilistic framework cannot well model and manage the imprecision of data.

The belief function theory [3–7] offers a rigorous mathematical formalism to model uncertain and imprecise information produced by a source of evidence. Some data classifiers have already been developed based on belief functions in the past. For instance, the evidential version of K-nearest neighbors method (EK-NN) has been proposed in [12] based on DST, for working only with the specific classes and the extra ignorant class defined by the union of all the specific classes. The meta-class defined by the union of several specific classes (say $w_i \cup w_j$, $w_i \cup w_j \cup w_k$, etc) is very important and useful to explore the partial imprecision inherent of the data set. However, it has not been considered completely in the existing evidential classifiers developed so far. We propose in this chapter a new case-based classifier working with credal classification, where both the meta-classes and the outlier class are taken into account to fully characterize the uncertainty and imprecision inherent in the data set. In this new method, the sample (the object to assign) is classified using its neighborhood of the training data space, and the K nearest neighbors (KNNs) in each class are used. A total of $c \times K$ (c being the number of classes) neighbors is used to classify the object. This new method is called a *belief* $c \times K$ *neighbors* (BCKN) classifier. In BCKN, $c \times K$ basic belief assignments (BBA's) will be constructed according to the distance between the object and its selected neighbors. A global fusion of these BBA's is done to decide the class, or the meta-class to assign for the object. The credal classification of BCKN can produce specific class, meta-class and outlier class.

An object that is very close to a particular class of data will be committed to this specific class. An object too far from all the training samples will be naturally considered as an outlier (noise),

which is helpful for the outlier detection in some applications. If the object is close to several specific classes (e.g. when lying in the overlapping zone of several different classes), then this object will be committed to the meta-class defined by the union of these specific classes. The meta-class reveals the imprecision in the classification of this object, and can also reduce the misclassifications. Of course the commitments are done in a soft manner thanks to the computation of proper basic belief masses as it will be explained in details in the section 3.2. Such credal classification (a classification based on soft assignments represented by belief functions) is very interesting in many applications, specially those related to defense and security (like in target classification and tracking) because it is generally preferable to get a more robust (and eventually partially imprecise) classification result that could be precisiated later with additional techniques or resources, than to obtain directly with high risk a wrong precise classification from which an erroneous fatal decision would be drawn. This is the main reason why we develop such type of classifiers.

If some samples are committed to the meta-classes, it implies that the used attributes information for classification is insufficient to get the specific classification for these samples. Thus, the output of BCKN can be considered as an interesting source of information to be fused with some other available complementary information sources (when available) for getting more precise classification results in the multi-source information fusion systems. Of course, other sophisticated and generally more costly techniques, like those applied in the military applications, could also be used to classify more precisely the objects in the meta-classes. The use of such additional sophisticated techniques highly depends on the importance of the consequences of the decision to take. The objects in a meta-class are usually a small subset of the total data set. So the price for the specific classification of these objects invoking costly sophisticated techniques can be acceptable for only a limited number of objects, but not for the whole data set at the very beginning of the classification process. Thus, BCKN method provides a way to select the objects (in meta-class) that need a particular attention which should be treated cautiously, as far as important decisions to take are under concern (like in a military targeting process by example).

This chapter is organized as follows. The details of the proposed method BCKN are presented in section 3.2. Several experiments are given in the section 3.3 to show how BCKN performs with respect to other classical methods. Concluding remarks are given in the last section.

3.2 BELIEF $C \times K$ NEIGHBORS CLASSIFIER

3.2.1 Principle of *BCKN*

For the classification of an input sample (the object), we choose K nearest neighbors (KNNs) in training data space of each class. A total of $c \times K$ (c being the number of classes in the whole training data set) neighbors is used in BCKN method. The sources of evidence associated with each class are constructed using these neighbors information, and they have the same weight in the fusion process since they use the same number of neighbors in each class. In a fusion process based on belief functions, the sources of evidence involved in the fusion are assumed to have the same reliability and importance, otherwise some discounting techniques must be applied [3, 116]. The class of the input sample to classify will depend on the global fusion of these sources of evidence. The credal classification of BCKN can produce specific classes, meta-classes and ignorant (outlier) class. A specific class consists of the data points that are very close to the training samples labeled by this class. A meta-class is defined by the union of several specific classes. All objects that are simultaneously close to the specific classes involved in a meta-class will be committed to the meta-class. The ignorant class contains the objects that are too far from all the training samples. The two main steps of BCKN approach are described in details in the next subsections.

3.2.2 The determination of basic belief assignments

Let us consider an object $\mathbf{y}_s \in Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_h\}, s = 1, \ldots, h$ to classify over a c-class frame $\Omega = \{w_1, \ldots, w_c\}$ with a given training data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. w_0 represents the unknown class included in Ω for the exhaustiveness (closure) of the frame. w_0 is used to distinguish the ignorant class denoted by Ω discriminating the objects too far from all the training samples and the meta-class $w_1 \cup \ldots \cup w_c$ describing the objects lying in the overlapping zone of all the singleton classes, as it will be shown in our experiments in section 3.3.

The KNNs of $\mathbf{y}_s$ in each class should be found at first, and there are $c \times K$ neighbors selected in a c-class problem. The BBA's associated with $\mathbf{y}_s$ can be determined using the distances between $\mathbf{y}_s$ and its $c \times K$ neighbors. The L_2-distance (Euclidean distance) between $\mathbf{y}_s$ and one of its neighbors $\mathbf{x}_i$ labeled by w_g is given by:

$$d_{si} = ||\mathbf{y}_s - \mathbf{x}_i|| \tag{3.1}$$

The smaller the distance d_{si} indicates that $\mathbf{y}_s$ more likely belongs to the class of $\mathbf{x}_i$. If $\mathbf{y}_s$ is far from $\mathbf{x}_i$, it means that $\mathbf{x}_i$ provides little useful information regarding the class of $\mathbf{y}_s$. In this work, we adopt a simple and rational way for the determination of BBA's[1]. The BBA's about $\mathbf{y}_s$ are defined for $i = 1, \ldots, c \times K$ and $\mathbf{x}_i \in w_g$ by

$$\begin{cases} m_{si}(w_g) = e^{-\gamma d_{si}} \\ m_{si}(\Omega) = 1 - e^{-\gamma d_{si}} \end{cases} \tag{3.2}$$

where $\gamma > 0$ in eq. (3.2) is a tuning parameter that is used to determine the proper BBA's. If γ takes a very small value, most of the mass of belief is focused on the specific class w_g, even when the object $\mathbf{x}_i$ is quite far from the neighbors in w_g (it means $\mathbf{x}_i$ is not likely in w_g). If γ takes a very big value, the ignorant class Ω will always take the most mass of belief, which is inefficient for the classification problem. γ can be determined according to the average distances between each pair of training samples in the same class. The bigger average distance should lead to a smaller γ value, and so we compute it as

$$\gamma = \frac{1}{\bar{d}} \tag{3.3}$$

with

$$\bar{d} = \frac{1}{cn_i(n_i - 1)} \sum_{i=1}^{c} \sum_{j=1}^{n_i} \sum_{l=1}^{n_i} ||\mathbf{x}_j - \mathbf{x}_l|| \tag{3.4}$$

where c is the number of classes in the data set, and n_i is the number of training samples in class w_i. $\mathbf{x}_j, \mathbf{x}_l$ are the training samples in the class w_i.

According to the BBA model given by eq. (3.2), if d_{si} is very small, most of the mass of belief will be committed to the class w_g of $\mathbf{x}_i$. This indicates that the object $\mathbf{y}_s$ is very likely in the class of $\mathbf{x}_i$. Otherwise, the most mass of belief will be put on the ignorant element Ω to reflect that $\mathbf{x}_i$ has little impact (plays almost a neutral role) in fact on the classification of $\mathbf{y}_s$. So the classification of one object mainly depends on the neighbors that are close to this object. $c \times K$ BBA's corresponding to the $c \times K$ selected neighbors of $\mathbf{y}_s$ in each class $w_g, g = 1, \ldots, c$ can be constructed using this BBA construction model.

[1] There exist other methods of construction of BBA's [117], but they need more tuning parameters and have a higher computation complexity which makes them not easy to use.

3.2.3 The fusion of the basic belief assignments

The fusion results of the $c \times K$ BBA's will be used for the credal classification of the object. The $c \times K$ BBA's can be classified into c groups according to the labels of the neighbors from which the BBA's have been obtained. The BBA's in the same group are all associated with the same class, whereas the BBA's from the different groups corresponding to different classes can highly conflict. So these BBA's are proposed to be fused following the two steps:

- Step1 (sub-combination step): We combine all the BBA's belonging to the same group, and this sub-combination is applied for all the available groups.
- Step 2 (global fusion step): Then, we combine the c BBA's resulting from the previous sub-combination Step 1.

These two steps are explained in more details in the next subsections.

- **Step1: The sub-combination of BBA's in the same group**

DS rule is usually considered as acceptable in most situations where the BBA's are not too conflicting. However, DS rule has several serious limitations as reported [45,46]. It is not appropriate to use DS rule here for combination of the BBA's in the same group because of the particular structure of the BBA's which yields a very fast convergence towards a singleton as stated in the following lemma which is consistent with the Example 2.1 in Chapter 2.

Lemma 3.1: Let us consider a group of K BBA's defined on 2^{Ω} having the following structure $m_{si}(w_g) = \epsilon_i$ and $m_{si}(\Omega) = 1 - \epsilon_i$, where ϵ_i are small positive values, $i = 1, \ldots, K$. Let's denote $\epsilon = \min[\epsilon_1, \ldots, \epsilon_K]$. The combined mass of belief obtained by the DS fusion of the K BBA's $m_{si}(.)$ will be focused on w_g because one always gets $m_{DS}(w_g) > 0.5 > m_{DS}(\Omega)$ as soon as $K > \frac{-\ln 2}{\ln(1-\epsilon)}$.

Proof: In applying the DS rule for combining the K BBA's, we get

$$\begin{cases} m_{DS}(w_g) = 1 - \prod_{i=1}^{K}(1-\epsilon_i) \\ m_{DS}(\Omega) = \prod_{i=1}^{K}(1-\epsilon_i) \end{cases} \tag{3.5}$$

Whence the the bigger K leads to the bigger $m_{DS}(w_g)$. The value of $m_{DS}(w_g)$ will converge very quickly to 1 when K increases. Since $\epsilon = min[\epsilon_1, \ldots, \epsilon_K]$, one always has

$$(m_{DS}(w_g) = 1 - \prod_{i=1}^{K}(1-\epsilon_i)) > 1 - (1-\epsilon)^K$$

which is always greater than 0.5 when $1 - (1-\epsilon)^K > \frac{1}{2}$, or equivalently when $K > \frac{-ln2}{ln(1-\epsilon)}$. This completes the proof. ■

The lemma 3.1 indicates that when the value of K is large enough, the object $\mathbf{y}_s$ will be considered very likely to belong to w_g (according to the combination results of DS rule) even if $\mathbf{y}_s$ is quite far from these K neighbors (i.e. the belief on the specific class w_g is very small). Such DS rule behavior goes against the intuition and is unacceptable. For example, if $\epsilon = \epsilon_i, i = 1, \ldots, K$ is a small value (say $\epsilon = 0.2$) indicating that $\mathbf{y}_s$ is far from the K neighbors, then $m_{DS}(w_g) > 0.5 > m_{DS}(\Omega)$ as soon as $K \geq 4$. Obviously, such combination result is not very reasonable and counter intuitive. If $\mathbf{y}_s$ is quite far from the K neighbors of class w_g ($\mathbf{y}_s$ could however be very close to

the neighbors of w_p class, $p \neq g$), it means that $\mathbf{y}_s$ doesn't very likely belong to w_g, and the most of the belief should (in our opinion) better be committed to the ignorant class Ω after combining efficiently the K BBA's. We have proved in Lemma 1 that DS rule produces unsatisfactory results and we propose to use the simple averaging fusion rule instead of combining the K BBA's in the same group. This rule is defined for $g = 1, \ldots, c$ by

$$\begin{cases} m_s^g(w_g) = \frac{1}{K} \sum_{i=1}^{K} m_{si}(w_g) \\ m_s^g(\Omega) = \frac{1}{K} \sum_{i=1}^{K} m_{si}(\Omega) \end{cases} \tag{3.6}$$

With this averaging fusion rule, the mass of belief on w_g always lies in the following bounds $\min[m_{s1}(w_g), \ldots, m_{sK}(w_g)] \leq m_s^g(w_g) \leq \max[m_{s1}(w_g), \ldots, m_{sK}(w_g)]$. If $\mathbf{y}_s$ is far from a group of K neighbors labeled by w_g (say $m_{si}(w_g) < 0.5, i = 1, \ldots, K$), then the combination results of $m_s^g(w_g)$ will be still very small as $m_s^g(w_g) < 0.5 < m_s^g(\Omega)$, which is logical and consistent with our analysis.

- **Step 2: The global fusion of sub-combination results**

The resulting BBA's of step 1 related to the different groups are combined altogether in Step 2 for the final credal classification of the object $\mathbf{y}_s$. In this global fusion process, we consider not only the specific classes and the ignorant class, but also the possible meta-classes (i.e. partial ignorant classes) for the objects that are difficult to classify correctly into a particular class. The partial conflicting belief (e.g. $m_s(w_i \cap w_j) = m_s^i(w_i) m_s^j(w_j)$ when $w_i \cap w_j = \emptyset$) produced by the conjunction of beliefs of different exhaustive specific classes reflects the ambiguity degree (difficulty) of the classification of the objects in the involved specific classes (e.g. w_i and w_j). Therefore, the mass $m_s^i(w_i) m_s^j(w_j)$ will be committed preferentially to the corresponding meta-class (e.g. $w_i \cup w_j$) rather than being eliminated through a global normalization procedure to avoid counter-intuitive behaviors as those observed with DS rule.

If all the conflicting beliefs are kept and committed to the associated meta-classes (as done classically in DP rule), then too many objects will be assigned to the meta-classes. This is not a very efficient data classification solution because we will lose a lot of specificity in the final result. For example, let us consider a pair of BBA's $m_s^1(w_1) = 0.99, m_s^1(\Omega) = 0.01$ and $m_s^2(w_2) = 0.5 + \epsilon, m_s^2(\Omega) = 0.5 - \epsilon$. If $\epsilon = 0$, both the focal elements w_1 and $w_1 \cup w_2$ will be considered most likely to be true in the fusion results of DP rule. If $\epsilon > 0$, the meta-class $w_1 \cup w_2$ will get the most belief after the combination of the two BBA's by DP rule because of the existing partial conflict belief $m_s(w_1 \cup w_2)$. Nevertheless, this object is more likely in w_1 than in w_2, since the belief on w_1 in $m_s^1(\cdot)$ is much bigger than the belief on w_2 in $m_s^2(\cdot)$. The classes w_1 and w_2 seem not so undistinguishable for $\mathbf{y}_s$ in such condition in fact. Thus, it is not very reasonable to commit this object into the meta-class $w_1 \cup w_2$. That is why in BCKN, we propose to select the meta-classes that should be kept conditionally according to the current context.

In the c pieces of sub-combination results, the biggest mass of belief on specific class is first identified, that is $m_s^{\max}(w_{\max}) = \max[m_s^1(w_1), \ldots, m_s^c(w_c)]$. The class $w_{\max}$ corresponds to the most likely class of $\mathbf{y}_s$. If $(m_s^{\max}(w_{\max}) - m_s^i(w_i)) \leq t, i = 1, \ldots, c$ (t being a given threshold), then the class w_i will be also considered as potentially likely true. In fact the classes w_i and $w_{\max}$ are almost undistinguishable for the classification of $\mathbf{y}_s$ with respect to the given threshold t, and it is with high risk of error for the assignment of this object to one specific class. Therefore, the object $\mathbf{y}_s$ should be cautiously committed to a set of classes $\psi_{\max} \triangleq \{w_i | m_s^{\max}(w_{\max}) - m_s^i(w_i) \leq t\}$ with big mass of belief. It says that $\mathbf{y}_s$ likely belongs to one of specific classes in $\psi_{\max}$, but these specific classes cannot be well distinguished for $\mathbf{y}_s$. In order to deal with all the classes in an equal manner, all the meta-classes having a cardinality less or equal to $|\psi_{max}|$ will be selected, and their

corresponding conflicting beliefs will be preserved and committed to the mass of the corresponding meta-classes. The set of the selected meta-classes is denoted by Ψ.

Example 3.1: Let's consider $\Omega = \{w_1, w_2, w_3\}$. If $w_{\max} = w_1$ and $\psi_{\max} = \{w_1, w_2\}$, then all the meta-classes whose cardinality is not bigger than $|\{w_1, w_2\}| = 2$ will be kept. Therefore, the selected meta-classes are the elements of the set[2] $\Psi = \{w_1 \cup w_2, w_1 \cup w_3, w_2 \cup w_3\}$. If the belief on $w_{\max}$ is much bigger than that on any other classes, none meta-class needs to be preserved in order to avoid the high imprecision of the solution. The guidelines for tuning the threshold parameter t are discussed in the sequel.

In this work, the global fusion rule of Step 2 of our BCKN method has been inspired by DS rule (2.4) and DP rule (2.7). It is mathematically defined by the formulas (3.7) and (3.8). The sub-combination results associated with the $\mathbf{y}_s$ and different classes can be fused sequentially by

$$m_s^{1,g}(A) = \begin{cases} \sum\limits_{B_1, B_2 \in 2^\Omega | B_1 \cap B_2 = A} m_s^{1,g-1}(B_1) m_s^g(B_2), & \text{for } A \notin \Psi \\ \sum\limits_{B_1, B_2 \in 2^\Omega | B_1 \cup B_2 = A} m_s^{1,g-1}(B_1) m_s^g(B_2), & \text{for } A \in \Psi \end{cases} \tag{3.7}$$

$m_s^{1,g}(.)$ is the unnormalized combination results of $m_s^1(.), \ldots, m_s^g(.), g = 1, \ldots, c$. By convention, one takes $m_s^{1,1}(.) = m_s^1(.)$. It is worth to note that this combination rule is close to DP rule (2.7) in its principle, but the summation of the combined BBA is not one (i.e. here one can have $\sum m_s^{1,g}(.) \leq 1$) if some partial conflicting beliefs are not preserved. In DP rule, the focal element A can be any subset of Ω, whereas in our (restricted version of DP) rule a focal element A can be only a specific class, the ignorant class or just a selected meta-class in Ψ (but not all possible meta-classes). This unnormalized combination rule is not associative in general, but it is associative here because of the very particular structure of BBA's $m_s^i(.)$ satisfying the BCKN model. This will be shown in the example 3.2.

The unnormalized fusion results obtained by (3.7) will then be normalized once all the BBA's have been combined. The mass of conflicting beliefs that have not been committed to the meta-classes must be redistributed to the other focal elements to get a normalized final BBA. In this work, the masses of conflicting beliefs are added together to compute the level of the total conflict which is then redistributed to all the focal elements (including specific class and meta-class) by the classical normalization procedure (as done in the DS rule). More precisely, the normalization of the fusion results is done by

$$m_s(A) = \frac{m_s^{1,c}(A)}{\sum\limits_j m_s^{1,c}(B_j)} \tag{3.8}$$

where $m_s^{1,c}(.)$ is the unnormalized BBA obtained after combining sequentially all the BBA's $m_s^g(.)$ for $g = 1, 2, \ldots, c$ with the formula (3.7).

If it is known that there is no outlier in the application under concern, then the mass on the ignorant class can be proportionally redistributed to other focal elements using eq. (3.8). If none meta-class is selected, the global fusion rule reduces to DS rule, since none of the conflicting beliefs can be transferred to meta-classes. Whereas, if all the meta-classes are preserved, the global fusion rule behaves like DP rule, and all the partial conflicting beliefs masses are transferred onto these meta-classes. This global fusion rule can be considered as a compromise between DS rule and DP rule, since we only select a subset of all possible meta-classes on which the conflicting beliefs masses will be redistributed. The selection of the admissible meta-classes used in BCKN method depends on the current context. The global fusion results can be used for the classification making support. The belief function $Bel(.)$, the plausibility function $Pl(.)$ and pignistic probability $BetP(.)$ can be

[2] In BCKN, the meta-classes involving w_0 like $w_0 \cup w_i$ are not taken into account.

used for final hard (binary) assignment of the objects to a specific class when it is really necessary. Such final hard assignment is not the purpose of BCKN since we do prefer to use the credal classification as a mean to understand the inherent structure of the data to classify, and this will help to request specific extra resources to better precisiate the result for some important objects.

In Fig. 3.1, the flowchart of BCKN is presented to explicitly illustrate how the proposed method works.

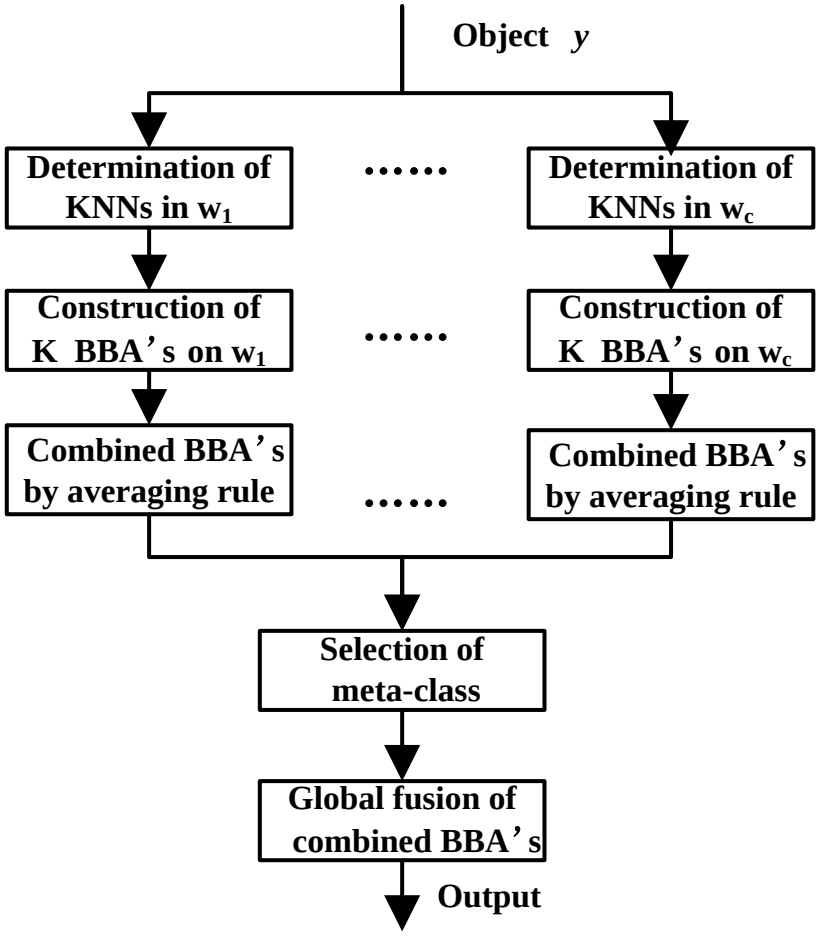


Figure 3.1 : Flowchart of the proposed BCKN method.

The pseudo-code of the BCKN is also given in Table 3.1 for convenience.

Table 3.1 : Belief $c \times k$ neighbors algorithm

Input:	Training samples: $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in $\mathbb{R}^p$ Objects to classify: $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_h\}$ in $\mathbb{R}^p$
Parameters:	K: number of nearest neighbors $t > 0$: threshold for meta-class
	for s=1 to h Select the K nearest neighbors of $\mathbf{y}_s$ in each class Construction of $c \times K$ BBA's using (3.2); Combination of BBA's from neighbors with same label by (3.6); Selection of meta-classes according to sub-combination results; Global fusion of these sub-combination results by (3.7) and (3.8); Credal classification of $\mathbf{y}_s$ based on the global fusion results. **end**

3.2.4 Guidelines for choosing the threshold parameter t

The BCKN method requires to choose the threshold parameter t for the contextual selection of meta-classes. The tuning of this parameter is very important in the application of BCKN. We provide here simple guidelines for the choice of this threshold t. The bigger threshold t can produce the fewer misclassifications, but it usually brings the larger meta-classes which is not efficient for maintaining an acceptable specificity of the classification result. Thus, the tuning of t depends on the expected compromise we want between the imprecision and the misclassification of the results. t can be optimized using the cross-validation (e.g. leave-one-out) in training data space with the given K value. t can be also tuned by a grid-search in $[0, 1]$. The optimal choice of t should correspond to the compromise we want between the imprecision and misclassification, which is application dependent. The following example shows how BCKN works.

Example 3.2: Let us consider the frame of classes $\Omega = \{w_0, w_1, w_2, w_3\}$ and the given value of $K = 2$. In the training data space of each class[3] $w_i, i = 1, 2, 3$, $K = 2$ nearest neighbors are searched at first. Then, the $K \times c = 2 \times 3 = 6$ BBA's of the object $\mathbf{y}_s$ to classify are determined using the distances between $\mathbf{y}_s$ and the $K \times c = 6$ neighbors. Let's suppose in this example that these BBA's are given by:

$$
\begin{aligned}
m_{s1}(w_1) &= 0.9, \quad m_{s1}(\Omega) = 0.1 \\
m_{s2}(w_1) &= 0.8, \quad m_{s2}(\Omega) = 0.2 \\
m_{s3}(w_2) &= 0.9, \quad m_{s3}(\Omega) = 0.1 \\
m_{s4}(w_2) &= 0.7, \quad m_{s4}(\Omega) = 0.3 \\
m_{s5}(w_3) &= 0.4, \quad m_{s5}(\Omega) = 0.6 \\
m_{s6}(w_3) &= 0.2, \quad m_{s6}(\Omega) = 0.8
\end{aligned}
$$

Thus $m_{s1}(.)$ and $m_{s2}(.)$ strongly support w_1, $m_{s3}(.)$ and $m_{s4}(.)$ support w_2, and $m_{s5}(.)$ and $m_{s6}(.)$ support moderately the class w_3. The combination results of the BBA's in the same group using

[3] In this work, the training samples are all considered with specific labels, and w_0 is the potential unknown class for some objects to test. So none training samples belongs to w_0.

the averaging rule eq. (3.6) gives

$$m_s^1(w_1) = \frac{m_{s1}(w_1) + m_{s2}(w_1)}{2} = 0.85, \quad m_s^1(\Omega) = \frac{m_{s1}(\Omega) + m_{s2}(\Omega)}{2} = 0.15$$
$$m_s^2(w_2) = \frac{m_{s3}(w_2) + m_{s4}(w_2)}{2} = 0.8, \quad m_s^2(\Omega) = \frac{m_{s3}(\Omega) + m_{s4}(\Omega)}{2} = 0.2$$
$$m_s^3(w_3) = \frac{m_{s5}(w_3) + m_{s6}(w_3)}{2} = 0.3, \quad m_s^3(\Omega) = \frac{m_{s5}(\Omega) + m_{s6}(\Omega)}{2} = 0.7$$

We see that $m_s^{\max} = m_s^1(w_1) = 0.85$. If we choose the value of $t = 0.1$, one gets $\psi_{\max} = \{w_1, w_2\} \triangleq w_1 \cup w_2$ since $m_s^{\max} - m_s^2(w_2) < 0.1$. Then the meta-classes having cardinality no bigger than $|\Psi_{\max}| = |\{w_1, w_2\}| = 2$ should be selected from the power-set 2^Ω. Therefore, the selected meta-classes are $\Psi = \{w_1 \cup w_2, w_1 \cup w_3, w_2 \cup w_3\}$. Because of the particular[4] structure of the BBA's, the unnormalized combination rule (3.7) is associative as we can verify in this simple example. Indeed, if A is a specific class or the ignorant class Ω, then one always has from eq. (3.7)

$$m^{1,3}(A) = m_1(A)m_2(\Omega)m_3(\Omega) = m^{1,2}(A)m_3(\Omega) = m_1(A)m^{2,3}(\Omega) = m^{1,3}(A)$$

If A is a selected meta-class, say $A = w_1 \cup w_2$, then one gets from eq. (3.7)

$$\begin{aligned} m^{1,3}(w_1 \cup w_2) &= m_1(w_1)m_2(w_2)m_3(\Omega) = m^{1,2}(w_1 \cup w_2)m_3(\Omega) \\ &= m_1(w_1)m^{2,3}(w_2) = m^{1,3}(w_1 \cup w_2) \end{aligned}$$

Such result is similar when considering $A = w_1 \cup w_3$ and $A = w_2 \cup w_3$.

Finally the result obtained by the global fusion rule (3.7) of Step 2 is

$$\begin{aligned} m_s^{1,3}(w_1) &= m_s^1(w_1)m_s^2(\Omega)m_s^3(\Omega) = 0.1190 \\ m_s^{1,3}(w_2) &= m_s^1(\Omega)m_s^2(w_2)m_s^3(\Omega) = 0.0840 \\ m_s^{1,3}(w_3) &= m_s^1(\Omega)m_s^2(\Omega)m_s^3(w_3) = 0.0090 \\ m_s^{1,3}(\Omega) &= m_s^1(\Omega)m_s^2(\Omega)m_s^3(\Omega) = 0.0210 \\ m_s^{1,3}(w_1 \cup w_2) &= m_s^1(w_1)m_s^2(w_2)m_s^3(\Omega) = 0.4760 \\ m_s^{1,3}(w_1 \cup w_3) &= m_s^1(w_1)m_s^2(\Omega)m_s^3(w_3) = 0.0510 \\ m_s^{1,3}(w_2 \cup w_3) &= m_s^1(\Omega)m_s^2(w_2)m_s^3(w_3) = 0.0360 \end{aligned}$$

These masses are then normalized according to (3.8), and we get

$$\begin{aligned} &m_s(w_1) = 0.1495 \qquad m_s(w_2) = 0.1055 \qquad m_s(w_3) = 0.0113 \\ &m_s(w_1 \cup w_2) = \mathbf{0.5980} \qquad m_s(w_1 \cup w_3) = 0.0641 \\ &m_s(w_2 \cup w_3) = 0.0452 \qquad m_s(\Omega) = 0.0264 \end{aligned}$$

The result indicates that the sample $\mathbf{y}_s$ most likely belongs to the meta-class $w_1 \cup w_2$, since $w_1 \cup w_2$ gets the most mass of belief. We can see that the belief granted to w_1 are similar to the belief granted to w_2 with respect to the given threshold t. It indicates that $\mathbf{y}_s$ is close to both the classes w_1 and w_2, and w_1 and w_2 are not very distinguishable for making a precise classification of $\mathbf{y}_s$. What we can only reasonably infer is that $\mathbf{y}_s$ belongs to $\{w_1, w_2\}$. So the meta-class $w_1 \cup w_2$ can be a good (acceptable) compromise for the classification of $\mathbf{y}_s$, which reduces the risk of misclassification. This solution is also consistent with our intuition. Such fusion result can be considered as a useful mean to ask for other complementary information sources if a more precise classification is absolutely necessary for the problem under consideration.

[4] The focal elements of each BBA are nested.

3.2.5 Expressive power of BCKN

The expressive power of a method can be seen as its ability to identify and manipulate complex propositions. The expressive power of the classification methods can be represented by the focal elements they can generate in the classification results. Let us examine and compare the expressive power of credal classification in BCKN with respect to probabilistic classification and classification in classical evidential methods. Let us consider a finite frame of discernment $\Omega = \{w_0, w_1, \ldots, w_c\}$ with $c > 1$ specific classes. Probabilistic classification provides only a Bayesian BBA (a probability measure) which can focus only on the c possible focal elements (singletons) of Ω. So the expressive power of probabilistic classification is c. In the classical evidential methods [12], it can provide a positive mass only on the c singletons of Ω and also the ignorance class Ω. So its expressive power is $c+1$. In BCKN, the credal classification can provide a positive mass on the c singletons of Ω, on the ignorance class Ω, and on $2^c - c - 1$ meta-classes as well. So the expressive power of credal classification in BCKN is 2^c. It is worth to note that the meta-class in BCKN is conditionally selected according to the given threshold t under the current context. The meta-class is not included when all the objects can be clearly classified, but meta-class is necessary when some objects are difficult to classify correctly. One sees that the credal classification produces more enlarged classifications and has a better expressive power than classical methods. The expressive power of BCKN goes from $c+1$ (no meta class) to 2^c. The more expressive it is, the more computationally costly it is. The following example shows the expressive power of different classifications.

Example 3.3: Let us consider a 3-classes data set. The three specific classes are w_1, w_2, and w_3. Below are the feasible classes expressed by the different methods.

- probabilistic classification: w_1, w_2, and w_3;
- classical evidential classification: w_1, w_2, w_3, and Ω;
- credal classification: w_1, w_2, w_3, $w_1 \cup w_2$, $w_1 \cup w_3$, $w_2 \cup w_3$, $w_1 \cup w_2 \cup w_3$ and Ω.

However, the computation burden of BCKN is generally bigger than the classical neighbor-based methods. In K-NN and EK-NN, there are K neighbors involved in the classification of one object, whereas it requires $c \times K$ (c being the number of the classes) neighbors in BCKN. The K BBA's corresponding to the K neighbors are simply combined using the DS rule to classify the object in EK-NN. Nevertheless, the combination of the $c \times K$ BBA's in BCKN should follow the two steps: 1) the sub-combination of the BBA's associated with the same class and 2) the global fusion of these sub-combination results in BCKN. So the computational complexity of BCKN seems bigger than EK-NN and K-NN. This is the necessary price we have to pay for the enlarged credal classification of the uncertain and imprecise data.

3.3 Experiments

BCKN has been tested in several experiments to evaluate its performance with respect to K-NN, EK-NN, Classification And Regression Tree (CART), Artificial Neural Networks (ANN) and Support Vector Machine (SVM) methods. In the following three experiments (i.e. Experiment 1, 2 and 3.), the tuning of parameters in these different methods are introduced as follows. The different methods have been programmed and tested with Matlab^{TM} software. The parameters of EK-NN were automatically optimized using the method introduced in [118]. In ANN, we use the feed-forward back propagation network with $epochs = 500$ and $goal = 0.001$. In SVM, we selected a Gaussian Radial Basis Function kernel with $\sigma = 0.125$. The tuning threshold t in BCKN is optimized using the training data. The optimized value corresponds to a suitable compromise

between error rate and imprecision rate (for example, imprecision rate is no more than five percent and it is no bigger than the error rate). The t value has been found by a grid search with 10^{-4} step width in the range $[0, 1]$. This optimization procedure can be done off-line.

In order to explicitly show the use of meta-class introduced in BCKN, the objects are directly committed to the class that receives the maximal mass of belief. In this work, we use both the common error rate, and a new concept of imprecision rate (related with the meta-classes) to evaluate the performance of BCKN. For one object originated from w_i, if it is classified into A with $w_i \cap A = \emptyset$, it will be considered as an error. If $w_i \cap A \neq \emptyset$ and $A \neq w_i$, it will be considered with imprecise classification. The error rate denoted by Re is calculated by $Re = N_e/T$, where N_e is number of objects wrongly classified, and T is the total number of the objects tested. The imprecision rate denoted by Ri_j is calculated by $Ri_j = N_{I_j}/T$, where N_{I_j} is number of objects committed to the meta-classes with the cardinality value j.

It is worth noting that the x-axis corresponds to the first dimension of test and training data, and y-axis corresponds to the second dimension in Fig. 3.2–3.3. In Fig. 3.4 and 3.5, the x-axis represents the number of K value used in BCKN method, and y-axis represents the error rate(or imprecision rate for BCKN).

3.3.1 Experiment 3.1 (with artificial data sets)

This experiment consists of two particular tests (numerical simulations) and shows how BCKN works and its difference with respect to EK-NN and K-NN methods. The tuning of parameters of these methods have been presented in the beginning of Section 3.3.

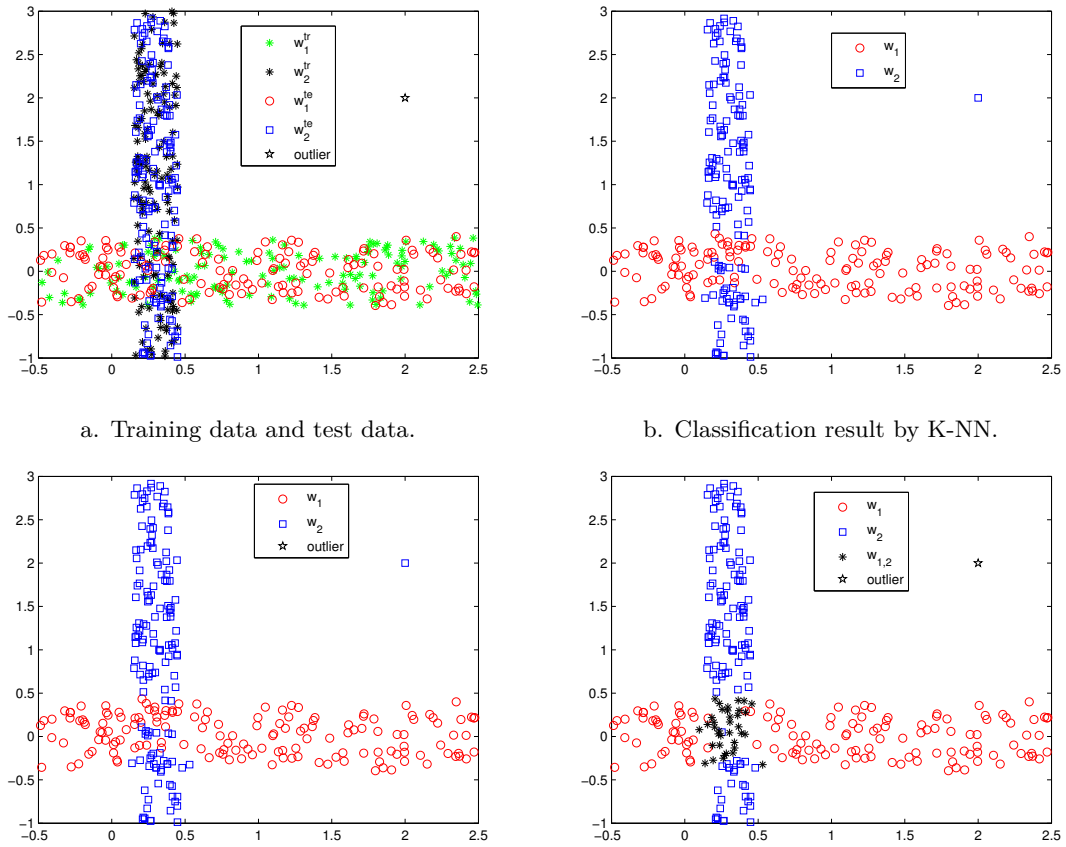
3.3.1.1 Test 1

BCKN is tested here using two particular 2-D data classes w_1 and w_2 that are obtained from two uniform distributions as shown by Fig. 3.2-a. Each class has 200 training samples and 200 test samples, and one more noisy sample (outlier) is included in the test samples. The uniform distributions of the two classes are characterized by :

	x-label interval	y-label interval
w_1	(-0.5, 2.5)	(-0.4, 0.4)
w_2	(0.15, 0.45)	(-1, 3)

A particular value of $K = 11$ is selected here, since it produce good results for all the three methods. So $K = 11$ neighbors are selected by K-NN and EK-NN, whereas there are $c \times K = 2 \times 11 = 22$ neighbors used in BCKN. The classification results of the tested objects by the different methods are given by Fig. 3.2-b–3.2-d. For notation conciseness, we have denoted $w^{te} \triangleq w^{test}$, $w^{tr} \triangleq w^{training}$ and $w_{i,\ldots,k} \triangleq w_i \cup \ldots \cup w_k$.

As we can see on Fig. 3.2-a, some tested objects originating from w_1 and w_2 can belong to the crossed (overlapped) zone and such objects are really hard to classify into a particular class w_1 or w_2. However, K-NN and EK-NN just commit these objects in the overlapping zone to two specific classes as shown on Fig. 3.2-b and on Fig. 3.2-c because of the limitation of probabilistic framework, and this can cause many misclassifications (i.e. EK-NN produces 26 errors, and K-NN also produces 26 errors). In BCKN, these objects are automatically reasonably classified into the



a. Training data and test data.

b. Classification result by K-NN.

c. Classification result by EK-NN.

d. Classification result by BCKN ($t = 0.005$).

Figure 3.2 : Classification results by K-NN, EK-NN and BCKN.

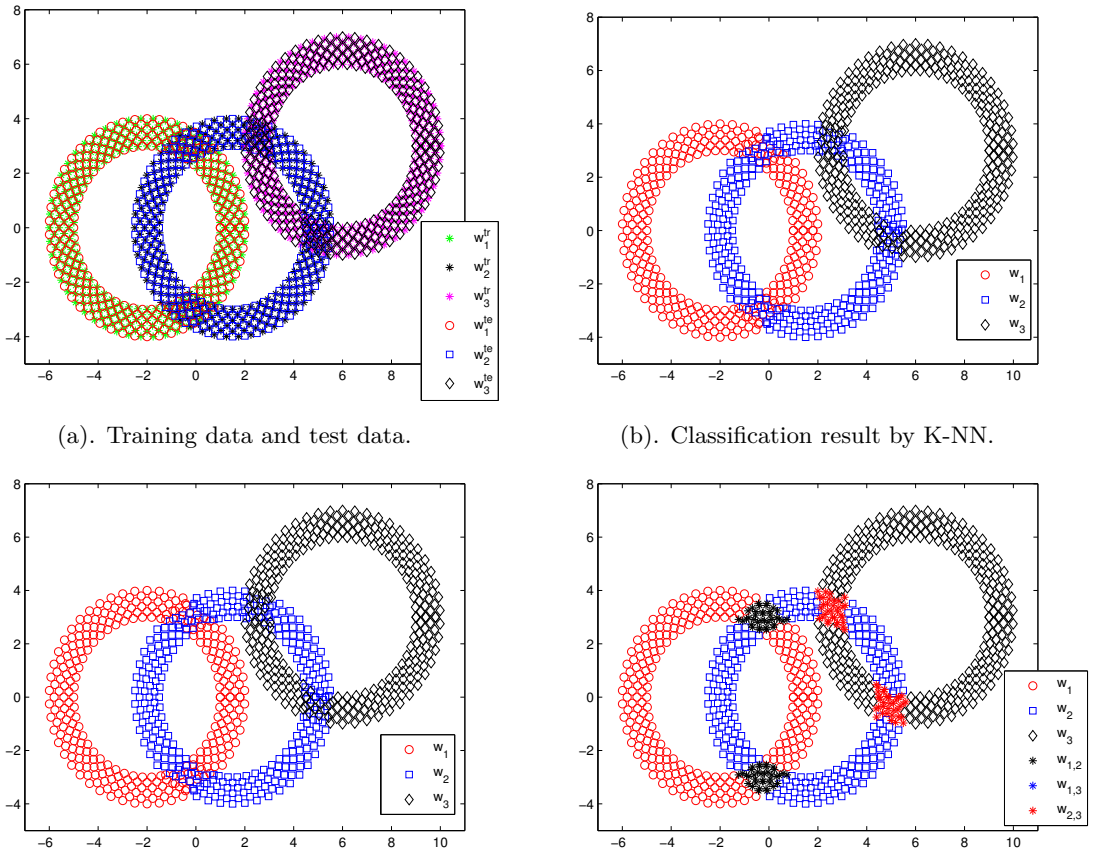
meta-class $w_1 \cup w_2$ thanks to the belief functions framework as shown on Fig. 3.2-d. BCKN is thus able to effectively reduce misclassification (i.e. BCKN produces 3 errors, and 36 points in the meta-class). One object labeled by black pentagram is far from the others. Therefore, it is considered as outlier by BCKN. Whereas, this noisy point is not detected by other methods. This example shows the interest of the credal classification provided by the BCKN approach.

3.3.1.2 Test 2

A 3-class 2-D data set composed by three rings shown by Fig. 3.3-(a) is used in this example. Each class contains 303 training samples and 303 objects for testing. The radiuses and centers of the three rings are given by:

	center	radius interval
w_1	(-2, 0)	[3, 4]
w_2	(1.5, 0)	[3, 4]
w_3	(6, 3)	[3, 4]

We have also taken $K = 11$ in this second test. The classification results of test data by K-NN, EK-NN and BCKN are respectively shown in Fig. 3.3-(b)-(d).



(a). Training data and test data.

(b). Classification result by K-NN.

(c). Classification result by EK-NN.

(d). Classification result by BCKN ($t = 0.002$).

Figure 3.3 : Classification results of a 3-class data set by K-NN, EK-NN and BCKN.

We can see that the three rings intersect, and the objects in the overlapping (intersecting) zones are impossible to classify correctly. In the classification results of K-NN and EK-NN, all the objects are committed to a particular class as shown on Fig. 3.3-(b), (c). K-NN and EK-NN generate both 109 misclassifications. In BCKN, the objects in the overlapping zones are reasonably automatically associated to meta-classes as shown on Fig. 3.3-(d). The BCKN produces only 4 misclassifications, but it commits 141 objects in the meta-classes. This example shows the effectiveness of BCKN for dealing with ambiguous data in a complex situation.

3.3.2 Experiment 3.2 (with 4-class data set)

In this second experiment, we compare the performances of BCKN with respect to the performances of EK-NN, K-NN, CART, ANN and SVM on a 4-classes problem. The data set is generated from

three 2D Gaussian distributions characterizing the classes w_1, w_2, w_3 and w_4 with the following means vectors and covariance matrices:

$$\mu_1 = (-5, 0), \Sigma_1 = [1, 0; 0, 6]$$

$$\mu_2 = (5, 0), \Sigma_2 = [1, 0; 0, 6]$$

$$\mu_3 = (0, 5), \Sigma_3 = [6, 0; 0, 1]$$

$$\mu_4 = (0, -5), \Sigma_4 = [6, 0; 0, 1]$$

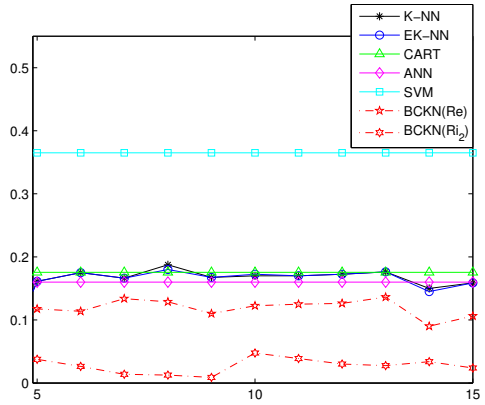
There are 4×100 test samples, and the training sets contain $4 \times N$ samples (for $N = 100, 200, 300$). Values of K ranging from 5 to 15 neighbors have been tested. For each pair (N, K), the reported error rates and imprecision rates are averages of 10 trials performed with 10 independent random generation of the data sets. The mean of the classification error and imprecision rates with different numbers of training samples (for $N = 100, 200, 300$) have been calculated, and the classification results by different methods are shown on Fig. 3.4. The average error rate Re_a, imprecision rate Ri_a and execution time (second) of BCKN, K-NN, EK-NN with $K = 5, 6, \ldots, 15$, as well as the error rate and computing time of CART, ANN and SVM are given in Table 3.2. It is worth noting that there are $c \times K = 4K$ neighbors involved in the classification by BCKN. The outliers have not been introduced in this experiment, and the belief on ignorant class has been proportionally redistributed to other available classes. In this experiment, none object is committed to the meta-class with cardinality value of three or four, and that is why we have only considered Ri_{a2}. The tuning of parameters in these different methods have been introduced in the beginning of Section 3.3.

Table 3.2 : The statistics of classification results by different methods (in %).

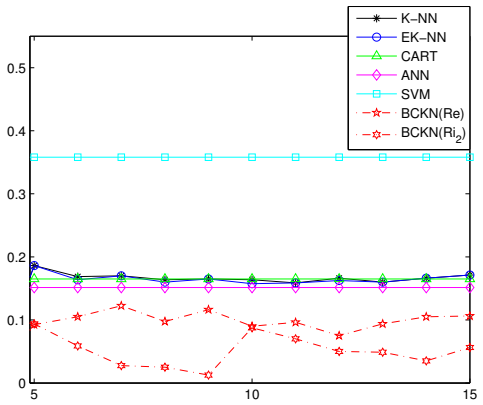
		$N = 100$	$N = 200$	$N = 300$
K-NN	Re_a	16.86	16.73	15.75
	time	0.0220	0.0362	0.0496
EK-NN	Re_a	16.77	16.56	15.90
	time	0.0695	0.0872	0.1099
CART	Re_a	17.55	16.50	16.35
	time	0.2387	0.3182	0.4306
ANN	Re_a	16.00	15.15	15.10
	time	3.5475	7.7657	7.7751
SVM	Re_a	36.50	35.80	35.35
	time	2.1466	11.7625	67.8916
BCKN	Re_a	11.91	10.00	8.98
	Ri_{a2}	2.73	5.14	6.77
	time	0.9169	1.8826	2.9016

In Fig. 3.4 and Fig. 3.5, the X-axis corresponds to the K values, and the Y-axis corresponds to the classification error rate R_e expressed in $[0, 1]$ (and also the imprecision rate Ri_2 for BCKN) of the classification methods.

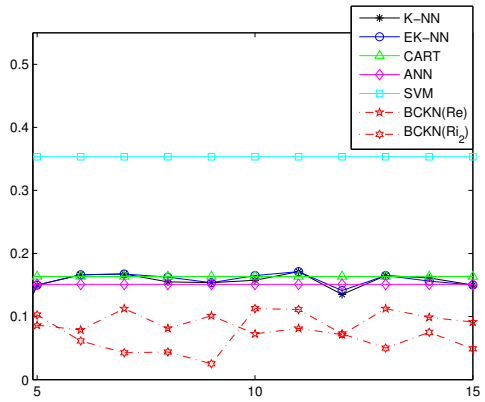
We can observe on Table 3.2 and Fig. 3.4 that BCKN produces the smallest error rate. In fact, the objects in class w_1 and w_2 are partly overlapped with the samples in w_3 and w_4. The objects



(a). Classification results with $N = 100$



(b). Classification results with $N = 200$



(c). Classification results with $N = 300$

Figure 3.4 : Classification results of the 4-class problem by different methods.

in the overlapping zones are very difficult to classify, and most of them are wrongly classified by the classical methods. Whereas, these objects that are difficult to correctly classify are mostly committed to the associated meta-classes (i.e. $w_1 \cup w_3$, $w_1 \cup w_4$, $w_2 \cup w_3$ and $w_2 \cup w_4$) by BCKN. That is why BCKN produces the fewest errors but brings naturally some imprecision of classification (i.e. meta-class).

We can see that BCKN takes more execution time than K-NN and EK-NN on Table 3.2, and it indicates that the computational complexity (burden) of BCKN is bigger than the other classical neighbor-based methods. This is the price we pay for the enlarged credal classification, which can provide more useful information in the classification than other classical methods.

3.3.3 Experiment 3.3 (with real data sets)

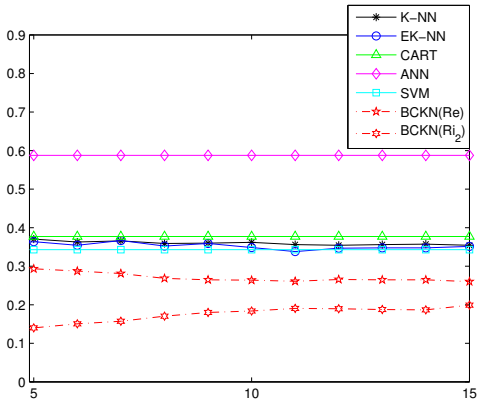
Four well-known data sets available from UCI [95] (the Wine data set, the Iris data set, the Breast cancer and Yeast data sets) are widely used by the scientific community to test data classification methods. So we have also used these real data sets to evaluate the performance of BCKN with respect to other classical methods. The basic information about these data sets are given in Table 3.3. Three classes (CYT, NUC and $ME3$) are selected in Yeast data set to the evaluate our

method, since these three classes are close and hard to classify.

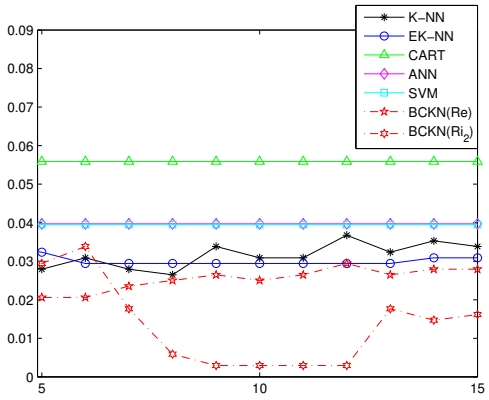
The k-fold cross validation is performed on the three data sets by different classification methods, and k generally remains a free parameter [119]. We use the common 10-fold cross validation here. The tuning parameter t is optimized using the training samples in each fold. The classification results by different methods with different values of K ranking from 5 to 15 are respectively shown on Fig. 3.5-(a)–(d). The average error rate Re_a, imprecision rate Ri_a (for BCKN) and execution time (second) of the different methods including K-NN, EK-NN, CART, ANN, SVM and BCKN are given in Table 3.4. The outlier class is absent in this real data sets, and the belief on the ignorant class is proportionally distributed to the other focal elements. The parameters in these different methods are tuned following the way introduced in the beginning of Section 3.3.

Table 3.3 : Basic information of the real data sets used for the test.

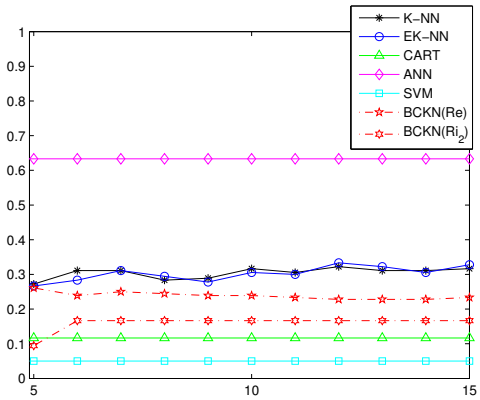
name	classes	attributes	instances
Wine	3	13	178
Breast cancer	2	9	683
Iris	3	4	150
Yeast	3	8	1055



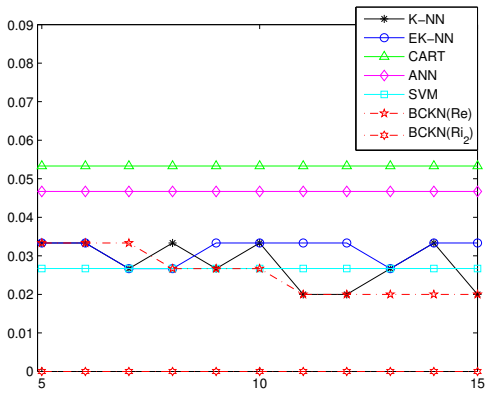
(a). Classification results of Yeast data.



(b). Classification results of Breast cancer data.



(c). Classification results of Wine data.



(d). Classification results of Iris data.

Figure 3.5 : Classification results of real data sets by different methods.

Table 3.4 : The statistics of the classification results for different real data sets (in %).

		Yeast	Breast cancer	Wine	Iris
K-NN	Re_a	35.97	3.16	30.45	2.79
	time	0.0143	0.0098	0.0030	0.0024
EK-NN	Re_a	35.24	2.99	30.25	3.15
	time	0.2261	0.1326	0.0228	0.0186
CART	Re	37.71	5.59	11.67	5.33
	time	0.8034	0.1934	0.1045	0.0811
ANN	Re	58.76	3.97	63.33	4.67
	time	6.7049	6.4584	3.3072	2.9905
SVM	Re	34.29	3.95	5.00	2.67
	time	2.1528	1.8861	0.2792	0.2308
BCKN	Re_a	27.05	2.54	23.84	2.55
	Ri_{a2}	17.59	1.34	16.01	0
	time	0.7484	0.2714	0.0211	0.0156

In these tests based on real data sets, none object is committed to the meta-class with cardinality value of three, and we have just taken Ri_{a2} for the evaluation. In the classification of Breast cancer data set and Yeast data set, the error rate of BCKN is smaller than the error rates obtained with other classical methods since the samples difficult to classify are automatically committed to the meta-classes by BCKN. For the Iris data set, although none object is committed to meta-class by BCKN, BCKN still provides a smaller error rate than with other methods. In the classification of Wine data set, BCKN produces a lower error rate than with K-NN, EK-NN and ANN methods, but SVM and CART obtain better results than the other neighbor-based methods[5]. BCKN requires a bit more time-consuming than K-NN and EK-NN on Table 3.4, which indicates that the computation burden of BCKN is bigger than K-NN and EK-NN. In Wine, Yeast and Breast cancer data sets, there are some objects belonging to meta-classes. The BCKN results clearly indicate that the attributes used in these three real data sets are in fact insufficient for making the correct specific classification for the objects in the meta-classes. We should treat these objects more cautiously and other complementary information sources will be necessary to get better specific results (if necessary). If we are forced to make a hard classification of these objects, we must be ready to take a high risk of misclassification, although a small part of the objects in meta-classes may be correctly classified in the hard classification. Our tests and analysis illustrate the interest and the potential of this new BCKN approach in real classification problems.

3.4 Conclusion

A new belief $c \times K$ neighbors (BCKN) classifier has been developed in this chapter to deal with the uncertain and imprecise data, and it works with credal classification based on the belief functions. The main advantage of this approach is the classification of the objects done according to the context. With BCKN, the object can be either in the specific classes, or in the meta-classes (i.e. the union of several specific classes), or eventually in the ignorant class. The BCKN credal

[5]CART and SVM being based on very different principles, it is difficult to draw a firm conclusion to establish if they outperforms or not BCKN in general. This question remains an interesting topic for future research.

classification allows to reduce the error rate by introducing the meta-class, which characterizes the partial imprecision of classification, and it allows also to well detect the outliers thanks to the ignorant class. The output of the BCKN classifier can be used as a primary source of information to orient the need of other complementary means of analysis when more precise results on the ambiguous objects are necessary. The comparative analysis of BCKN method with respect to other classical methods through several experiments (using both synthetic data sets and real data sets) has shown its real ability to reduce the classification errors by increasing judiciously the imprecision rate that one accepts in the applications. In practice, a suitable compromise between the error rate and imprecision rate must always be found by optimizing the choice of threshold parameter entering in the BCKN approach. BCKN is able to well deal with the general and complicate cases, but its computation burden is higher than with K-NN and EK-NN due to the enlarged credal classification. The reduction of the computational burden of the BCKN will be investigated in next chapter when the situation of classification is simple.

4 Credal classification rule for uncertain data using prototype of each class

4.1 Introduction

In the previous chapter, a belief $c \times K$ neighbors (BCKN) classifier working with credal classification has been developed to deal with uncertain data by considering all possible meta-classes in the process. In BCKN, the distances between the object and all the training samples should be calculated, and such method requests a high computational burden which is usually the main drawback of all K-NN alike methods [113–115]. In this chapter, we propose a new straightforward and more simple mathematical solution, called Credal Classification Rule (CCR), for directly computing the basic belief assignments of uncertain data for their credal classification.

The interest of credal classification mainly resides in its ability to commit objects to the meta-classes rather than to the specific classes when the information is insufficient for making it correctly. By doing so, we preserve the robustness of the result and we reduce the risk of misclassification errors. Of course the price to pay is the increase of the non-specificity of the classification. In some applications, like in target classification and tracking, it is very crucial to maintain such robustness than to provide immediately (with high risk of error) a precise classification. The credal classification can be very helpful to manage external (possibly costly) complementary resources in order to reduce some particular ambiguities. Our approach is very helpful for requesting (or not) a complementary information sources (if possible and available) in order to get more precise reliable classification results, and to reduce dramatic errors in the final decision-making process.

In this new CCR approach, each specific class is characterized by the corresponding class center (i.e. prototype) computed from the training data. The center of a meta-class is calculated based on the centers of specific classes included in the meta-class. In the multi-class classification problem, there are usually only few (not all) classes that partly overlap, and most classes that are in fact far from each other can be well separated. The meta-class defined by the union of the classes that are far from each other are not useful in such applications. In order to reduce the computational complexity, we just need to select the useful meta-classes according to the context of the application under concern. The belief mass assignment of the object to classify with each specific class is determined based on the Mahalanobis distance between the object and the corresponding specific class center. Intuitively, the object committed to a specific class should be very close its center. If the object to classify is assigned to a meta-class, it means that the true class of the object is among the specific classes included in the meta-class but we don't know which one precisely. The ratio of the maximum distance of the object to the involved specific classes' centers, over the minimum distance, is introduced to measure the degree of distinguishability of these classes. Thus, the belief mass of a meta-class is determined from the distance between the object and the center of meta-class and its corresponding ratio value. An object will be committed to a meta-class with a high belief mass as soon as it is located at (almost) the same distances of several specific classes centers. Because in that case, it means that the object is very difficult to be correctly classify into a specific class. CCR provides credal classification results with low

computational burden due to the simple working principle.

We state in Section 4.2 the principles of CCR and the mathematical computation of BBA's for the credal classification. In Section 4.3, we present some classification results based on artificial and real data sets, and we compare the performances of the CCR with respect to well-known classification methods. Conclusions are given in Section 4.4.

4.2 Credal classification rule (CCR)

In this section we present in details the Credal Classification Rule (CCR) for classifying uncertain data. CCR provides a simple and an efficient way to compute the belief mass of the assignment of the object with the specific classes, with the meta-classes (which characterize the partial imprecision of class), and with the outlier class (i.e. the full ignorant class). The CCR consists of two main steps: 1) the determination of centers of the specific and meta classes, and 2) the construction of the BBA's based on the distances between the object and each class center.

4.2.1 Determination of the centers of classes

Let us consider a given set of data $Y = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n\}$, where the vectors $\mathbf{y}_i$ $(i = 1, \ldots, n)$ have to be classified over a frame of discernment $\Omega = \{w_1, \cdots, w_h\}$ using the training data set $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_g\}$. The element w_0 is explicitly included in the frame Ω to represent the unknown extra class for the closure of the frame.

The center of each specific class of Ω can be obtained in many ways[1]. For instance, one can use a given data pdf model, or the average of training data, or the centers produced by an unsupervised clustering (estimation) method (e.g. FCM, EM, etc). In this work, the center of each specific class is simply defined by the mean value of the training data $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_g\}$ in the corresponding class. It is assumed that $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_h\}$ are given, and correspond to the centers of the specific classes $w_1, \ldots, w_h$. For $j = 1, 2, \cdots, h$, the center $\mathbf{c}_j$ is defined $\forall \mathbf{x}_i \in w_j$ by

$$\mathbf{c}_j = \frac{1}{S_j} \sum_{\mathbf{x}_i \in w_j} \mathbf{x}_i \tag{4.1}$$

where S_j is the number of training samples in the class w_j.

The interest of the credal classification is the taking into account of the meta-classes that are used to model the imprecision of the class of the object to classify. The clustering center of meta-class is usually defined by the simple mean value of the involved specific classes' centers as done in [23]. This is mainly for the convenience and simplicity of linear optimization of the objective function. In fact, the arithmetic mean value of the specific classes' centers generally does not take the same distance to the each center of the associated specific class. Because of this, the centers of the involved specific classes is not precisely indistinguishable for the center of the meta-class according to the distance measure. In this work, we propose a new method to determine the center of the meta-classes, which should fairly reflect the real impossibility to distinguish the involved specific classes for the object belonging to this meta-class.

Basically in our approach, an object to classify will be committed to a meta-class (e.g. $w_i \cup w_j \ldots \cup w_k$), as soon as all the specific classes (e.g. w_i, w_j, ..., w_k) become undistinguishable for this object according to the distance measures. Therefore, we argue that the center of a meta-class

[1] It is worth noting that there is no class center corresponding to the outlier class w_0. The meta-classes involving w_o do not enter in CCR because w_0 plays the role of the default (closure) class which will contain all data points that cannot be reasonably associated within $2^{\Theta \setminus \{w_o\}}$.

must be located at the same distances of all the centers of the specific classes included in the meta-class under consideration.

For instance, let us consider the simplest meta-class (e.g. $U = w_i \cup w_j$) having a cardinality equals to two, e.g. $|U| = 2$. The meta-class center, denoted $\mathbf{c}_U$, should be at the same distance to all the specific classes' centers include in U, which are c_i and c_j. Therefore, the following condition must be satisfied

$$d(c_U, c_i) = d(c_U, c_j) \tag{4.2}$$

Eq. (4.2) represents only one constraint, and it can produce only one solution of $\mathbf{c}_U$ when the dimension of the vector $\mathbf{c}_U$ (i.e. the number of the attributes of data) is exactly one. If the dimension of $\mathbf{c}_U$ is bigger than one, there are many possible solutions for $\mathbf{c}_U$. Then, we will select the one which is closest to all the centers of the specific classes included in U, and given by $\arg[\min_{\mathbf{c}_U} \sum_{w_j \in U} (d(\mathbf{c}_U, \mathbf{c}_j))]$ because the meta-class center should be also simultaneously close to all the involved specific classes as much as possible.

It is worth noting that Mahalanobis distance (i.e. the normalized Euclidean distance) is used in this work to deal with the anisotropic data sets. This distance between two vectors $\mathbf{c}_U$ and $\mathbf{c}_i$ is given by:

$$d_{Ui} \triangleq d(\mathbf{c}_U, \mathbf{c}_i) = \sqrt{\sum_{k=1}^{N} \frac{(\mathbf{c}_U(k) - \mathbf{c}_i(k))^2}{\delta_i(k)^2}} \tag{4.3}$$

where N is the number of dimensions (attributes/features) of $\mathbf{c}_U$ and $\mathbf{c}_i$, and $\delta_i(k)$ is the standard deviation of the training data of class w_i in its k-th dimension.

The object committed to a meta-class (e.g. $U = w_i \cup w_j$) indicates that it must truly belong to one of the specific classes included in this meta-class, but these specific classes are not very distinguishable for this object. So the meta-class center should be closer to the centers of these involved specific classes than to other incompatible classes' centers[2]. Therefore, the following condition must be satisfied

$$\max_{w_i \in U} d_{Ui} < \min_{w_j \notin U} d_{Uj} \tag{4.4}$$

If the condition given by Eq. (4.4) is fulfilled, then one considers that the meta-class U must be kept as a potential solution of the classification, i.e. as a focal element of the BBA. Otherwise, if the meta-class center is closer to the center of an incompatible specific class $w_k \notin U$ than to the center of a specific class $w_i \in U$, it indicates that the objects close around the center c_U should belong more likely to the specific class $w_k \notin U$ rather than to $w_i \in U$. In such case, this meta-class U cannot be considered as effective[3] for the classification solution, and the center $\mathbf{c}_U$ should be eliminated to reduce the computational burden by reducing the number of focal elements of the BBA.

Figure 4.1 illustrates the selection of the meta-class. One considers a three classes problem with $\Omega = \{w_1, w_2, w_3\}$ and the corresponding set of centers $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ as shown on Fig. 4.1. One sees that the class w_2 partly overlaps w_1 and w_3, whereas w_1 and w_3 are well separated. The meta-class center $\mathbf{c}_{1,2}$ is more close to $\mathbf{c}_1$ and $\mathbf{c}_2$ than to $\mathbf{c}_3$. So the meta-class[4] $w_1 \cup w_2$ is

[2] The elements A and B are considered incompatible if $A \cap B = \emptyset$, and compatible if $A \cap B \neq \emptyset$.

[3] because it is very likely that the specific classes (e.g. w_i and w_j) in U are separated by the class w_k. The classes w_i and w_j in fact do not overlap, and so none object need to be assigned to the meta-class $U = w_i \cup w_j$.

[4] Actually according to the Fig.4.1, the overlapped zone between w_1 and w_2 should better correspond to $w_1 \cap w_2$. Because in this work we don't allow an object to belong simultaneously to several distinct classes, the object in the overlapped zone is supposed to belong to one class only, but we cannot exactly determine precisely which class (w_1 or w_2). So the meta-class $w_1 \cup w_2$, which is more coherent with our interpretation here than $w_1 \cap w_2$, is used to represent the overlapped zone.

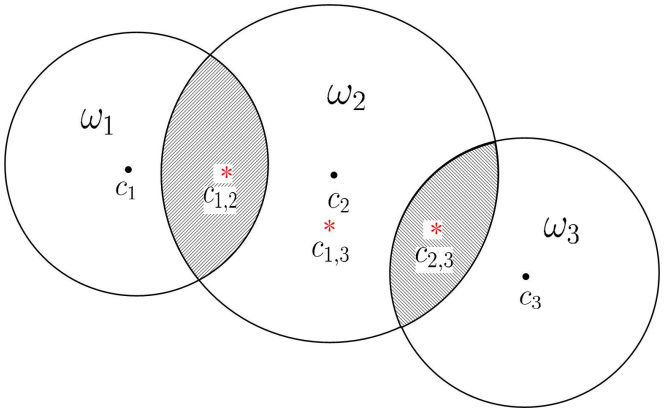


Figure 4.1 : Simple illustration of the meta-class selection.

considered acceptable and its center $\mathbf{c}_{1,2}$ should be kept for the determination of the mass of belief. For a similar reason, the meta-class $w_2 \cup w_3$ with the center $\mathbf{c}_{2,3}$ is also acceptable. However, the center $\mathbf{c}_{1,3}$ is closer to the incompatible class' center$\mathbf{c}_2$ than to $\mathbf{c}_1$ and to $\mathbf{c}_3$ which indicates that the objects around $\mathbf{c}_{1,3}$ will more likely belong to the class w_2. Consequently, the meta-class $w_1 \cup w_3$ will not be taken into account in the credal classification, and its center $\mathbf{c}_{1,3}$ is useless for the determination of the bba because one will take $m(w_1 \cup w_3) = 0$ in that case.

Let us consider the more general situation with the cardinality value of the meta-class bigger than two (i.e. $|U| \geq 3$). If the meta-class U is accepted[5] in the credal classification, it indicates that all the specific classes included in U should be undistinguishable for the objects committed to this meta-class. So all the subsets (i.e. the sub-meta-classes) of U should be also acceptable before entering the calculation of meta-class center $\mathbf{c}_U$. If one meta-class $A \subset U$ is considered unacceptable, it means that several specific classes in A can be distinguished by all the objects, and there is no necessity to preserve the meta-class A as a focal element of the BBA. In that case, the meta-class U of course becomes unacceptable (useless), and we do not need to calculate its center. If all the subsets of U are acceptable, then one can go for the computation of the center $\mathbf{c}_U$ to determine $m(U) > 0$.

Because the center $\mathbf{c}_U$ must be located at the same Mahalanobis distance with respect to all centers of the specific classes included in U, the following conditions must hold

$$d(\mathbf{c}_U, \mathbf{c}_i) = d(\mathbf{c}_U, \mathbf{c}_j), \forall w_i, w_j \in U, i \neq j. \tag{4.5}$$

Since one can obtain a set of $|U| - 1$ independent constraints from Eq. (4.5), there will be only one solution of $\mathbf{c}_U$ when the number of the available attributes of data is equal to $|U| - 1$. If the number of attributes is bigger than $|U| - 1$, there exist many solutions for $\mathbf{c}_U$. If so, we will select the solution which is closest to all the centers of the specific classes included in U, and given by $\arg[\min_{\mathbf{c}_U} \sum_{w_j \in U} (d(\mathbf{c}_U, \mathbf{c}_j))]$. If the dimension of $\mathbf{c}_U$ is smaller than $|U| - 1$, one has to solve an optimization problem to seek the solution for $\mathbf{c}_U$ that should be satisfied with all the constraints as much as possible, such as for $\forall w_i, w_j \in U, i \neq j$, $\arg[\min_{\mathbf{c}_U} \sum_{w_i, w_j \in U} (d(\mathbf{c}_U, \mathbf{c}_i) - d(\mathbf{c}_U, \mathbf{c}_j))^2]$. This can be done using any classical nonlinear optimization method. In this work, we seek the solution using the classical nonlinear least squares estimate method [120].

[5] then U is a focal element of the BBA.

Moreover, $\mathbf{c}_U$ should be also satisfied with the constraint given by Eq. (4.4). Otherwise, this meta-class cannot be included in the credal classification results. In real applications, many unacceptable meta-classes will be eliminated through this step, and we just consider only the selected acceptable meta-classes as true focal elements of the BBA. By doing so, we greatly reduce the computational complexity, which is very appealing for most engineering applications.

4.2.2 Construction of BBA's

Let us consider a particular object $\mathbf{y}_s \in Y, s = 1, \ldots, n$ to classify over the frame of discernment $\Omega = \{w_0, w_1, \ldots, w_h\}$ using the framework of the belief functions. The mass of belief of the specific class (e.g. w_i) should depend on the Mahalanobis distance between the object and the corresponding center of class, and the bigger distance generally leads to the smaller mass of belief. If $\mathbf{y}_s$ is closer to a specific class center (e.g. $\mathbf{c}_i$), it indicates that $\mathbf{y}_s$ belongs very likely to the class w_i as done in the classical way. So the initial mass of $\mathbf{y}_s$ of a singleton class should be a monotone decreasing function (denoted by $f_1(\cdot)$) of the distance between the object and the corresponding class center, which is denoted

$$\tilde{m}(w_i) = f_1(d(\mathbf{y}_s, \mathbf{c}_i)), \forall i = 1, \ldots, h \tag{4.6}$$

The credal classification approach offers the possibility that the object belongs with different masses of belief to all the specific classes, and also to some meta-classes as well. The meta-classes are introduced as a means for modeling the imprecision of the class of the object. To reduce the computational burden, we have shown in the previous step devoted to the determination of meta-class center how some unacceptable meta-classes can be reasonably ignored. Moreover, we can reasonably assume that the object is close to the true class it belongs to in general. Consequently, the object should not very likely belong to the classes very far away of its true class. Based on this remark, we also consider (for the construction of the BBA) the compatibility of the meta-classes according to the ascending order of the distances between the object and all the centers of specific classes.

If the specific classes are listed in the ascending order of the distances of $\mathbf{y}_s$ to the centers as $(w_i, w_j, w_k, \ldots, w_g)$. It means that $\mathbf{y}_s$ belongs most likely to w_i, then to w_j, w_k, ..., w_g. Thus, we just need to consider only the nested meta-classes $w_i \cup w_j$, $w_i \cup w_j \cup w_k$, ..., $w_i \cup w_j \cup w_k \ldots \cup w_g$ because the object $\mathbf{y}_s$ will not very likely belong to the other meta-classes.

For example, let us consider the simple frame $\Omega = \{w_1, w_2, w_3\}$, and an object $\mathbf{y}_s$ which is the most close to w_1, and then to w_2, and then to w_3. In that case, we select only the following nested meta-classes $w_1 \cup w_2$ and $w_1 \cup w_2 \cup w_3$ as potential focal elements. The meta-class $w_2 \cup w_3$ is not considered as compatible because the true class of $\mathbf{y}_s$ cannot be reasonably compatible with $w_2 \cup w_3$ only (because $\mathbf{y}_s$ is in fact the most close to w_1). Moreover, if some selected compatible nested classes appear finally unacceptable (according to the step of meta-class center determination), they will be ignored in the construction of the BBA for the credal classification of the object.

Once all the acceptable meta-classes of the object $\mathbf{y}_s$ have been determined, we can proceed with the computation of the mass of these meta-classes (i.e. the focal elements of cardinalities greater than one). The principle of construction of the mass for a meta-class U is based on the following considerations:

- If an object is committed to a meta-class U, then of course it should be very close to the center $\mathbf{c}_U$ of this meta-class.
- the ratio $\gamma = d_{\max}/d_{\min}$ of the maximum distance $d_{\max}$ of the object to the centers of the specific classes included in U over the minimum distance $d_{\min}$ is used to measure the degree of

distinguishability among the classes in U. The smaller ratio indicates a poor distinguishability degree among the classes in U from the object. The object committed to a meta-class must have a small ratio value (close to one) indicating that the involved specific classes are not very distinguishable for the object. So the value of the ratio γ will be used to put more or less mass of belief to the meta-class U.

Based on these considerations, the mass of belief of assignment of the object $\mathbf{y}_s$ with the meta-class U is mathematically defined as

$$\tilde{m}(U) = f_2(d(\mathbf{y}_s, \mathbf{c}_U), \gamma_U) \tag{4.7}$$

where

$$d(\mathbf{y}_s, \mathbf{c}_U) = \frac{1}{|U|} \sum_{w_i \in U} \sqrt{\sum_{k=1}^{N} \frac{(\mathbf{y}_s(k) - \mathbf{c}_U(k))^2}{\delta_i(k)^2}} \tag{4.8}$$

$$\gamma_U = \frac{\max_{w_i \in U} d(\mathbf{y}_s, \mathbf{c}_i)}{\min_{w_i \in U} d(\mathbf{y}_s, \mathbf{c}_i)} \tag{4.9}$$

The smaller value of $d(\mathbf{y}_s, \mathbf{c}_U)$ and γ_U will yield bigger mass of belief $\tilde{m}(U)$, and vice versa. Hence, $f_2(.)$ should be a monotone decreasing function with respect to $d(\mathbf{y}_s, \mathbf{c}_U)$ and γ_U.

To get good results, the functions $f_1(.)$ and $f_2(.)$ must be determined according to the application under concern. Unfortunately, we do not have found yet general guidelines for the selection of these functions. Here, we have chosen the exponential decreasing function because it is commonly used in many engineering applications [12, 117].

In summary, the (unnormalized) masses of belief for the specific classes and the acceptable meta-classes are finally given by:

$$\tilde{m}(w_i) = e^{-d(\mathbf{y}_s, \mathbf{c}_i)} \tag{4.10}$$

$$\tilde{m}(U) = e^{-\lambda_U \gamma_U d(\mathbf{y}_s, \mathbf{c}_U)}, \quad \text{for } |U| \geq 2 \tag{4.11}$$

with $\lambda_U = \eta |U|^{\alpha}$. The quantity $|U|^{\alpha}$ is a penalized parameter for the meta-classes having a big cardinality value. In most cases, the classification of the object is imprecise only among a small number of specific classes, and there are usually only few objects to assign with the meta-classes having big cardinalities. Thus, bigger cardinalities generate stronger penalization. η is a tuning parameter used to control the number of objects committed to the meta-classes. In practice, we always have to find a good compromise between the error rate and the imprecision rate. The guidelines for tuning the parameters are given at the end of this Section.

The outlier class w_0 is also taken into account to deal with the case where the potential outliers (noise) can be involved. The object will be considered as outlier if it is far from all the other classes according to a given outlier threshold t. The mass of the object in the outlier class is defined by:

$$\tilde{m}(w_0) = e^{-t} \tag{4.12}$$

$\forall A \subseteq \Omega$, the previous unnormalized mass of belief $\tilde{m}(.)$ is normalized as follows

$$m(A) = \frac{\tilde{m}(A)}{\sum_{B \subseteq \Omega} \tilde{m}(B)} \tag{4.13}$$

This normalized BBA $m(.)$ is then used for the credal classification of the object $\mathbf{y}_s$.

• **Guidelines for tuning the parameters in the CCR approach**

The parameters α, η and t can be optimized using the training data with the cross-validation method (e.g. leave-one-out) before the application of CCR. The bigger penalized parameter α will lead to smaller number of the objects in the meta-class with big cardinality, and the suitable value can be found according to the classification results of the training data. Generally, one can take $\alpha \in [1, 3]$, e.g. 1 or 2. The parameter η is used to control the number of objects in the meta-classes. The bigger value of η will produce smaller number of objects committed to the meta-classes. It is recommended to take $\eta \in (0, 1)$, but the exact value of η can be tuned according to the imprecision degree (i.e. the rate of the objects in the meta-classes) of the classification results one can accept in the application under concern. The outlier threshold t should be determined according to the outlier rate one expects in the classification. The bigger t will cause smaller number of outliers, and we generally recommend to take $t \in [2, 5]$.

4.3 Evaluation of CCR on artificial and real data sets

In this section we present four experiments to evaluate and compare the performances of CCR with respect to four classical methods: 1) the Classification And Regression Tree (CART) [62], 2) the Artificial Neural Networks (ANN) [61], 3) the EK-NN [12], and 4) the BCKN. The experiment 4.1 based on artificial data sets, is presented to show the difference between the credal classification and the classical methods. The experiment 4.2 allows to evaluate the performance of CCR with respect to the other methods based on a 4-class artificial data set. The experiment 4.3 is used to illustrate the efficiency of CCR for dealing with the large scale data set. The experiment 4.4 is based on four real-data sets from UCI [95]. It shows the advantage of CCR over the other methods. The different methods in the experiments have been programmed and tested with Matlab^{TM} software.

In order to show the ability of CCR to deal with the meta-classes, the class of each object is decided according to the maximal mass of belief criterion. In CCR, the error rate Re and imprecision rate Ri_j defined in the same way as in last chapter are introduced to evaluate the performance of CCR.

Please note that in Fig. 4.2 and 4.3, the x-axis and y-axis respectively represent the first and the second dimension of test and training data.

4.3.1 Experiment 4.1 (with artificial data sets)

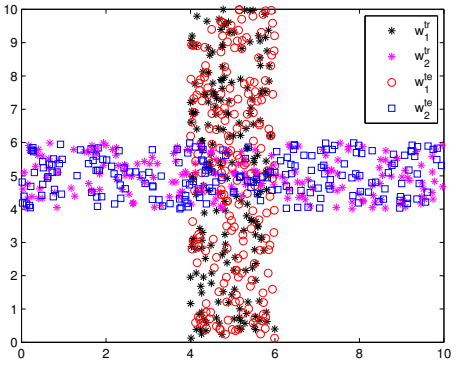
4.3.1.1 Test 1: A 2-class problem with artificial data

Two classes of artificial data set w_1 and w_2 are obtained from two uniform distributions as shown by Fig. 4.2-(a). Each class has 200 training samples and 200 test samples. The uniform distributions of the samples of the two classes are characterized by the following bounds:

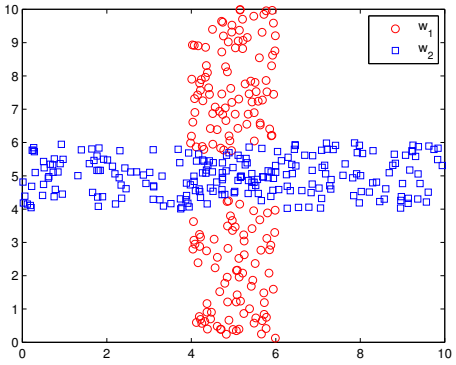
	x-label interval	y-label interval
w_1	(4, 6)	(0, 10)
w_2	(0, 10)	(4, 6)

Three classical classifiers (CART, ANN and EK-NN) are compared with the proposed CCR method. A particular value of $K = 9$ is selected here for EK-NN, since it provides good results. The other parameters in EK-NN are optimized using the method introduced in [118]. In ANN, we

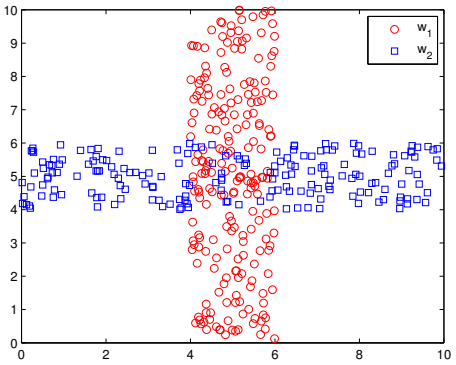
used the feed-forward back propagation network with $epochs = 500$ and $goal = 0.001$ in all the experiments. In CCR, one has taken $\alpha = 1$, $t = 2$, and tested two different values of η to show its influence on the performances of CCR. The classification results of the objects with the different methods are given in Fig. 4.2-b–4.2-f. For notation conciseness, we have denoted $w^{te} \triangleq w^{test}$, $w^{tr} \triangleq w^{training}$ and $w_{i,\ldots,k} \triangleq w_i \cup \ldots \cup w_k$.



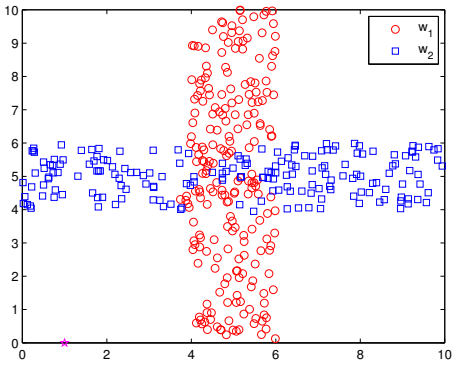
(a) Original data.



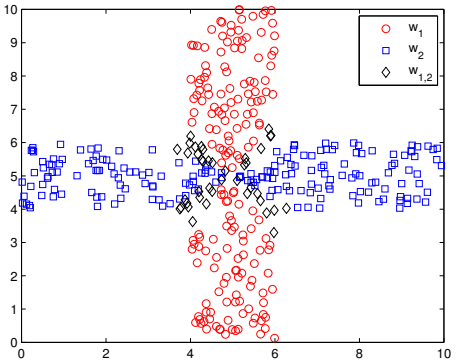
(b) Classification result by ANN ($Re = 9.00$).



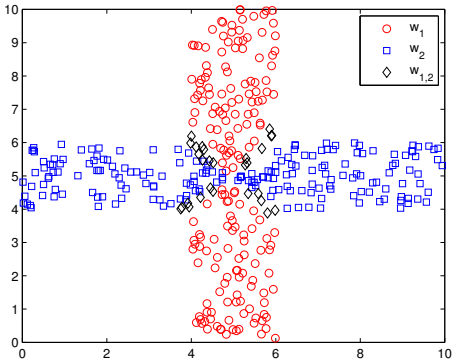
(c) Classification result by CART ($Re = 10.75$) .



(d) Classification result by EK-NN ($Re = 12.25$).



(e) Classification result by CCR with $\eta = 0.25$ ($Re = 4.25, Ri_2 = 10.25$) .



(f) Classification result by CCR with $\eta = 0.3$ ($Re = 5.25, Ri_2 = 7.00$).

Figure 4.2 : Classification results obtained by different methods for a 2-class problem.

The misclassification rate obtained by the different methods is indicated in the title of each subfigure. The objects of classes w_1 and in w_2 are distributed over two overlapping areas following a cross shape as shown in Fig. 4.2-(a). Obviously, all objects belonging to the middle of the cross area are really difficult to associate with a particular class. However, EK-NN, CART and ANN just commit these objects into a specific class w_1 or w_2 as shown in Fig. 4.2-(b)-(d). Such classification methods generate many misclassification errors (the error rate is about ten percent). CCR provides one more meta-class $w_1 \cup w_2$ as shown in Fig. 4.2-(e),(f). The classes w_1 and w_2 are undistinguishable for all the objects located in the intersecting (overlapping) zone. Thus, it is more judicious and prudent to assign these objects to the meta-class $w_1 \cup w_2$. By doing this, one greatly reduces the number of misclassification, and also deeply reveals the imprecision degree of class of the objects. Once the tuning parameter $\eta = 0.25$ increases to $\eta = 0.3$, the imprecision rate will decrease but meanwhile the error rate will increase. So one should find a good compromise between the error rate and imprecision rate by tuning η in the training data space.

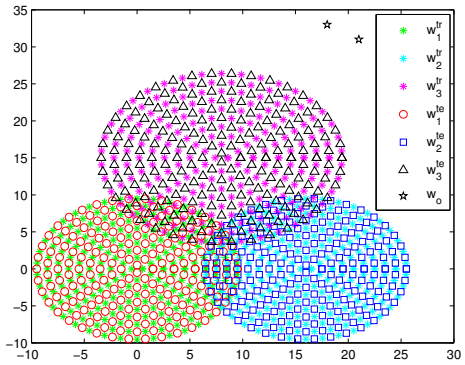
4.3.1.2 Test 2: A 3-class problem with artificial data

In this second test, we consider a particular 3-classes data set in a round shape as shown in Fig. 4.3-(a). This data set consists of 615 training data points and 617 test data points including two noisy data (outliers). The radii of the circles for w_1, w_2 and w_3 are $r_1 = 10, r_2 = 10, r_3 = 12$. The centers of three circles are located at $\mathbf{c}_1 = (0, 0)$, $\mathbf{c}_2 = (16, 0)$ and $\mathbf{c}_3 = (8, 15)$. CCR is applied for the classification of this particular data set and it is compared with the CART, ANN and EK-NN classification methods. A particular value of $K = 9$ is also selected in EK-NN. In CCR, we have chosen the tuning parameters $\alpha = 2$, $t = 2$ and $\eta = 0.4$. The classification results obtained by the different methods are shown in Fig. 4.3-(b)-(e).

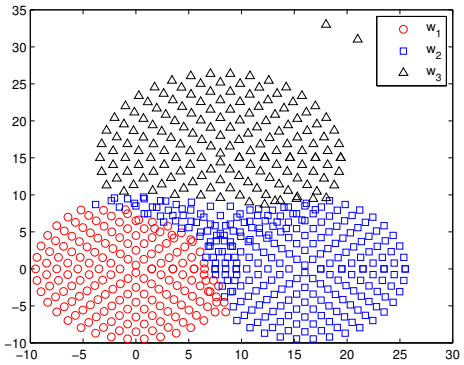
The error rate and the imprecision rate of classification results obtained by the different methods are also given in the title of each subfigure. In Fig. 4.3-(a), one sees that the three classes w_1, w_2 and w_3 partly overlap on their borders, and the points belonging to the overlapped zones are really difficult to classify correctly due to their ambiguity. Moreover, two noisy points far from the other data are included in the test data set. As shown in Fig. 4.3-(b)-(d), ANN, CART and EK-NN produce only three singleton clusters w_1, w_2 and w_3. Thus, most of the points in the overlapped zone are probably misclassified because of the inherent limitation of the framework adopted for these methods. These classifiers cannot detect the noisy data (outliers), and they all commit the noisy data into the class w_3. CCR produces more reasonable credal classification results in comparison with other methods. The points in the middle of w_1 and w_2, w_2 and w_3 and w_1 and w_3 are respectively committed to $w_1 \cup w_2$, $w_2 \cup w_3$ and $w_1 \cup w_3$ as shown in Fig. 4.3-(e) because these points are really difficult to classify correctly into a particular class. All of the three classes overlap in their middle, and the points in this zone are prudently committed to the meta-class $w_1 \cup w_2 \cup w_3$ because their classes are totally imprecise with respect to w_1, w_2 and w_3. CCR is also able to well detect the outliers. This example clearly shows the potential interest of the credal classification done by this new CCR approach.

4.3.2 Experiment 4.2 (with artificial data sets)

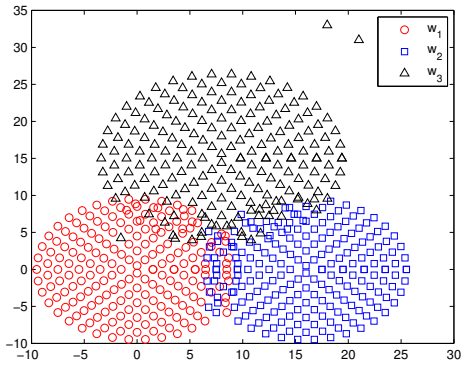
In this second experiment, the statistics of the performances of CCR are compared with CART, ANN, EK-NN and BCKN on a 4-class artificial data set, which is generated from four 2D Gaussian



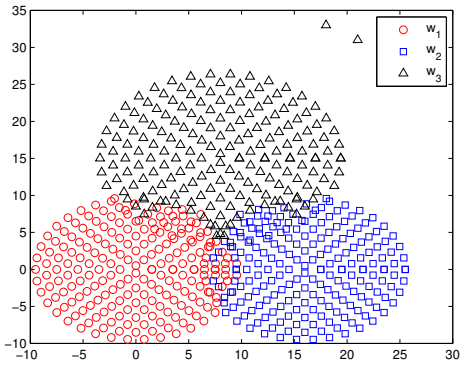
(a) Original data.



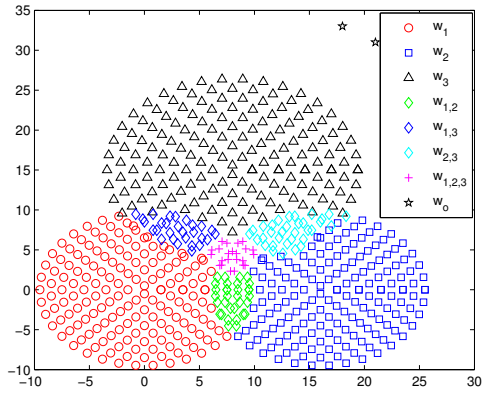
(b) Classification result by ANN ($Re = 14.42$).



(c) Classification result by CART ($Re = 10.86$) .



(d) Classification result by EK-NN ($Re = 12.68$).



(e). Classification result by CCR with $\eta = 0.4$ ($Re = 1.13, Ri_2 = 15.07, Ri_3 = 3.73$).

Figure 4.3 : Classification results by different methods for a 3-class problem.

distributions characterizing the classes w_1, w_2, w_3 and w_4 with the following means vectors

$$\mu_1 = (0, 0), \mathbf{\Sigma}_1 = 2 \cdot \mathbf{I}$$

$$\mu_2 = (7, 0), \mathbf{\Sigma}_2 = 2.5 \cdot \mathbf{I}$$

$$\mu_3 = (15, 0), \mathbf{\Sigma}_3 = 3 \cdot \mathbf{I}$$
$$\mu_4 = (22, 0), \mathbf{\Sigma}_4 = 2 \cdot \mathbf{I}$$

There are 3×200 test objects, and the training sets contain $3 \times N$ samples (for $N = 200, 300, 500$).

For EK-NN and BCKN methods, the values of K ranging from 5 to 15 neighbors have been tested. The error rates Re, the imprecision rates Ri_j, and the computation time (in seconds) have been averaged over 10 Monte Carlo runs (i.e. 10 independent random generation of the data sets). The results obtained with the different classifiers are shown in Table 4.1. The BCKN and CCR have been tuned to get a good compromise between the misclassification error and the imprecision of the results.

Table 4.1 : Classification results for a 4-class problem with different methods (in %).

		N=200	N=300	N=500
ANN	Re	13.40	12.73	13.08
	time	10.8327	11.8592	13.0573
CART	Re	14.17	14.23	14.55
	time	0.0265	0.0374	0.0546
EK-NN	Re	11.20	11.02	10.97
	time	0.5171	0.6518	1.1950
BCKN	Re	10.19	9.74	9.26
	Ri_2	1.50	2.28	2.95
	time	3.9126	5.9837	9.7273
CCR	Re	8.45	8.06	7.50
	Ri_2	6.53	6.20	6.12
	time	0.0125	0.0140	0.0156

The meta-classes with cardinalities bigger than two are not considered in this application, that is why we did just mention Ri_2 in Table 4.1. One sees in Table 4.1 that CCR and BCKN produce smaller error rate than other methods. This is normal because the objects that are difficult to classify correctly have been assigned to the associated meta-classes. In general, the classification results of BCKN and CCR are similar. The error rate for CCR is a bit lower than for BCKN, but in counterpart the imprecision rate for BCKN is lower than for CCR. However, BCKN requires much more computational time than CCR, which shows that the computational burden of BCKN is much bigger than CCR. CCR consumes much less time than any other tested methods which indicates that CCR has the least computational complexity which offers a strong advantage for some engineering applications with respect to other methods.

4.3.3 Experiment 4.3 (with large scale artificial data sets)

The performance of CCR for dealing with large scale data sets (i.e. big number of samples with high-dimensional features) is evaluated in this experiment by comparing CCR with several other classical methods[6] (ANN, CART and EK-NN).

In this experiment, an artificial data set with four class w_1, w_2, w_3 and w_4 is generated from four 30D Gaussian distributions with the means vectors and covariance matrices as follows:

$$\mu_1 = \mathbf{zeros}(\mathbf{1}, \mathbf{30}), \mathbf{\Sigma}_1 = 10 \cdot \mathbf{I}$$

[6]It is well known that the K-NN based methods (e.g. EK-NN, BCKN, etc) are usually not very effective for dealing with the big data set due to the large computation burden. We have shown that BCKN can produce results similar to CCR, but it requires more computational time than CCR, EK-NN and CART. So we just use the EK-NN method here to compare its performance with CCR.

$$\mu_2 = 5 \cdot \mathbf{ones}(\mathbf{1}, \mathbf{30}), \mathbf{\Sigma}_2 = 10 \cdot \mathbf{I}$$

$$\mu_3 = 20 \cdot \mathbf{ones}(\mathbf{1}, \mathbf{30}), \mathbf{\Sigma}_3 = 15 \cdot \mathbf{I}$$

$$\mu_4 = 30 \cdot \mathbf{ones}(\mathbf{1}, \mathbf{30}), \mathbf{\Sigma}_4 = 15 \cdot \mathbf{I}$$

Here $\mathbf{zeros}(\mathbf{1}, \mathbf{30})$ represents the 30-dimensional vector with value of zero in each dimension, and $\mathbf{ones}(\mathbf{1}, \mathbf{30})$ is the 30-dimensional vector with value of one in each dimension, and $\mathbf{I}$ denotes the 30×30 identity matrix.

In each class, we use the same number (i.e. n) of training samples and test samples. So there are totally $N = 4 \times n$ training samples and $N = 4 \times n$ test samples, and we take $N = 8000, 40000, 200000, 1000000$. In EK-NN, the values of K ranging from 5 to 15 neighbors are tested. The parameters have been tuned to get a good compromise between the misclassification error and the imprecision of the results by CCR. The error rates Re, the imprecision rates Ri_j, and the computation time (in seconds) are the average value over 10 Monte Carlo runs. The results produced by the different classifiers are illustrated in Table 4.2 where 'NA' means 'Not Applicable'.

Table 4.2 : Classification results large scale data with different methods (in %).

	ANN (Re, time)	CART (Re, time)	EK-NN (Re, time)	CCR (Re, Ri_2, time)
N=8000	(33.09, 15.6313)	(29.59, 1.2168)	(8.46, 47.5023)	(5.26, 5.84, 0.2340)
N=40000	(35.04, 58.9684)	(26.66, 6.4428)	(8.25, 1669.1)	(5.15, 6.41, 1.1544)
N=200000	(33.93, 241.7703)	(24.34, 35.1470)	NA	(5.11, 6.24, 5.8032)
N=1000000	NA	(22.25, 200.3053)	NA	(5.14, 6.16, 29.0162)

We can see that CCR produces the lowest error rate with some partial imprecision results, since it assigns some objects that are hard to be correctly classified into the proper meta-classes. Meanwhile, CCR consumes the shortest operation time. EK-NN can obtain the reasonable classification results, but it requires the longest running time, which is the main drawback of the K-NN based methods. EK-NN is even not applicable when the number of samples is big (i.e. N=200000 and N=1000000), since it takes too long time, which is not convenient in many cases where the high speed of execution is necessary. ANN and CART cause much higher error rate than CCR and EK-NN, and they are also much more time-consuming than CCR. ANN is not applicable for the big data set (i.e. N=1000000) because of its high computational burden. So it indicates that CCR is effective for dealing with the large scale data set thanks to its low computational and complexity burden.

4.3.4 Experiment 4.4 (with real data sets)

Four well-known real data sets obtained from UCI Machine Learning Repository [95] (the Iris, Seeds, Wine and Yeast data sets) have been tested in this experiment to evaluate the performances of CCR compared with CART, ANN, EK-NN and BCKN. For the Yeast data set, three classes named as CYT, NUC and ME3 are selected here, since these three classes are close and difficult to discriminate. The main characteristics of the four data sets are summarized in Table 4.3 below. All the detailed information can be found on UCI repository archive at `http://archive.ics.uci.edu/ml/`.

The k-fold cross validation is performed on the four data sets by different classification methods. In previous chapter on BCKN, the classification results of the classifiers (i.e. K-NN, EK-NN, CART, SVM, ANN, SVM, BCKN) have already been shown using the 10-fold cross validation. In

Table 4.3 : Basic information of the real data sets

name	classes	attributes	instances
Iris	3	4	150
Seeds	3	7	210
Wine	3	7	255
Yeast	3	9	683

this chapter, we use the different 2-fold cross validation[7] here, since it has the advantage that the training and test sets are both large, and each sample is used for both training and testing on each fold. The tuning parameter of CCR and BCKN were optimized using the training samples. The classification results including Re and Ri_j of BCKN and EK-NN are calculated with values of K ranging from 5 to 15. The reported error rates Re, the imprecision rates Ri_j, and the computation time (in seconds) for the different methods are given in Table 4.4.

Table 4.4 : Classification results of real data with different methods (in %).

		Iris	Seeds	Wine	Yeast
ANN	Re	4.00	16.665	33.71	51.42
	time	4.5708	9.3289	8.4553	9.9841
CART	Re	5.33	11.90	8.89	37.73
	time	0.0156	0.0234	0.0312	0.0936
EK-NN	Re	3.98	10.57	28.94	36.51
	time	0.0094	0.0296	0.0135	0.2366
BCKN	Re	4.00	8.66	25.81	27.46
	Ri_2	0	3.33	4.39	14.22
	time	0.0348	0.0803	0.0668	1.6214
CCR	Re	2.00	7.14	3.37	25.31
	Ri_2	6.67	6.19	0	19.43
	time	0.0000	0.0000	0.0078	0.0156

In these tests, none object is committed to the meta-class with cardinality value of three, and that is why we have just given Ri_2 in Table 4.4. From the table 4.4, one sees that CCR and BCKN produce the smaller error rate than other classical methods. It is normal because the objects difficult to classify correctly have been reasonably and automatically committed to the associated meta-classes. It shows that the credal classification can effectively reduce error occurrences, and the meta-classes indicate that the attributes information is not good enough to obtain the correct specific class of some objects. In that case, some other complementary sources of information, or techniques, will be necessary if one wants to precisely discriminate the objects committed to the meta-classes with high belief mass value (if a precise classification is absolutely required). The CCR and BCKN methods provide similar performances for the Iris, Seeds and Yeast data sets according to the compromise between error rate and imprecision rate. For the Wine data set, CCR yields the lowest error rate due to its inherent working principle which is very different of the other classifiers. It is worth noting that BCKN requires a very long running time due to the heavy computational load. The proposed CCR method requires less computational time than the other methods. This shows again that CCR working with credal classification can deal efficiently with uncertain data using belief functions with a serious computational complexity advantage over other methods.

[7] More precisely, the samples in each classes are randomly assigned to two sets S_1 and S_2 having equal size. Then we train on S_1 and test on S_2, and reciprocally.

4.4 Conclusions

A new simple and effective credal classification rule (CCR) based on the belief functions has been proposed in this chapter to deal with the classification of uncertain data when the data classes can be well characterized using the prototype vectors. CCR strengths the robustness of results by reducing the misclassification errors thanks to the introduction of meta-classes. The CCR approach is also able to detect the outliers in the data sets. In CCR, each specific class corresponds to a center (i.e. prototype) obtained using the training data, and the center of meta-class is located at the equal Mahalanobis distances to all the centers of the involved specific classes. Mahalanobis distance is used here to deal with the anisotropic data sets. The acceptable meta-classes are selected according to the current context and distance ratios, and all the unacceptable meta-classes are automatically rejected to reduce the number of focal elements and the computational complexity. A tuning parameter has been introduced in CCR to control the number of objects in the meta-classes. The output of CCR can be used efficiently to alert the classification system designer that other complementary information sources are necessary to remove (or reduce) the ambiguity of the classification of some particular data points. Several experiments using both the artificial and real data sets have been presented to evaluate the performance of CCR with respect to other methods. Our results show that CCR is able to provide good credal classification results with a relatively low computational complexity with respect to other methods.

5 Credal classification of incomplete patterns

5.1 Introduction

Missing (unknown) data is a common problem encountered in the classification problem, and a number of methods [92, 93] have emerged for classifying incomplete data (pattern) with missing values. The estimation strategy [94] is usually adopted for missing values in many cases, and then the incomplete patterns with estimated values are classified. We develop a new method for classification of incomplete data based on the estimation of missing values. There exist many methods for estimating missing values. In the often used mean imputation (MI) method [100, 101], the missing values are simply replaced by the mean of all known values of that attribute. In the K-nearest neighbor imputation (KNNI) method [98, 99], the missing values are estimated using the K-nearest neighbors of the object (incomplete pattern), but KNNI requires a big computation burden. In fuzzy c-means imputation (FCMI) method [121,122], the missing values are filled based on the clustering centers produced by FCM and the distances between the object and the centers. There are also other methods of imputation, such as the SOM imputation [103], the regression imputation [96], the multiple imputation approach [102], etc. Most methods (except multiple imputation) produce only one precise estimation for the missing value, and they are not able to well reflect the uncertainty about the prediction of the missing values. In the multiple imputation method [102], the missing values are imputed M times to produce M complete data sets based on an appropriate model with random variation, but the model is hard to obtain sometimes. The multiple imputation approach mainly focus on the imputation of the missing values, whereas this work is devoted to the classification of incomplete pattern.

In some applications, the missing data of attribute may have several different possible estimated values, and the classification result of the incomplete pattern (test sample) with different estimations can be distinct. As a simple example, Fig. 5.1 shows a 2-class problem with 2 dimensions of attributes corresponding to x-coordinate and y-coordinate.

In the example of Fig. 5.1, it is assumed that the attribute values in x-coordinate are all missing, and these incomplete patterns can be denoted by $[?, y]$. Then, the classification of the objects mainly depends on the only attribute value in y-coordinate. For the objects labeled by red square, the class B and A are indistinguishable based on the known value in y-coordinate. If the estimation of missing value is below the upper bound of the B class in x-coordinate, these red square points are likely committed to class B. Whereas, if the estimated value is bigger than the lower bound of class A in x-coordinate, these points are very likely assigned to class A. It is similar for the objects labeled by green circles, which cannot be clearly distinguished by class A and C. Such conflict (uncertainty) of classification is caused by the lack of information of the missing values, and it is hard to correctly classify the object in such condition because the known (available) attributes information is really insufficient for making a specific classification. The belief function theory [3–7] is appealing for dealing with such uncertain and imprecise information [2]. In the previous works, the classification methods developed based on belief functions [76] were all

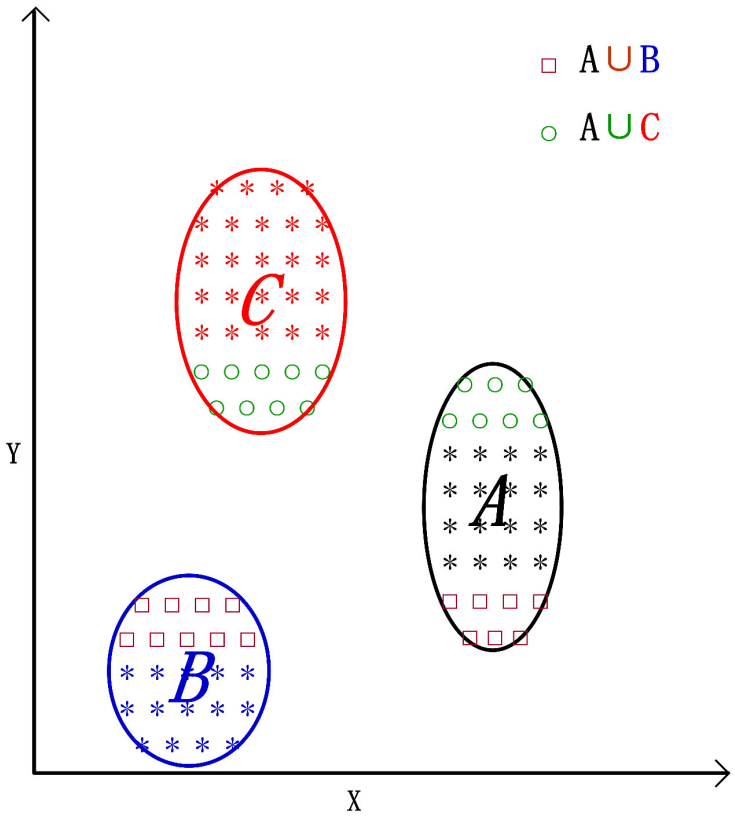


Figure 5.1 : Simple illustration of incomplete pattern classification.

designed for classifying complete patterns only, and the missing data aspect was not taken into account.

A new efficient prototype-based[1] credal classification (PCC) method for incomplete patterns working with belief function framework is proposed here. The missing values, which cannot be precisely determined from the incomplete available information, can play a crucial rule in the classification of the pattern, and different estimations of these missing values can lead to distinct classification results. If one uses the classical methods to commit the pattern to a particular class (based on the highest probability value), one will generate very likely misclassifications. In some applications, it is primordial to get a robust (even partially imprecise) classification result, which could be refined later by additional techniques, rather than to obtain the specific result with high risk of error that may bring fatal collateral damages. For this reason, PCC can well model such partial imprecision (uncertainty) thanks to meta-class introduced in credal classification. The object hard to be correctly classified due to the imprecision caused by missing values will be reasonably committed to the proper meta-class defined by the union (disjunction) of several specific classes (e.g. $A \cup B$ in Fig. 5.1) that the object likely belongs to. This approach allows us

[1]The estimation of missing data in this new method is based on the prototypes of the classes.

to reduce the misclassification error rate and to reveal the imprecision of classification.

In PCC, the prototypes of all classes obtained by the training data (complete pattern) are used to estimate the missing values of the incomplete pattern. Thus in a c-class problem, one has to deal with c estimations of the missing values. The object with each of the c estimated values will be classified using a standard classifier, and PCC will produce c pieces of classification results represented by basic belief assignments (BBA's). These c pieces of results have different weighting factors (determined by the distances between the object and the prototypes) playing the role of discounting factor of the BBA's. The global fusion of the c discounted results will be adopted for obtaining the final credal classification of this object. In the fusion process, meta-classes will be conditionally kept for the uncertain objects that are hard to correctly classify. Conflicting beliefs are very important to capture the imprecision (ambiguity) degree of classification, and they will be also selected and transferred to the corresponding meta-classes depending on the current context.

This chapter is organized as follows. The new prototype-based credal classification method is presented in the section 5.2. The proposed method PCC is then tested in section 5.3 and compared with several other classical methods, followed by conclusions.

5.2 Prototype-based Credal classification method

A new prototype-based credal classification (PCC) method is proposed in this chapter to deal with incomplete patterns based on evidential reasoning. PCC method provides multiple possible estimations of missing values according to class prototypes obtained by the training samples. For a c-class problem, it will produce c probable estimations. The object with each estimation is classified using any standard classifier working with complete patterns. Then, it yields c pieces of classification results, but these results take different weighting factors depending on the distance between the object and the corresponding prototype. So the c classification results should be discounted with different weights, and the discounted results are globally fused for the credal classification of the object. If the c classification results are quite consistent on the decision of class of the object, the fusion result will naturally commit this object to the specific class that is supported by the classification results. However, it can happen that high conflict among the c classification results occurs, and it indicates that the class of this object is quite imprecise (ambiguous) only based on the known attribute values. In such case, it is very difficult to correctly classify the object in a particular (specific) class, and it becomes more prudent and reasonable to assign the object to a meta-class (partial imprecise class) in order to reduce the error rate. The classification of the uncertain object in meta-class can be eventually precisiated (refined) using some other (costly) techniques or with extra information sources. So PCC approach prevents us to take erroneous fatal decision by robustifying the specificity of the classification result whenever it is essential to do it.

5.2.1 Classification of incomplete patterns with c estimations

Let us consider a test data set $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ to be classified using the training data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_H\}$ in the frame of discernment $\Omega = \{\omega_1, \ldots, \omega_c\}$. Because we focus on the classification of the incomplete data (test sample) in this work, one assumes that the test samples are all incomplete data (vector) with single or multiple missing values, and the training data set Y consists of a set of complete patterns[2].

[2]In some applications, there exist incomplete patterns in the training data set. If the training samples with missing values take a very small amount say less than 5%, they can be ignored in the classification. If the rate of the incomplete patterns is big, then the missing values are usually estimated at first, and the classifier will be trained using the edited set, i.e. complete data portion and incomplete patterns with estimated values. In this

The prototype of each class i.e. $\{\mathbf{o}_1, \ldots, \mathbf{o}_c\}$ is calculated using the training data at first, and $\mathbf{o}_g$ corresponds to class ω_g. There exists many methods to produce the prototypes. For example, the K-means method can be applied for each class of the training data, and the clustering center is chosen for the prototype. The simple arithmetic average vector of the training data in each class can also be considered as the prototype, and this method is adopted here for its simplicity. Mathematically, the prototype is computed for $g = 1, \ldots, c$ by

$$\mathbf{o}_g = \frac{1}{T_g} \sum_{\mathbf{x}_j \in \omega_g} \mathbf{x}_j \tag{5.1}$$

where T_g is the number of the training samples in the class ω_g.

Once each class prototype is obtained, we use the value of the prototype to fill the missing values of the object in the same attribute dimension. Because one has considered c possible classes with their prototypes, one gets c versions of estimated values. For the object $\mathbf{y}_i$ with some missing component values, the c versions of estimations of the missing component values y_{ij} of $\mathbf{y}_i$ are given by

$$y_{ij}^g = o_{gj} \tag{5.2}$$

where o_{gj} is the j-th component of the prototype $\mathbf{o}_g, g = 1, 2, \ldots, c$.

When working with a n-dimensional incomplete pattern, it can happen that more than one component (attribute value) of the pattern is missing. In our work we estimate these missing components by the corresponding components coming from the same prototype. More precisely, we do not consider hybrid cases[3]. For example, let us consider the following 3D incomplete pattern $\mathbf{y}_i = [?, y_{i2}, ?]$ to be classified in a 3-class problem using the three prototypes $\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3$. The edited pattern with three versions of estimated values are given by $\mathbf{y}_i^g = [o_{g1}, y_{i2}, o_{g3}], g = 1, 2, 3$, and the hybrid cases like $[o_{g1}, y_{i2}, o_{h3}], g \neq h$ are considered irrelevant for our analysis.

From each complete estimated vector $\mathbf{y}_i^g$, $g = 1, 2 \ldots, c$, we can draw a classification result using any standard classifier working with the complete pattern. At this step, the choice of the classifier, denoted $\Gamma(.)$, is left to user's preference. For instance, one can use for $\Gamma(.)$ the artificial neural network (ANN) [61] or evidential neural network (ENN) approach [38], or the K-NN [113], or the EK-NN [12], etc. The c pieces of sub-classification results for $\mathbf{y}_i$ are given for $g = 1, \ldots, c$ by

$$\mathbf{P}_i^g = \Gamma(\mathbf{y}_i^g | Y) \tag{5.3}$$

where $\Gamma(.)$ represents the chosen classifier, and $\mathbf{P}_i^g$ is the output (i.e. classification result) of the classifier when using the prototype of class ω_g to fill the incomplete pattern $\mathbf{x}_i$. $\mathbf{P}_i^g$ can be a Bayesian BBA if the chosen classifier works under probability framework (e.g. K-NN, ANN), and it can also be a regular bba with having some mass of belief committed to the ignorant class Ω if the classifier works under belief functions framework (e.g. EK-NN, ENN).

In this new PCC approach, we propose to combine these c pieces of classification results in order to get a credal classification of the incomplete pattern. These c pieces of classification results are considered as c distinct sources of evidences. Because the distances between the object and the c prototypes are usually different, some discounting technique must be applied to weight differently the impact of these sources of evidences in the global fusion process. If the distance of the object to prototype is big according to the known attribute values, it means that the estimation of missing values using this prototype is not very reliable. So the bigger distance d_{ij} generally leads to the smaller discounting factor α_j. A rational way that has been widely applied in many works is

work, we want to focus on the classification of the incomplete patterns as test samples. So the training samples are all assumed complete.

[3]Hybrid case means that if two (or more) components are missing in a pattern, there are replaced by the components coming from different prototypes.

adopted here to estimate at first the weighting factor w_i^g. For $g = 1, \ldots, c$, this factor w_i^g is defined by

$$w_i^g = e^{-d_{ig}} \tag{5.4}$$

where

$$d_{ig} = \sqrt{\frac{1}{p}\sum_{s=1}^{p}\left(\frac{y_{is} - o_{gs}}{\delta_{gs}}\right)^2} \tag{5.5}$$

with

$$\delta_{gs} = \sqrt{\frac{1}{T_g}\sum_{\mathbf{x}_i \in \omega_g}(x_{is} - o_{gs})^2} \tag{5.6}$$

y_{is} is value of $\mathbf{y}_i$ in s-th dimension, and x_{is} is value of $\mathbf{x}_i$ in s-th dimension. p is the number of dimensions of known values of $\mathbf{y}_i$. The coefficient $1/p$ is necessary to normalize the distance value because each test data can have a different number of dimensions of missing values. δ_{gs} is the average distance of all training data belonging to class ω_g to the prototype o_g in s-th dimension, and it is introduced mainly for dealing for the anisotropic data set. T_g is the number of training samples in the class ω_g.

From these weighting factors w_i^g for $g = 1, \ldots, c$, one then defines the relative reliability factors (discounting factor) α_i^g by

$$\alpha_i^g = \frac{w_i^g}{w_i^{\max}} \tag{5.7}$$

where $w_i^{\max} = \max(w_i^1, \ldots, w_i^c)$.

The discounting method proposed by Shafer in [3] is applied here to discount the BBA of each source of evidence according to the factors α_i^g. More precisely, the discounted masses of belief are obtained for $g = 1, \ldots, c$ by

$$\begin{cases} m_i^g(A) = \alpha_i^s P_i^g(A), & A \subset \Omega \\ m_i^g(\Omega) = 1 - \alpha_i^g + \alpha_i^g P_i^g(\Omega) \end{cases} \tag{5.8}$$

In Eq. (5.8), the focal element A usually represents a specific class in Ω because most classical classifiers work with probability framework, and thus they just consider specific classes as an admissible solution of the classification. Nevertheless, some classifiers based on DST, like EK-NN and ENN can generate results on specific classes and also on the full ignorant class Ω as well. $P_i^g(A)$ is the probability (or belief mass) committed to the class A by the chosen classifier.

5.2.2 Global fusion of the c discounted classification results

The c classification results obtained according to the c prototypes may strongly support different classes that the object should belong to. For instance, several sources of evidence could support that the object is most likely in class A, whereas some others could strongly support the class B, with $A \cap B = \emptyset$. In practice, the conflict usually exists in global fusion process. The maximum of belief function $Bel(.)$ given in Eq. (2.2) is used as criteria[4] for the decision making of the class which is strongly supported by the classification results, and the c pieces of results can be divided into several distinct groups $G_1, G_2, \ldots, G_r$ according to the classes they strongly support.

The classification results in the same group are combined at first, and then these sub-combination results are globally fused for the credal classification. The classification results in the same group

[4]The plausibility function $Pl(.)$ can also be used here, since $Bel(.)$ and $Pl(.)$ have a straight corresponding relationship in such particular bba's structure.

are generally not in high conflict. Therefore, one proposes to apply DS rule, defined in Eq. (2.4), to fuse these results, because DS rule offers a reasonable compromise between the specificity of the result and the level of complexity of the combination.

For $G_s = \{\mathbf{m}_i^j, \ldots, \mathbf{m}_i^k\}$, the fusion results of the BBA's in the group G_s using DS rule[5] are given for a focal element $A \in 2^\Omega$ by:

$$\mathbf{m}_i^{\omega_s}(A) = [\mathbf{m}_i^j \oplus \ldots \oplus \mathbf{m}_i^k](A) \tag{5.9}$$

where $\oplus$ represents the DS combination defined in Eq. (2.4). Since DS rule is associative, these BBA's can be combined sequentially using Eq. (2.4) and the sequential order doesn't matter.

These sub-combined BBA's $\mathbf{m}_i^{\omega_s}(.)$, for $s = 1, \ldots, r$, will then be globally fused to get the final BBA of credal classification. In the global fusion process, these sub-combination results of the different groups of sub-classification results can be in high conflict because of the distinct classes they strongly support. Because DS rule is known to produce counter-intuitive results specially in high conflicting situations [40, 44–46, 123, 124] due to its way of redistributing the conflicting beliefs, we propose to use another fusion rule to circumvent this problem. We recall that in DS rule the conflicting masses of belief are redistributed to all focal elements by the classical normalization step of Eq. (2.4). In our context, the partial conflicting information are very important to characterize the degree of uncertainty and imprecision of the classification caused by the missing values, and they should be preserved and transferred to the corresponding meta-classes specially in the high conflicting situation. Nevertheless, if all the partial conflicts are always unconditionally kept in the fusion results, they generate a high degree of imprecision of the result, which is not an efficient solution of the classification. To avoid this drawback, we make a compromise between the misclassification error rate and the imprecision degree we want to tolerate. This compromise is obtained by selecting the conflicting beliefs that need to be transferred to the corresponding meta-classes. The selection is done conditionally and according to the current context following the method explained in the sequel.

For simplicity and notation convenience, we assume that the resulting sub-combined BBA of group G_s is focused on the the class ω_s. That is $m_i^{\omega_s}(\omega_s) = \max(m_i^{\omega_s}(.))$[6], for $s = 1, \ldots, r$. This indicates that ω_s is strongly supported by the BBA's in group G_s. Moreover, the class $\omega_{\max}$ is the most believed class of the object if one has

$$m_i^{\omega_{\max}}(\omega_{\max}) = \max(m_i^{\omega_1}(\omega_1), \ldots, m_i^{\omega_g}(\omega_g)) \tag{5.10}$$

We remind that ω_{max} is the class having the biggest $m(.)$ value among all the classification groups, whereas $\omega_s, s = 1, \ldots, g$ just takes the biggest $m(.)$ value in the group G_s. In practice, it can happen that the belief $m_i^{\omega_s}(\omega_s)$ of the strongest class of the group G_s can be very close (or equal) to $m_i^{\omega_{\max}}(\omega_{\max})$ with $\omega_s \neq \omega_{\max}$. In such case, the object can also potentially belong to ω_s with a high likelihood. So we must consider all the very likely specific classes as potential solution of the classification of the object $\mathbf{y}_i$. The set of these potential classes is denoted Λ_i and it is defined by

$$\Lambda_i = \{\omega_s | m_i^{\omega_{\max}}(\omega_{\max}) - m_i^{\omega_s}(\omega_s) < \epsilon\} \tag{5.11}$$

where $\epsilon \in [0, 1]$ is a chosen threshold. Because all classes in Λ_i can very likely correspond to the real (unknown) class of $\mathbf{y}_i$, they appear not very distinguishable with respect to the threshold ϵ. This

[5] In the previous classifier BCKN, the simple averaging rule is used in the fusion of BBA's in the first step due to the particular structure of BBA's (with only single class and ignorant class). In PCC, there exists multiple single classes and the ignorant class in each BBA, and DS rule works quite well under these conditions.

[6] In fact, there are just single classes and total ignorant class involved in the classification result here. So the value of $m(.)$ for single class is equal to the value of $Bel(.)$ here.

means that a strategy of classification of the object $\mathbf{y}_i$ based only on one specific class of Λ_i is very risky, and all elements of Λ_i must be considered as acceptable in fact. To reduce misclassification errors with such type of strategy, we propose to keep all the subsets of Λ_i in the fusion process and we deal with the involved meta-classes.

If the beliefs of the other classes (e.g. ω_f) are all much smaller than $m_i^{\omega_{\max}}(\omega_{\max})$ as $m_i^{\omega_{\max}}(\omega_{\max}) - m_i^{\omega_f}(\omega_f) > \epsilon$, it means that the class $\omega_{\max}$ is generally distinct for the object with respect to the other classes (e.g. ω_f). Then, there is no necessity to keep the meta-class in such case.

The global fusion rule for these sub-combination results is defined by: $\forall B_i \subseteq \Omega$

$$\tilde{m}_i(A) = \begin{cases} \text{for } A \in \Omega \text{ with} |A| = 1, \text{or } A = \Omega \\ \quad \sum\limits_{\bigcap_{g=1}^{r} B_g = A} m_i^{\omega_1}(B_1) \cdots m_i^{\omega_r}(B_r), \\ \text{for } A \subseteq \Lambda_i, \text{with } |A| \geq 2 \\ \quad \sum\limits_{\substack{\bigcap_{i=1}^{|A|} B_i = \emptyset \\ \bigcup_{i=1}^{|A|} B_i = A}} [m_i^{\omega_1}(B_1) \cdots m_i^{\omega_s}(B_s) \prod\limits_{g=|A|+1}^{r} m_i^{\omega_g}(\Omega)] \end{cases} \tag{5.12}$$

In Eq. (5.12), r is the number of the groups of the classification results. $|A|$ is the cardinality of the hypothesis A, and it is equal to the number of singleton elements included in A. For example, if $A = \omega_i \cup \omega_j$, then $|A| = 2$.

In the first part of Eq. (5.12), the conjunctive combination is exactly the same as the unnormalized DS rule in Eq. (2.4), and it is used to calculate the mass of belief of the specific classes and of the ignorant class[7], since the degree of assignment of the object to a specific class or to the ignorant class depends on the consensus of sub-combination results represented by BBA's. In the second part of Eq. (5.12), $m_i^{\omega_1}(B_1) \cdots m_i^{\omega_s}(B_s)$ represents the partial conflicting beliefs produced in the fusion of the S sub-combined BBA's. This product characterizes in fact the joint belief that the object simultaneously belongs to these specific exhaustive and incompatible classes $B_i, i = 1, \ldots, s$. $m_i^{\omega_g}(\Omega)$ denotes the ignorance, and it plays a neutral role in the fusion process. Therefore, the product of them (i.e. the whole second part of Eq. (5.12)) reflects the imprecision (uncertainty) degree of classification of the object with these different specific classes in the global fusion of all the sub-combined BBA's. For this reason, one reasonably commits it to the meta-class composed by the union (disjunction) of these classes as $A = \bigcup_{i=1}^{s} B_i$. If none meta-class is selected, the second part of formula can be ignored, and the Eq. (5.12) will reduce to DS rule after the normalization step given in Eq. (5.13).

Because not all partial conflicting masses of belief[8] are transferred into the meta-classes through the global fusion formula (5.12), the combined bba is normalized as follows before making a decision:

$$m_i(A) = \frac{\tilde{m}_i(A)}{\sum\limits_{B_j} \tilde{m}_i(B_j)} \tag{5.13}$$

[7] The ignorant class represents the outlier (noisy) class.

[8] In Eq. (5.12), the partial conflicts with redundant elements (e.g. with mass $m_1(\omega_1)m_2(\omega_1)m_3(\omega_2)$) are neither committed to the most redundant specific class (e.g. ω_1), nor assigned to a meta-class (e.g. $\omega_1 \cup \omega_2$) because they contain some partial consensus (e.g. $m_1(\omega_1)m_2(\omega_1)$ for ω_1). So it seems more appropriate to distribute such type of partial conflicting beliefs to all the focal elements through the final classical normalization step according to Eq. (5.13).

The credal classification of the object can be made directly based on this final normalized combined result BBA's, and the object will be assigned to the focal element (a specific class or a meta-class) with maximal mass of belief. The maximum of belief $Bel_i(.)$ of the singleton (specific) class, or the maximum of plausibility $Pl_i(.)$, or the maximum of pignistic probability $BetP_i(.)$ drawn from the global combined BBA $m_i(.)$ are usually used as the criteria for making hard (specific) classification, but the hard classification is not recommended in such uncertain case. The credal classification based on the BBA's is preferred here since it can well reflect the inherent imprecision (ambiguity) degree of the classification due to the missing values.

The flowchart of PCC is presented in Fig. 5.2 to explicitly show how PCC works and the pseudo-code of the PCC is given in Table 5.1 for convenience.

Table 5.1 : Prototype-based Credal classification method.

Input:
Training samples: $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_H\}$ in $\mathbb{R}^p$
Incomplete test samples: $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ in $\mathbb{R}^p$
Parameters:
$\epsilon \in [0, 1]$: threshold of meta-class
for i=1 to N
Calculate prototypes using Eq. (5.1);
Classify c versions of edited $\mathbf{x}_i$ with estimated values;
Determine the weighting factors by Eq.(5.7);
Discount the c classification results using Eq. (5.8);
Subcombination of consistent classification results by Eq. (5.9);
Select meta-classes according to Eq. (5.11);
Global fusion of the subcombination results by Eqs. (5.12), (5.13);
end

Guideline for choosing the meta-class threshold ϵ: In the applications, the threshold ϵ of PCC must be tuned according to the number of objects in the meta-classes. A small ϵ value generally leads to fewer objects in the meta-classes, but it may cause more misclassifications for the uncertain objects. A big ϵ value yields more objects in the meta-classes and leads to high imprecision degree, which is not an efficient solution for the classification. So ϵ should be tuned according to the imprecision degree of the fusion results that one accepts.

The following simple example shows how PCC works.

Example 5.1: Let us consider a 3-D object $\mathbf{y}_i = [y_{i1}, ?, ?]$ with the missing values in the 2nd dimension and 3rd dimension to be classified over the frame of classes $\Omega = \{\omega_1, \omega_2, \omega_3\}$. It is assumed that the prototypes $O = \{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}$ of the three classes can be calculated using the training data as:

$$\begin{aligned}\mathbf{o}_1 &= [o_{11}, o_{12}, o_{13}]\\ \mathbf{o}_2 &= [o_{21}, o_{22}, o_{23}]\\ \mathbf{o}_3 &= [o_{31}, o_{32}, o_{33}]\end{aligned}$$

So the object with three versions of estimation of the missing value is obtained by:

$$\begin{aligned}\mathbf{y}_i^1 &= [y_{i1}, o_{12}, o_{13}]\\ \mathbf{y}_i^2 &= [y_{i1}, o_{22}, o_{23}]\\ \mathbf{y}_i^3 &= [y_{i1}, o_{32}, o_{33}]\end{aligned}$$

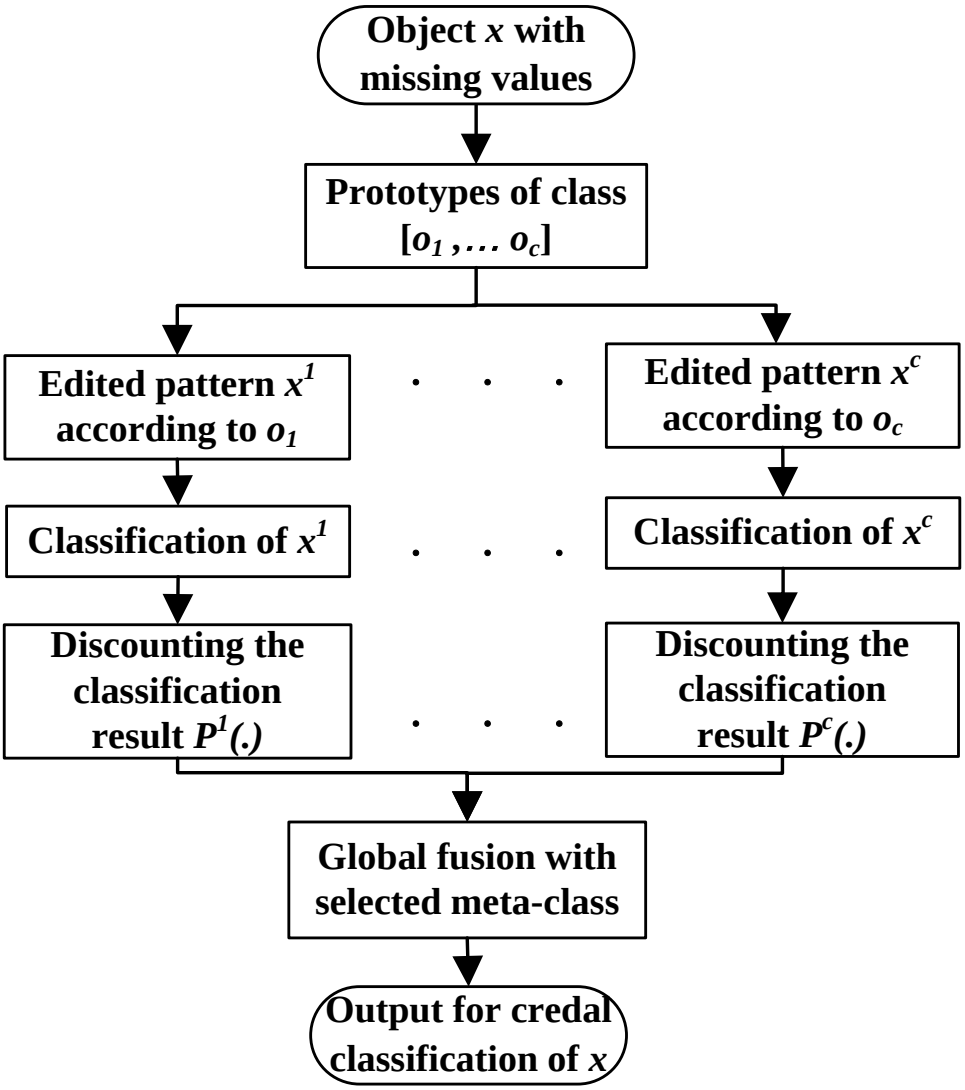


Figure 5.2 : Flowchart of the proposed PCC method.

The patterns with three estimated values are respectively classified using a standard classifier, and

the classification results represented by the probability membership are given by[9]:

$$P_i^1(\omega_1) = 0.8, \quad P_i^1(\omega_2) = 0.2.$$
$$P_i^2(\omega_1) = 0.1, \quad P_i^2(\omega_2) = 0.8, \quad P_i^2(\omega_3) = 0.1.$$
$$P_i^3(\omega_1) = 0.5, \quad P_i^3(\omega_2) = 0.2, \quad P_i^3(\omega_3) = 0.3.$$

The relative weighting factor of each classification result is calculated according to the distance between $\mathbf{y}_i$ and the three prototypes using Eq. (5.7). For simplicity and convenience, they have been randomly chosen as follows for this example:

$$\alpha_i^1 = 1, \quad \alpha_i^2 = 0.9, \quad \alpha_i^3 = 0.3$$

Then, each classification result $P_i^k(.)$, $k = 1, \ldots, 3$ can be discounted using Eq. (5.8), and the discounted BBA's are given by

$$\begin{aligned}
\mathbf{m}_i^1(.): &\ m_i^1(\omega_1) = 0.8, \quad m_i^1(\omega_2) = 0.2.\\
\mathbf{m}_i^2(.): &\ m_i^2(\omega_1) = 0.09, \quad m_i^2(\omega_2) = 0.72,\\
&\ m_i^2(\omega_2) = 0.09, \quad m_i^2(\Omega) = 0.1.\\
\mathbf{m}_i^3(.): &\ m_i^3(\omega_1) = 0.15, \quad m_i^3(\omega_2) = 0.06,\\
&\ m_i^3(\omega_3) = 0.09, \quad m_i^3(\Omega) = 0.7.
\end{aligned}$$

Because of the particular choice of $\alpha_i^1 = 1$ the BBA $m_i^1(.)$ is not discounted in this example.

The belief functions $Bel_i(.)$ corresponding to each BBA $m_i(.)$ are obtained using Eq. (2.2) and are given by

$$Bel_i^1(\omega_1) = 0.8, \quad Bel_i^1(\omega_2) = 0.2.$$
$$Bel_i^2(\omega_1) = 0.09, \quad Bel_i^2(\omega_2) = 0.72, \quad Bel_i^2(\omega_3) = 0.09.$$
$$Bel_i^3(\omega_1) = 0.15, \quad Bel_i^3(\omega_2) = 0.06, \quad Bel_i^3(\omega_3) = 0.09.$$

For the singleton (specific) class, $m_i^1(.)$ and $m_i^3(.)$ put the most belief on class ω_1, whereas $m_i^2(.)$ commits most of mass to the class ω_2. It means that the object likely belongs to class ω_1 with the estimation from prototype $\mathbf{o}_1$ and $\mathbf{o}_3$, but it is very probably classified into ω_2 with the estimation according to $\mathbf{o}_2$. This uncertainty (conflict) is mainly caused by the lack of discriminant information inherent of the missing values. Then, the three BBA's can be divided into the two following groups: $G_1 = \{m_i^1(.), m_i^3(.)\}$ and $G_2 = \{m_i^2(.)\}$.

The sub-combination results of each group of BBA's using DS rule (2.4) are:

$$\begin{aligned}
\mathbf{m}_i^{\omega_1}(.): &\ m_i^{\omega_1}(w_1) = 0.8173, \quad m_i^{\omega_1}(w_2) = 0.1827.\\
\mathbf{m}_i^{\omega_2}(.): &\ m_i^{\omega_2}(w_1) = 0.09, \quad m_i^{\omega_2}(w_2) = 0.72,\\
&\ m_i^{\omega_2}(w_3) = 0.09, \quad m_i^{\omega_2}(\Omega) = 0.1.
\end{aligned}$$

Then one gets: $Bel_i^{\omega_{\max}}(\omega_{\max}) = Bel_i^{\omega_1}(\omega_1) = 0.8173$ and $Bel_i^{\omega_2}(\omega_2) = 0.72$. If the meta-class threshold is chosen as $\epsilon = 0.3$, we get $Bel_i^{\omega_1}(\omega_1) - Bel_i^{\omega_2}(\omega_2) < \epsilon$, and thus $\Lambda_i = \{\omega_1, \omega_2\}$. So the meta-class $\omega_1 \cup \omega_2$ will be kept, and the conflicting mass of belief produced by the conjunctive combination $m_i^{\omega_1}(w_1)m_i^{\omega_2}(w_2) + m_i^{\omega_1}(w_2)m_i^{\omega_2}(w_1)$ will be transferred to $\omega_1 \cup \omega_2$.

The global fusion of BBA's $\mathbf{m}_i^{\omega_1}(.)$ and $\mathbf{m}_i^{\omega_2}(.)$ using Eq. (5.12) yields the following unormalized combined BBA

$$\begin{aligned}
\tilde{\mathbf{m}}_i(.): &\ \tilde{m}_i(\omega_1) = 0.1553, \quad \tilde{m}_i(\omega_2) = 0.1498,\\
&\ \tilde{m}_i(\omega_1 \cup \omega_2) = 0.6049.
\end{aligned}$$

[9] In this example, we just want to show the main steps of the PCC method. The classification results $P_i^k(.)$ and the relative factors α_i^k have been arbitrarily chosen here for convenience.

As we see, the BBA $\tilde{\mathbf{m}}_i(.)$ is not a normalized BBA because some conflicting masses of belief are voluntarily discarded of the redistribution on the meta-classes. After the normalization step, we finally get:

$$\begin{aligned}\mathbf{m}_i(.): \ & m_i(\omega_1) = 0.1707, \quad m_i(\omega_2) = 0.1646, \\ & m_i(\omega_1 \cup \omega_2) = \mathbf{0.6647}.\end{aligned}$$

One sees that the biggest mass of belief is committed to the meta-class $\omega_1 \cup \omega_2$. This result indicates that the classes ω_1 and ω_2 are not very distinguishable based only on the known attribute information, and the object must quite likely belong to ω_1 or ω_2 according to the different estimations of the missing values. In this simple example, it is difficult to commit the object to a particular class. If one had to take a specific class decision, one would very probably make a mistake. So the hard classification is not recommended in such case, and the object will be committed to the meta-class $\omega_1 \cup \omega_2$ by PCC approach, which is prudent and reasonable behavior consistent with the intuitive reasoning. Some additional sources (if available) need to be used and combined with the available information to get a more precise classification result.

5.3 EXPERIMENTS

Four experiments have been carried out to test and evaluate the performance of this new PCC method. The performances of PCC are compared to that of the mean imputation (MI) method, K-NN imputation (KNNI) method and FCM imputation (FCMI) method. In MI, the missing values are replaced using the mean value of the same attribute of the training samples. In KNNI, the missing values are estimated using its K-nearest neighbors in training data space. In FCMI, the missing values are imputed according to the clustering centers of FCM and the distances between the object and these centers [121, 122]. In this work, the EK-NN classifier [12] and evidential neural network classifier (ENN)[10] [38] are adopted as the standard classifier to classify the test samples with the estimated values in PCC, MI, KNNI and FCMI. In fact, many other standard classifiers can be applied here according to the actual request, and the selection of standard classifier is not the main purpose of this work. The parameters of EK-NN [118] and ENN [38] can be automatically optimized. In the applications of PCC, the tuning parameter ϵ can be automatically tuned according to the imprecision rate one can accept, and can also be optimized using cross validation in training data space where the attribute value is randomly missed in every dimension. In order to show the ability of PCC to deal with the meta-classes, the class of each object is decided according to the criterion of the maximal mass of belief.

In our simulations, the error rate denoted by Re and the imprecision rate denoted by Ri_j are still used to evaluate the performance of CCR. In the sequel experiments, the classification of object is generally uncertain (imprecise) among a very small number (e.g. 2) of classes, and we only take Ri_2 here since there is no object committed to the meta-class including three or more specific classes.

It is worth noting that in Fig. 5.3 and 5.4, the x-axis and y-axis respectively correspond the first and the second dimension of test and training data used in the experiments.

[10] In this work, it is considered that the uncertainty of classification is mainly caused by the missing values, and the meta-class is not necessarily taken into account in the classification of the complete data for simplicity of computation. So the proposed credal classifier in previous chapters is not used in this step.

5.3.1 Experiment 5.1 (with 2D 3-class data set)

This experiment shows the results of the credal classification obtained by PCC with respect to other classical methods, and the EK-NN is applied here to classify the test samples with the estimated values. We consider a particular 3-class data set $\Omega = \{\omega_1, \omega_2, \omega_3\}$ in the circular shape as shown in Fig. 5.3-a. Each class contains 305 training samples and 305 test samples. Thus, we consider $3 \times 305 = 915$ training samples and $3 \times 305 = 915$ test samples. The radius of the circle is $r = 3$, and the centers of three circles are given by the points $\mathbf{c_1} = (3,3)^T$, $\mathbf{c_2} = (13,3)^T$, $\mathbf{c_3} = (8,8)^T$, where T denotes the transposed vector. The values in the second dimension corresponding to y-coordinate of test samples are all missing, and the there is only one known value in the first dimension corresponding x-coordinate for each test sample. The different meta-class selection thresholds $\epsilon = 0.3$ and $\epsilon = 0.45$ have been applied in PCC to show their influences on the results. A particular value of $K = 9$ is selected in the classifier EK-NN and the K-NN imputation[11]. The classification results of the test objects by different methods are given by Fig. 5.3-b–5.3-f. For notation conciseness, we have denoted $w^{te} \triangleq w^{test}$, $w^{tr} \triangleq w^{training}$ and $w_{i,\ldots,k} \triangleq w_i \cup \ldots \cup w_k$. The error rate (in %) and imprecision rate (in %) for PCC have been given in the caption of each subfigure.

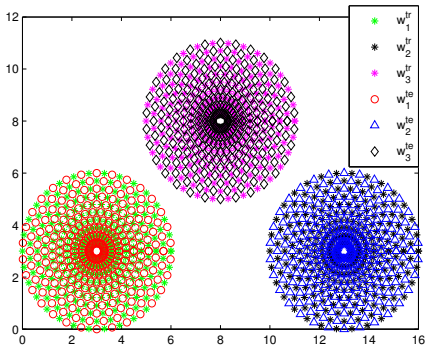
5

The values of the y-coordinate of the test samples are all missing, and the class of each test sample is determined only based on the value of x-coordinate. We can see from Fig. 5.3-(a) that the class ω_3 partly overlaps with the classes ω_1 and ω_2 on their margins with respect to x-coordinate. The objects lying in the overlapped zone are really difficult to be correctly classified into a particular class, since ω_1 and ω_3 (resp. ω_2 and ω_3) seem undistinguishable for these objects based on the values on x-axis. The mean, K-NN and FCM estimation methods provide only one value for the missing data, and then the EK-NN classifier is used to classify the test samples with this estimated value. The objects are all committed to a particular class by these methods with big error rate, and the results cannot well reflect the uncertainty and imprecision of classification caused by the missing values. With the PCC approach, most objects lying in the overlapped zones are reasonably assigned to the proper meta-classes $\omega_1 \cup \omega_3$ and $\omega_2 \cup \omega_3$. So PCC is able to reduce the error rate and well characterize the imprecision (ambiguity) of the classification thanks to the use of meta-class under belief functions framework. One can see that the increases of ϵ value leads to the decrease of error rate but meanwhile it brings the increase of imprecision rate. So we should find a good compromise between the error rate and imprecision rate. In real applications, ϵ can be optimized using the training data, and the optimized value should correspond to a suitable compromise between the error rate and imprecision rate. ϵ can also be tuned according to the imprecision rate one can accept in the classification.

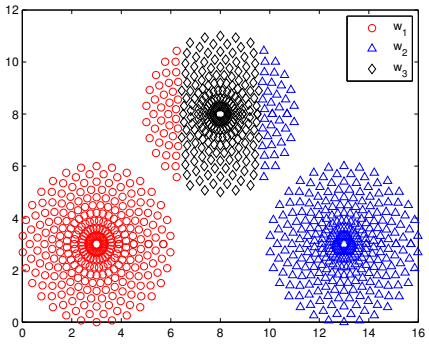
5.3.2 Experiment 5.2 (with 4-class data set)

A 4-class data set $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ obtained from four 2-D uniform distributions is used to test the performance of PCC with respect to other methods. Each class has 100 training samples and 100 test samples. The uniform distributions of the samples of the four classes are characterized by the following interval bounds:

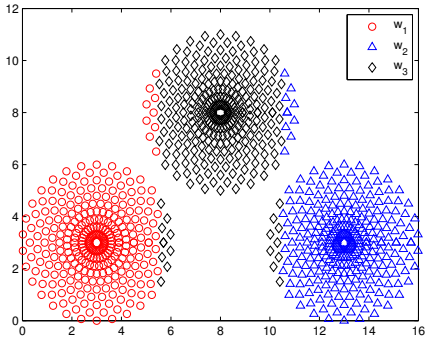
[11] In fact, the choice of K ranking from 7 to 15 does not affect seriously the results.



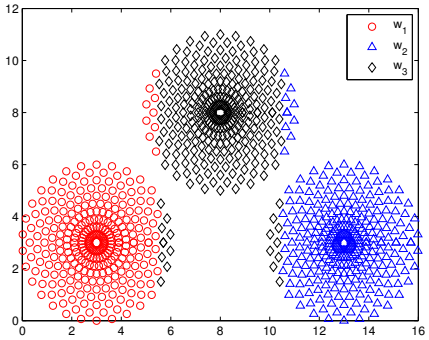
(a). Training data and test data.



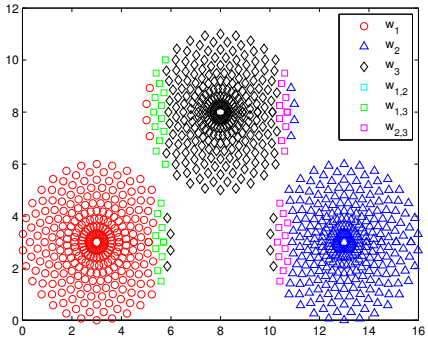
(b). Classification result by method with mean estimation ($Re = 8.52$).



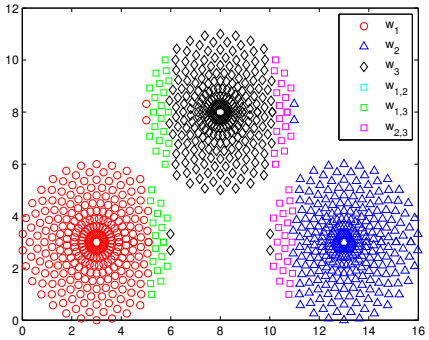
(c). Classification result by method with K-NN estimation ($Re = 4.15$).



(d). Classification result by method with FCM estimation ($Re = 4.15$).



(e). Classification result by PCC $\epsilon = 0.3$ ($Re = 1.75, Ri_2 = 4.81$).



(f). Classification result by PCC $\epsilon = 0.45$ ($Re = 0.87, Ri_2 = 8.31$).

Figure 5.3 : Classification results of a 3-class data set by different methods.

	x-label interval	y-label interval
w_1	(10, 20)	(5, 65)
w_2	(10, 20)	(110, 170)
w_3	(35, 45)	(50, 120)
w_4	(55, 65)	(150, 230)

We consider that the values in the first dimension corresponding to x-coordinate of test samples are all missing, and each test sample has only one value in the second dimension corresponding y-coordinate. Both EK-NN and ENN classifiers are employed here to classify the test samples with estimated values. The averages error rates and imprecision rates are given in Table 5.2 over 10 trials performed with 10 independent random generation of the data sets. For the K-NN based methods, the mean value of error and imprecision rates for $K \in [5, 20]$ are calculated. The parameter ϵ in PCC has been optimized to find the proper compromise between error and imprecision. The classification results (with EK-NN classifier) of a particular data set picked out from the random generations are displayed in Fig. 5.4 to clearly illustrate the use of meta-class in credal classification with respect to other classical methods.

Table 5.2 : Statistics of classification for 4-class data set by different methods (in %).

	MI	KNNI	FCMI	PCC
	Re	Re	Re	$\{Re, Ri_2\}$
EK-NN	28.37	13.04	16.53	{8.62, 13.87}
ENN	18.90	13.63	16.01	{8.75, 11.75}

In Table 5.2, we can see that error rates of PCC method with EK-NN and ENN are smaller than the other applied methods. Meanwhile, some incomplete patterns that are very difficult to classify into a specific class have been committed to the proper meta-class. For example, the objects labeled by blue square in Fig. 5.4-(e) can not be clearly distinguished by class w_1 and w_3 according to the only one known value in y-coordinate. So they are automatically committed to the meta-class $w_1 \cup w_3$ to reduce the risk of mistake. It is similar for some other objects that are committed to $w_2 \cup w_3$ and $w_2 \cup w_4$ shown in Fig. 5.4-(e). Nevertheless, these uncertain objects are all specifically classified into the particular class by other methods, which produce a lot of errors. Both EK-NN and ENN were applied in this experiment to classify the samples with the estimated values. The difference of their performances is quite small in this experiment, but the high computation burden is usually the main drawback of the K-NN method. So ENN can be a good choice especially for dealing with the large scaled data set.

5.3.3 Experiment 5.3 (with 4D 3-class data sets)

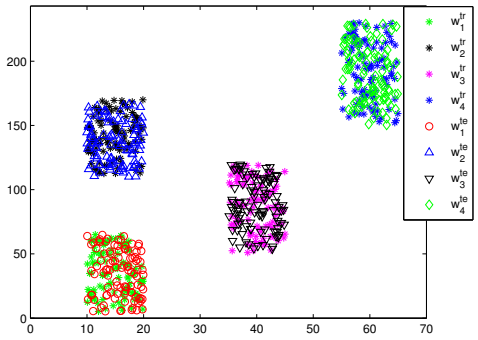
In this experiment, the statistics of the performances of PCC are compared with KNNI, MI and FCMI on a 3-class data set which is generated from three 4D Gaussian distributions characterizing the classes $\omega_1, \omega_2, \omega_3$. The 4D Gaussian distributions of probabilities used in our simulations have the following means vectors and covariance matrices ($\mathbf{I}$ is the 4×4 identity matrix):

$$\mu_1 = (1, 5, 10, 10)^T, \mathbf{\Sigma}_1 = 6 \cdot \mathbf{I}$$

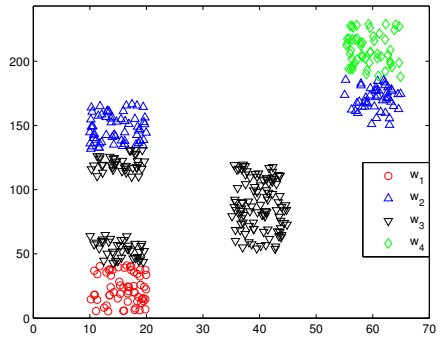
$$\mu_2 = (10, 3, 2, 1)^T, \mathbf{\Sigma}_2 = 5 \cdot \mathbf{I}$$

$$\mu_3 = (15, 15, 1, 15)^T, \mathbf{\Sigma}_3 = 7 \cdot \mathbf{I}$$

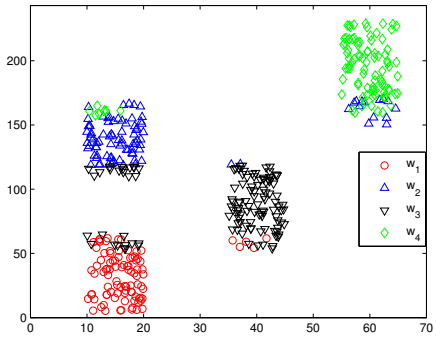
We have used N training samples, and N test samples (for $N = 100, 200$) in each class. Each test sample has n missing values (for $n = 1, 2, 3$), and the missing component value is chosen randomly in every dimension. The values of K ranging from 5 to 20 neighbors in EK-NN and KNNI have



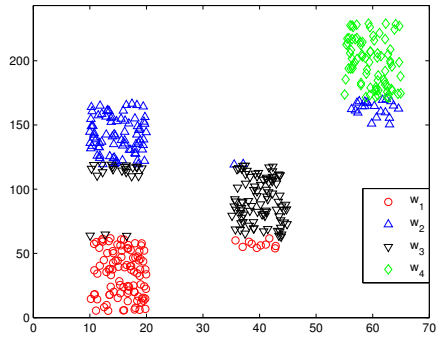
(a). Training data and test data.



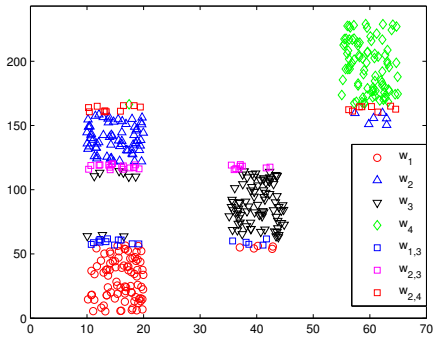
(b). Classification result by method with mean estimation ($Re = 31.25$).



(c). Classification result by method with K-NN estimation ($Re = 12$).



(d). Classification result by method with FCM estimation ($Re = 14$).



(e). Classification result by PCC ($Re = 5.25, Ri_2 = 15.25$).

Figure 5.4 : Classification results of a 4-class data set by different methods.

been tested. For each pair (N, n), the reported error rates and imprecision rates are the averages over 10 trials performed with 10 independent random generation of the data sets. The mean of the classification error and imprecision rates for $K \in [5, 20]$ are calculated and given in Table 5.3. In PCC method, the parameter ϵ has been optimized to obtain an acceptable compromise between error rate and the imprecision degree. EK-NN and ENN are respectively applied for the classification of the test samples with estimated values. In Table 5.3 and Table 5.5, we have denoted SC $\triangleq$ standard classifier, A $\triangleq$ EK-NN and B $\triangleq$ ENN for notation conciseness.

Table 5.3 : Statistics of classification for 3-class data set by different methods (in %).

(N, n)	SC	MI	KNNI	FCMI	PCC
		Re	Re	Re	$\{Re, Ri_2\}$
(100, 1)	A	19.24	18.28	18.17	{11.67, 12.60}
(100, 1)	B	19.33	17.99	17.50	{14.67, 5.33}
(100, 2)	A	25.94	24.57	24.27	{16.72, 12.01}
(100, 2)	B	25.20	24.32	24.13	{16.85, 11.00}
(100, 3)	A	41.85	41.00	40.06	{29.00, 14.39}
(100, 3)	B	40.57	39.91	38.43	{27.70, 15.07}
(200, 1)	A	19.62	18.83	18.81	{12.65, 11.34}
(200, 1)	B	17.83	17.60	17.50	{15.17, 4.50}
(200, 2)	A	28.06	25.00	24.90	{15.73, 14.86}
(200, 2)	B	23.85	23.52	23.22	{18.27, 8.83}
(200, 3)	A	41.34	40.86	39.59	{29.82, 14.86}
(200, 3)	B	39.90	38.99	37.85	{29.67, 15.13}

One can see from Table 5.3 that PCC produces the smallest error rate, since the objects that are very difficult to classify have been assigned to the proper meta-class. So there are some imprecision in the credal classification of PCC, and this represents well the uncertainty caused by the missing values. When the number of missing values is big (i.e. $n = 3$), the uncertainty degree of classification becomes high, and then the PCC method provides much better performances than the other methods. This experiment illustrates the interest of credal classification of incomplete data based on evidential reasoning. EK-NN and ENN are used as standard classifiers for dealing with the imputed patterns, and it seems that ENN generally has a bit better performance (with lower error rate and computation burden) than EK-NN here.

5.3.4 Experiment 5.4 (with real data sets)

In this experiment, we use the four real data sets (Breast cancer, Seeds, Yeast and Wine data sets) available from UCI Machine Learning Repository [95] to test the performance of PCC with respect to MI, KNNI and FCMI. Both EK-NN and ENN are still selected here as standard classifiers. Three classes (CYT, NUC and $ME3$) are selected in Yeast data set to the evaluate our method, since these three classes are close and difficult to classify. The basic information of the four data sets is given in Table 5.4.

The simple *2*-fold cross validation was performed on the four data sets by the different classification methods here. Each test sample has n missing (unknown) values, and they are missing completely at random in every dimension. The average error rate Re_a and imprecision rate Ri_a (for PCC) of the different classical methods with values of K ranging from 5 to 20 are given in Table 5.5.

Table 5.4 : Basic information of the used data sets.

name	classes	attributes	instances
Breast	2	9	699
Seeds	3	7	210
Wine	3	13	178
Yeast	3	8	1050

Table 5.5 : Classification results for different real data sets (in %).

data set	(n, SC)	MI	KNNI	FCMI	PCC
		Re	Re	Re	$\{Re, Ri_2\}$
Breast	(3, A)	4.71	6.10	3.95	{4.10, 3.38}
	(3, B)	4.25	3.95	3.81	{3.81, 2.34}
	(5, A)	8.20	8.15	5.07	{4.38, 4.69}
	(5, B)	6.44	5.76	5.27	{3.81, 6.00}
	(7, A)	38.33	14.35	13.00	{7.91, 8.05}
	(7, B)	14.64	11.54	11.42	{6.88, 12.44}
Yeast	(1, A)	37.59	38.13	38.54	{34.36, 6.95}
	(1, B)	37.71	36.70	36.19	{32.67, 6.19}
	(3, A)	45.08	44.29	45.95	{34.71, 18.00}
	(3, B)	42.10	40.90	41.33	{34.19, 14.95}
	(5, A)	51.16	50.95	51.11	{33.46, 31.01}
	(5, B)	49.33	49.22	46.00	{32.29, 27.62}
Seeds	(3, A)	21.03	9.68	12.46	{7.14, 3.72}
	(3, B)	21.43	11.19	13.33	{9.05, 2.86}
	(5, A)	33.49	12.54	20.08	{9.67, 6.70}
	(5, B)	31.43	12.14	20.00	{9.52, 9.05}
	(6, A)	40.71	25.87	21.75	{16.79, 12.77}
	(6, B)	39.52	25.71	20.95	{16.19, 14.76}
Wine	(3, A)	30.71	26.59	30.15	{26.05, 1.05}
	(3, B)	29.78	26.97	26.97	{26.97, 1.69}
	(6, A)	34.93	25.84	32.12	{26.62, 0.84}
	(6, B)	33.71	28.09	32.02	{27.53, 1.12}
	(10, A)	39.23	30.90	32.30	{25.84, 3.86}
	(10, B)	37.64	31.18	31.46	{27.53, 3.93}

The results of Table 5.5 clearly show that the PCC method generally produces lower error rate than the MI, KNNI and FCMI classification methods, but meanwhile it yields some imprecision in the classification result due to the introduction of meta-classes, which indicates that some incomplete objects are very difficult to classify because of lack of discriminant information. The increase of the number (i.e. n) of missing values in each test sample generally causes the increment of error rate in the classifiers. The imprecision rate becomes bigger in PCC, since the more missing values lead to the bigger imprecision (uncertainty) in the classification problem. So the credal classification including meta-class is very useful and efficient here to represent the imprecision degree and it can help also to decrease the misclassification rate. The PCC approach allows to indicate that the objects in meta-classes are really hard to be correctly classified, and they should be cautiously treated in the applications. If one wants to get more precise results, some other (possibly costly) techniques seem necessary to discriminate and classify such uncertain objects. EK-NN and ENN are respectively applied to classify the test samples with estimated values in MI, KNNI, FCMI and PCC methods. The computation burden of K-NN based methods like EK-NN is usually big, which is not very convenient for the real applications requiring high running speed. ENN provides a bit lower error rate than EK-NN in many cases, but the training procedure of ENN is a bit complicate. The proper standard classifier for dealing with complete data can be selected according to the actual application.

5.4 Conclusion

A new prototype-based credal classification (PCC) method has been presented in this chapter for classifying incomplete patterns thanks to the belief function framework. This PCC method allows the object (incomplete pattern) to belong not only to specific classes, but also to meta-classes (i.e. union of several specific classes) with different masses of belief. The meta-class is introduced to characterize the imprecision of classification due to the missing values, and it can also reduce errors. In a c-class problem, the c class prototypes obtained from training data are respectively used to estimate the missing values of the incomplete pattern. The object with each of the c estimations can be classified by any standard classifier, and it will produce c pieces of classification results with different weighting factors determined by the distances between the object and the prototypes. These results are respectively discounted according to their relative weights. The global fusion of these discounted results is adopted for credal classification of the object. If the c results are consistent on the classification, the object will be committed to a particular class that is strongly supported by the c results. However, the high conflict among these c results means that the class of the object is quite uncertain and imprecise only based on the known attributes information. In such case, the object becomes very difficult to classify correctly in a specific class and it is reasonably assigned to the proper meta-class defined by the union of the specific classes that the object is likely to belong to. Then the conflicting mass of belief is transferred conditionally to the selected meta-class. Once an object is committed to a meta-class, it means that the specific classes included in the meta-class seem undistinguishable for this object based on the known attributes. If one wants to get more precise result, some other (possibly costly) techniques or information sources must be developed and used. Four experiments with artificial and real data sets have been done to evaluate the performances of PCC with respect to other classical methods. The results show that PCC is able to reduce error rate, and well capture and represent the imprecision of classification caused by missing data.

6 Credal c-means clustering method

6.1 Introduction

Fuzzy c-means (FCM) [39] remains so far the most popular data clustering method, and it works with fuzzy partition under the probabilistic framework. In the clustering of imprecise data, some data points (objects) can be simultaneously close to several clusters, and these clusters may seem undistinguishable for the objects. Such imprecise data are really difficult to classify correctly into a particular cluster. The imprecision of information cannot be well captured by the probabilistic framework [2], whereas belief functions theory [3–6] also called evidence theory is able to well model the imprecision and uncertainty.

Credal partition [25] has been recently proposed for data clustering using belief functions, and it allows that the objects belong to not only the singleton clusters in $\Omega = \{w_1, \dots w_c\}$ but also any subsets of Ω (i.e. meta-clusters) with different masses of belief. An evidential C-Means (ECM) [23] clustering method as an extension of Fuzzy C-means under belief functions framework is proposed for the credal partition of object data. The relational version of ECM (RECM) for dealing with relational data [27] and constrained ECM (CECM) [28] taking into account the pairwise constraints information have been developed. In the ECM [23] method, each class is characterized by its class center, and the center of meta-class is the simple mean value of the involved specific classes' centers. The mass of belief committed to each class is proportional to the distance between the object and the corresponding class center. The bigger distance leads to the smaller mass of belief. However, when the different cluster centers are close, ECM will produce very counterintuitive results that some objects belonging to a singleton cluster can be wrongly committed into an incompatible meta-cluster whose center is close to the singleton cluster's center. The important contribution of ECM mainly lies in the introducing of meta-clusters to FCM, but the use of meta-cluster in ECM is still questionable. The limitations of ECM has been pointed out in Chapter 2.

We propose a new evidential version of FCM called credal c-means (CCM) to overcome the limitation of ECM, and a justified use of meta-clusters is presented. CCM also works with credal partition for the clustering of imprecise data based on belief functions. In CCM, the meta-cluster is considered as a kind of transition cluster among the different close singleton clusters. Thanks to meta-cluster, the credal partition provides an effective tool for the clustering of the imprecise data that are difficult to be correctly committed to a particular cluster, and it can reduce the error occurrences. If one object is considered in a meta-cluster, it must be simultaneously close to the singleton clusters included in the meta-cluster, which means this object is not likely to belong to the other incompatible clusters, and this mainly depends on the distances between the object and these singleton cluster centers. Meanwhile, these singleton clusters should be undistinguishable for the object, which indicates the objects are hard to be correctly committed into a particular singleton cluster, and this mainly depends on the distance between this object and the meta-cluster's center (i.e. the mean value of the involved singleton cluster centers). Thus, in the determination of the belief on the meta-cluster, we should take into account not only the distance to the meta-cluster's

center but also the distances to the involved singleton cluster centers.

One object in a meta-cluster means that it belongs to one of the singleton clusters included in the meta-cluster, but the available information used for making the classification is not sufficient enough to obtain a clear (specific) class of the object. CCM can well reveal the imprecision degree of the object belonging to different classes and can also reduce the misclassification errors due to the meta-cluster. CCM is also robust to noisy data (i.e. outlier) using the outlier cluster, and it is mainly determined according to a given outlier threshold. The singleton cluster in CCM corresponds to the objects very close to the center of this cluster, which is similar to FCM and ECM. The output of CCM is not always used to provide a final decision about classification of an object. In fact, it can be seen as an interesting source of information to be combined with some other complementary information sources in order to get more precise clustering results if necessary.

The objective function of CCM is defined following this basic principle. The clustering centers and the belief of each cluster for the objects can be obtained by the optimization (minimization) of this objective function. For a c-class data set, the credal partition produces 2^c clusters, and its computation complexity is very high when c is big. In real applications, the classification of imprecise objects are usually unspecific among several (a very small number, e.g. two or three) singleton clusters, and there are very few objects belonging to the meta-clusters with big cardinality. So a threshold t_c is introduced in CCM to eliminate the meta-clusters with big cardinality, which can effectively reduce the computation burden. If necessary, the credal partition can be simply reduced to fuzzy partition.

It is worth noting that the BBA's determination way in the unsupervised CCM method is quite different from the way used in supervised CCR introduced in Chapter 3. In CCM, there is no training samples available for classification, and the clustering analysis is mainly based on the objective function, which will be optimized to obtain the clustering centers and the BBA's of each object associated with different clusters. If meta-cluster center is considered with exact the same distance to all the involved specific classes' centers and distinguishability degree is used for the determination of BBA's as done in CCR, the objective function will become unlinear, and it could be too complicate to be simply optimized. We want to make the objective function linear, which is convenient for the optimization procedure. So the meta-cluster center is defined by the simple mean value of the involved specific clusters' centers in CCM, and the mass of belief on the meta-cluster directly depends on the distance of the object to the meta-cluster center and the distances to all the involved clusters' centers.

The details of the new CCM approach is presented in section 6.2. Some experiments are given in section 6.3 to illustrate the effectiveness of CCM with respect to FCM and ECM approaches, before concluding this chapter in section 6.4.

6.2 Credal c-Means (CCM) approach

6.2.1 The objective function of CCM

In order to circumvent the limitation of ECM, a new alternative evidential version of fuzzy c-Means, called credal c-means (CCM), is proposed for modeling and clustering the uncertain and imprecise data. The basic principle of CCM is as follows.

Let us consider a set of $n > 1$ objects. Each object also called a data point $\#i$ is represented by a given attribute vector $\mathbf{x}_i$ of dimension $p \geq 1$. These objects will be clustered over a given frame of discernment $\Omega = \{w_1, w_2, \ldots, w_c\}$ with the corresponding centers $\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_c}\}$, and the credal partition is generalized based on the power-set 2^Ω.

In clustering of a data set, if one object $\mathbf{x}_i$ is very close to a singleton cluster's center $\mathbf{v}_j$ but far from the others, it will be committed to this singleton cluster w_j. The mass of belief $m_{\mathbf{x}_i}(w_j)$ of $\mathbf{x}_i$ committed to w_j should be determined according to the distance $d_{\mathbf{x_i v_j}}$ between $\mathbf{x}_i$ and the center $\mathbf{v}_j$, which is similar to FCM and ECM.

If the object $\mathbf{x}_i$ is too far from all the clustering centers according to the given outlier threshold δ, it will be naturally considered as outlier.

If the object $\mathbf{x}_i$ is simultaneously close to several singleton cluster centers as $\mathbf{v_j}$, $\mathbf{v_{j+1}}$,..., $\mathbf{v_t}$, and these centers appear undistinguishable for $\mathbf{x}_i$ (it indicates the object is close to the mean value of these singleton cluster centers), then it will be very hard to correctly commit this object to a particular singleton cluster. So it should be better to consider this object in a meta-cluster represented by the disjunction of these several singleton clusters as $w_j \cup w_{j+1} \ldots \cup w_t$, which can well reveal the imprecision of the class of this object and can also decrease the misclassification errors. The belief committed to a meta-cluster should depend on the distances between the object and the centers of these singleton clusters included in the meta-cluster, as well as on the distance to the meta-cluster' center defined by the mean value of these singleton cluster centers.

If the frame of discernment Ω contains $|\Omega| = c$ elements, the credal partition over the power-set 2^Ω will produce $2^{|\Omega|} = 2^c$ clusters including c singleton clusters, $2^c - c - 1$ meta-clusters, and 1 outlier cluster. When Ω contains a large number of elements (i.e. $|\Omega| = c$ is a big value), the computation burden of credal partition will be very high, which is a serious problem for the application. For example, if $|\Omega| = 5$, it produces $2^{|\Omega|} = 2^5 = 32$ clusters. In the real applications, the class of imprecise objects are usually unspecific among very few singleton clusters (which means the cardinality of the associated meta-cluster is small), and there is even no objects belonging to the meta-cluster with big cardinality. So we do not have to consider all the meta-clusters in 2^Ω, and we can eliminate some meta-clusters with big cardinality according to a given threshold $t_c \in [2, 2^{|\Omega|}]$. If the cardinality of a meta-cluster A_j is bigger than the given threshold as $|A_j| > t_c$, A_j will be removed from 2^Ω. In CCM, the set of the selected available clusters is given by $S^\Omega = \{A_i || A_i| \leq t_c\} \subseteq 2^\Omega$[1]. t_c is typically chosen as a small integer number, say two or three. This can effectively reduce the size of the credal partition, and it has very little influence on the clustering results, which will be shown in our experiments. For example, if one has $|\Omega| = 5$, the original size of the credal partition is $2^{|\Omega|} = 2^5 = 32$ clusters, but if we take $t_c = 2$, we just select 16 clusters since the other 16 meta-clusters whose cardinality is bigger than $t_c = 2$ are eliminated.

The objective function of CCM denoted by J_{CCM} is designed according to this basic principle, and it is given by

$$J_{CCM}(M, V) = \sum_{i=1}^{n} \sum_{j/A_j \in S^\Omega} m_{ij}^\beta D_{ij}^2 \tag{6.1}$$

with

$$D_{ij}^2 = \begin{cases} \delta^2; & A_j = \emptyset \\ d_{ij}^2; & |A_j| = 1, \\ \frac{\sum_{A_k \in A_j} d_{ik}^2 + \gamma d_{ij}^2}{|A_j| + \gamma}; & |A_j| > 1. \end{cases} \tag{6.2}$$

where $M = (m_1, \cdots, m_n) \in \mathbb{R}^{n \times 2^{|\Omega|}}$ is the mass of belief matrix for all objects, and $V_{c \times p}$ is the matrix of clustering centers. In the objective function J_{CCM}, we just consider the selected clusters

[1] The cardinality of outlier cluster is usually defined by $|\emptyset| = 0$.

in S^{Ω}, and J_{CCM} should be satisfied with the following constraint:

$$\sum_{j|A_j \in S^{\Omega}} m_{ij} = 1 \tag{6.3}$$

d_{ij} is the distance between the data point $\mathbf{x}_i$ and the center of cluster A_j. If A_j is a singleton cluster, its center is $\mathbf{v_j}$. If A_j is a meta-cluster, its center is defined by the mean value of the singleton clusters included in A_j as eq. (2.18).

The tuning parameters β is a weighting exponent, and $\beta = 2$ can be used as default value as in FCM [39], and ECM [23] approaches. δ is a chosen outlier threshold, which is strongly data-dependent, and it can be determined according to the outlier rate one expects, for example five percent. The bigger δ causes the fewer outliers. γ is the weighting factor of the distance between the object and the center of the meta-clusters, and it is generally used to control the number of objects in the meta-clusters. The bigger γ usually leads to more objects in the meta-clusters and the fewer misclassification errors. So the γ should be tuned to find a good compromise between the imprecision (corresponding to the number of objects in the meta-clusters) rate and error rate of the clustering results. In practice, γ can be determined according to the imprecision rate that the user is ready to accept. We generally suggest to take $\gamma \in [0.5, 3]$ according to our experience acquired in different applications.

Objective function J_{CCM} can be well justified as follows:

- The belief of an object on outlier cluster represented by $\emptyset$ is mainly determined by the given outlier threshold δ.
- The belief of an object on a singleton cluster is proportional to the distance between the object and the center of the singleton cluster. The smaller distance leads to the bigger belief.
- The belief of an object on a meta-cluster is proportional to the average distance to the involved singleton cluster centers, and also to the distance to the meta-cluster's center with a tuning factor γ. If the object $\mathbf{x}_i$ is closer to the centers $\mathbf{v}_j$, $\mathbf{v}_{j+1}$,..., $\mathbf{v}_t$, it indicates that $\mathbf{x}_i$ has potentially more chance to belong to the classes w_j, w_{j+1}, ..., w_t than to other clusters. Meanwhile, if the distances between the object and meta-cluster's center $\mathbf{v} = \frac{\mathbf{v_j}+\cdots+\mathbf{v_t}}{t-j+1}$ is smaller, it reflects the fact that these involved singleton clusters are more likely to be undistinguishable for the object. Then the belief committed to the meta-cluster $w_j \cup \cdots \cup w_t$ will be bigger.

In CCM, the mass of belief matrix $M = (\mathbf{m}_1, \cdots, \mathbf{m}_n)$ and the clustering centers matrix $V_{c\times p}$ can be obtained by the minimization of the objective function J_{CCM}.

6.2.2 Minimization of the objective function J_{CCM}

To minimize J_{CCM}, we use Lagrange multipliers method. In the first step, the centers of the clusters V are considered fixed. Lagrange multipliers λ_i are used to solve the constrained minimization problem with respect to M as follows:

$$\mathcal{L}(M, \lambda_1, \ldots, \lambda_n) = J_{CCM}(M, V) - \sum_{i=1}^{n} \lambda_i \Big(\sum_{j|A_j \in S^{\Omega}} m_{ij} - 1 \Big) \tag{6.4}$$

By differentiating the Lagrangian with respect to the m_{ij} and λ_i and setting the derivatives to zero, we obtain:

$$\frac{\partial \mathcal{L}}{\partial m_{ij}} = \beta m_{ij}^{\beta-1} D_{ij}^2 - \lambda_i = 0 \tag{6.5}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \sum_{j|A_j \in S^\Omega} m_{ij} - 1 = 0 \tag{6.6}$$

We thus have from (6.5)

$$m_{ij} = (\frac{\lambda_i}{\beta})^{\frac{1}{\beta-1}} (\frac{1}{D_{ij}^2})^{\frac{1}{\beta-1}} \tag{6.7}$$

using (6.6) and (6.7)

$$(\frac{\lambda_i}{\beta})^{\frac{1}{\beta-1}} = \frac{1}{\sum\limits_{j|A_j \in S^\Omega} D_{ij}^{\frac{2}{\beta-1}}} \tag{6.8}$$

Returning in (6.7), one obtains the necessary condition of optimality for M:

$$m_{ij} = \frac{D_{ij}^{\frac{-2}{\beta-1}}}{\sum\limits_{A_k \in S^\Omega} D_{ik}^{\frac{-2}{\beta-1}}} \tag{6.9}$$

Using (6.2), we can get the following masses of belief respectively committed to different focal elements including singleton cluster, meta-cluster and outlier cluster:

$$m_{ij} = \frac{\delta^{\frac{-2}{\beta-1}}}{\sum(D)}; \quad A_j = \emptyset \tag{6.10}$$

$$m_{ij} = \frac{d_{ij}^{\frac{-2}{\beta-1}}}{\sum(D)}; \quad |A_j| = 1 \tag{6.11}$$

$$m_{ij} = \frac{(\frac{\sum\limits_{A_k \in A_j} d_{ik}^2 + \gamma d_{ij}^2}{|A_j| + \gamma})^{\frac{-1}{\beta-1}}}{\sum(D)}; \quad |A_j| > 1 \tag{6.12}$$

where the denominator in these formulas, denoted $\sum(D)$, is given by:

$$\sum(D) = \sum_{A_j = \emptyset} \delta^{\frac{-2}{\beta-1}} + \sum_{A_j \in S^\Omega s.t. |A_j|=1} d_{ij}^{\frac{-2}{\beta-1}} + \sum_{A_j \in S^\Omega s.t. |A_j|>1} (\frac{\sum\limits_{A_k \in A_j} d_{ik}^2 + \gamma d_{ij}^2}{|A_j| + \gamma})^{\frac{-1}{\beta-1}} \tag{6.13}$$

Now let us consider that M is fixed. The minimization of J_{CCM} with respect to V is an unconstrained optimization problem. The partial derivatives of J_{CCM} with respect to the centers

are given by:

$$\frac{\partial J_{CCM}}{\partial \mathbf{v}_l} = \sum_{i=1}^{n} \sum_{A_l \cap A_j \neq \emptyset} m_{ij}^{\beta} \frac{\partial D_{ij}^2}{\partial \mathbf{v}_l} \tag{6.14}$$

with

$$\frac{\partial D_{il}^2}{\partial \mathbf{v_l}} = 2(\mathbf{v_l} - \mathbf{x}_i), \quad |A_l| = 1 \tag{6.15}$$

$$\frac{\partial D_{ij}^2}{\partial \mathbf{v_l}} = \frac{2(\mathbf{v_l} - \mathbf{x}_i) + \frac{2\gamma}{|A_j|}(\frac{\sum_{A_g \in A_j} \mathbf{v_g}}{|A_j|} - \mathbf{x}_i)}{|A_j| + \gamma}, \quad A_l \in A_j, |A_j| > 1 \tag{6.16}$$

Thus,

$$\frac{\partial J_{CCM}}{\partial \mathbf{v}_l} = \sum_{i=1}^{n} 2m_{il}^{\beta}(\mathbf{v}_l - \mathbf{x}_i) + \sum_{i=1}^{n} \sum_{A_l \in A_j} m_{ij}^{\beta} \frac{2(\mathbf{v}_l - \mathbf{x}_i) + \frac{2\gamma}{|A_j|}(\frac{\sum_{A_g \in A_j} \mathbf{v}_g}{|A_j|} - \mathbf{x}_i)}{|A_j| + \gamma} \tag{6.17}$$

Setting these derivatives to zero gives c linear equations that can be written as:

$$(\sum_{i=1}^{n} m_{il}^{\beta} + \sum_{i=1}^{n} \sum_{A_l \in A_j} m_{ij}^{\beta} \frac{1 + \frac{\gamma}{|A_j|}}{|A_j| + \gamma})\mathbf{x}_i = \sum_{i=1}^{n} m_{il}^{\beta} \mathbf{v}_l + \sum_{i=1}^{n} \sum_{A_l \in A_j} m_{ij}^{\beta} \frac{\mathbf{v}_l + \frac{\gamma \sum_{A_g \in A_j} \mathbf{v}_g}{|A_j|^2}}{|A_j| + \gamma} \tag{6.18}$$

The system of linear equations can be written more concisely as:

$$B_{c \times n} X_{n \times p} = H_{c \times c} V_{c \times p} \tag{6.19}$$

where

$$B_{li} \triangleq m_{il}^{\beta} + \sum_{A_l \in A_j} m_{ij}^{\beta} \frac{1 + \frac{\gamma}{|A_j|}}{|A_j| + \gamma} \tag{6.20}$$

$$H_{ll} \triangleq \sum_{i=1}^{n} m_{il}^{\beta} + \sum_{i=1}^{n} \sum_{A_l \in A_j} m_{ij}^{\beta} \frac{1 + \frac{\gamma}{|A_j|^2}}{|A_j| + \gamma} \tag{6.21}$$

$$H_{lq} \triangleq \sum_{i=1}^{n} \sum_{A_l \in A_k, A_q \in A_k} m_{ik}^{\beta} \frac{\gamma}{|A_k|^2(|A_k| + \gamma)}, l \neq q \tag{6.22}$$

V is the solution of the above linear system (6.19), and it can be solved by using a standard linear system solver as

$$V_{c \times p} = H_{c \times c}^{-1} B_{c \times n} X_{n \times p} \tag{6.23}$$

Table 6.1 : Credal C-Means algorithm.

Input:	Data to cluster: $X = \{\mathbf{x_1}, \cdots, \mathbf{x_n}\}$ in $\mathbb{R}^p$
Parameters:	c: number clusters, $2 \leq c < n$
	t_c: meta-cluster threshold (by default take $t_c = 3$)
	$\delta > 0$: outlier threshold
	$\gamma > 0$ weight of the distance(by default take $\gamma = 1$)
	$\epsilon > 0$: termination threshold (by default take $\epsilon = 0.001$)
Initialization:	Choose randomly initial mass M_0
	t $\leftarrow 0$
	Repeat
	$t \leftarrow t + 1$
	Compute B_t and H_t by (6.20)-(6.22);
	Compute V_t by solving (6.23);
	Compute M_t using (6.10)-(6.12);
	until $\|\|V_t - V_{t-1}\|\| < \epsilon$

The pseudo-code of the CCM algorithm is given in Table 6.5 for convenience.

The initial BBA M_0 can be randomly generated, and the final clustering results are not very sensitive to the choice of the initialization of M_0 after the process of optimization. CCM is an extension of FCM with the credal partition, and its convergence is similar to FCM. These properties of CCM have been validated in our tests and applications.

In CCM, the number of clusters c can be selected according to some prior experience or knowledge. If no prior knowledge about the value of c is available, it can be determined by minimizing the validity index of a credal partition as the average normalized specificity following the idea proposed in [23]:

$$N^*(c) = \frac{1}{n \log_2(c)} \sum_{i=1}^{n} [\sum_{A \in 2^\Omega \setminus \emptyset} m_i(A) \log_2 |A|] \tag{6.24}$$

where $0 \leq N^*(c) \leq 1$. The validation of the method has been tested in [23].

In some applications, the approximations of credal partition to the fuzzy (probabilistic) partition may be useful. Then, the meta-clusters and the outlier cluster need to be eliminated, and their masses of belief should be redistributed to the other clusters. The Pignistic probability transformation $BetP(.)$ introduced by Smets in his Transferable Belief Model (TBM) [5,6] is commonly used to transfer a BBA to the probability measure, and it is rewritten here for convenience .

$$BetP(w) = \sum_{w \in A} \frac{m(A)}{|A|(1 - m(\emptyset))}, w \in \Omega \tag{6.25}$$

Besides that, the lower and upper bounds of imprecise probability associated with BBA's can be approximated using the belief function $Bel(.)$ as eq. (2.2), plausibility function $Pl(.)$ as eq. (2.3) in credal partition context [3]. $[Bel(A), Pl(A)]$ is interpreted as the interval characterizing the imprecision of the unknown underlying probability measure on A. $Bel(.)$ and $Pl(.)$ can also be used for decision-making support if necessary.

6.3 EXPERIMENTS

Four experiments have been applied to test our method. Experiment 1 is a particular 3-class data set, and it is used to explicitly show the use of CCM and the limitations of ECM. Experiment 2 is

a particular 4-class data set, and it is given to show again the limitations of ECM and also to show the interest of the meta-cluster threshold t_c in CCM. Experiment 3 is given to simply illustrate the potential use of the CCM in the supervised classification of remote sensing image. Experiment 4 with five real data sets is presented to evaluate the effectiveness of CCM with respect to ECM and FCM.

In order to show the interest of the use of meta-class in CCM, the decision of classification is made by a simple criterion where the class of the object receives the biggest mass of belief in the following experiments.

In our data clustering analysis, the masses of belief of the object is unknown at the beginning, and the initializations of mass of belief M are randomly generated in our experiments. This is a very common way used in the clustering methods (e.g. ECM, FCM, CCM). The initial mass values must be positive in [0, 1] and satisfy the constraint of eq. (6.3), which stipulates that the sum of mass values of one object associated with each cluster must be normalized to one. In fact, the initializations of M have no effect on the final results due to the convergence of the clustering methods, e.g. CCM, ECM and FCM. This behavior has always been demonstrated in our results of simulations. We have repeated all the tests many times with different random generations of M and the clustering results are very similar.

In CCM, the number of objects in meta-clusters corresponding to the imprecision level can be controlled by the tuning of the parameter γ. The bigger value of γ generally causes higher imprecision level and lower error rate. So one should find a proper compromise between the error and imprecision depending on the actual applications. The influence of different values of γ on the clustering results have been evaluated in our experiments.

In the following figures, as Fig. 6.1 and 6.2, the x-axis and y-axis respectively correspond the first and the second dimension of data (2-D vector) used in the experiments.

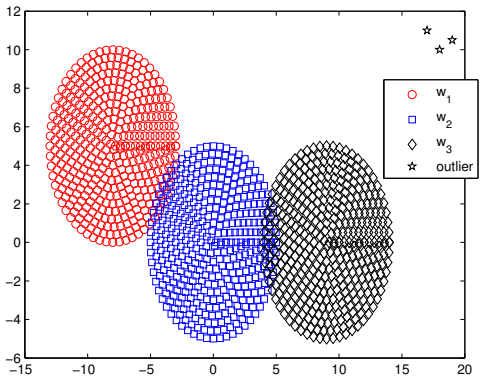
6.3.1 Experiment 6.1 (with 3-class artificial data set)

In the first experiment, we consider a particular 3-class data set in the round shape as shown in 6.1-a. This data set consists of 1245 data points including 3 noisy data. The radius of the circle is $r = 5$, and the centers of three circles are given by

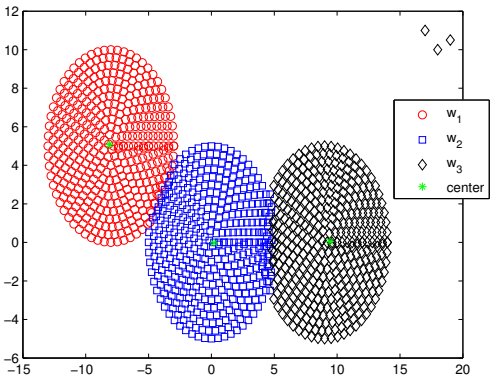
$$\begin{aligned} \mathbf{c_1}: &\quad [-8, \quad 5]' \\ \mathbf{c_2}: &\quad [0, \quad 0]' \\ \mathbf{c_3}: &\quad [9, \quad 0]' \end{aligned}$$

FCM, ECM and CCM are applied for the clustering this particular data set. The outlier threshold used in ECM and CCM is $\delta^2 = 49$. One takes different value of $\gamma = 1$ and $\gamma = 0.5$ in CCM, and $\alpha = 1$ and $\alpha = 0.5$ in ECM to test the effect of the tuning parameters on the results. In CCM, all the meta-clusters are kept (i.e. $t_c = 3$). The clustering results obtained with FCM, ECM and CCM are shown in Fig. 6.1-b-f. For the convenience of denotation, we defined $w_{i,\ldots,j} \triangleq w_i \cup \ldots \cup w_j$.

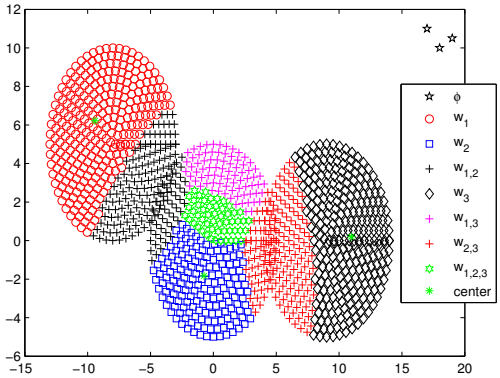
The clustering centers are an important criteria to evaluate the performance of clustering method. The clustering centers $\mathbf{v}$ obtained by different methods and their distances $\mathbf{d}(\mathbf{v}, \mathbf{c})$ to the original centers of the three circles are given in Table 6.2, and we also show the average distance $\overline{\mathbf{d}}(\mathbf{v}, \mathbf{c})$ between the clustering centers and the circles' centers for convenience.



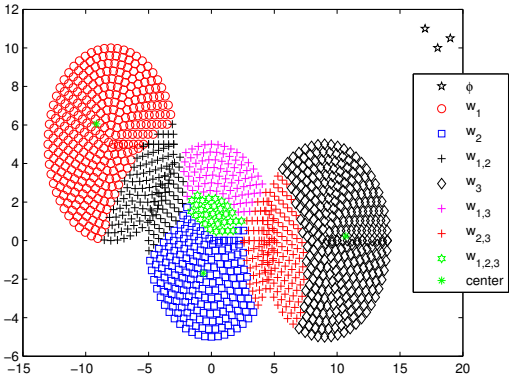
a. Original data set.



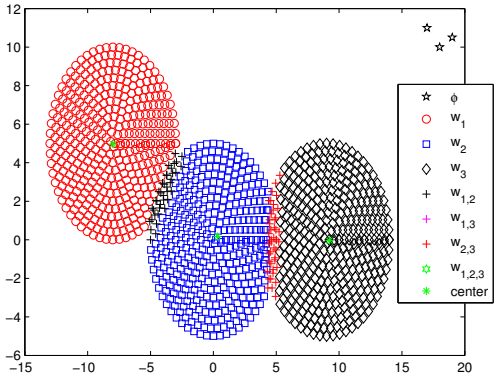
b. Clustering result by FCM.



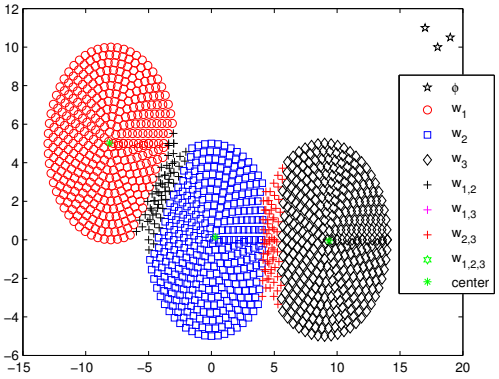
c. Clustering result by ECM with $\alpha = 0.5$.



d. Clustering result by ECM with $\alpha = 1$.



e. Clustering result by CCM with $\gamma = 0.5$.



f. Clustering result by CCM with $\gamma = 1$.

Figure 6.1 : Clustering results for the 3-class data set by different methods.

It is considered better if the clustering centers are closer to the given circles' centers. FCM and CCM provide the similar clustering centers, and it is good that their distances to the given circles' centers is very small. Whereas, clustering centers obtained by ECM are not so close to the given circles' centers, which reflects that the performance of ECM is not very good.

Table 6.2 : Clustering centers with different methods.

	V	$\mathbf{d}(\mathbf{v},\mathbf{c})$	$\overline{\mathbf{d}}(\mathbf{v},\mathbf{c})$
FCM	$\mathbf{v_1}$ (-8.1062, 5.0792)	0.1327	
	$\mathbf{v_2}$ (0.1997, -0.0248)	0.2012	0.2522
	$\mathbf{v_3}$ (9.4209, 0.0392)	0.4227	
ECM ($\alpha = 0.5$)	$\mathbf{v_1}$ (-9.4147, 6.2224)	1.8697	
	$\mathbf{v_2}$ (-0.7470, -1.8270)	1.9738	1.9407
	$\mathbf{v_3}$ (10.9682, 0.2019)	1.9785	
ECM ($\alpha = 1$)	$\mathbf{v_1}$ (-9.1464, 6.0584)	1.5603	
	$\mathbf{v_2}$ (-0.6453, -1.6931)	1.8119	1.6901
	$\mathbf{v_3}$ (10.6826, 0.2291)	1.6981	
CCM ($\gamma = 0.5$)	$\mathbf{v_1}$ (-7.9931, 4.9661)	0.0346	
	$\mathbf{v_2}$ (0.3307, 0.1789)	0.3760	0.2172
	$\mathbf{v_3}$ (9.2375, -0.0402)	0.2409	
CCM ($\gamma = 1$)	$\mathbf{v_1}$(-8.0995, 5.0176)	0.1010	
	$\mathbf{v_2}$(0.3072, 0.1407)	0.3379	0.2581
	$\mathbf{v_3}$ (9.3520, -0.0494)	0.3554	

We can see on Fig. 6.1-a, the class w_2 is partly overlapped with w_1 and w_3 on their borders, and these points in the overlapped zone are really difficult to be clearly classified. FCM produces 3 singleton clusters w_1, w_2 and w_3 based on the probability framework, and the points in the overlapped zone are all committed to a singleton cluster, which is likely to cause the misclassification errors. The three outliers quite far from the other data cannot be detected by FCM, and they are simply considered belonging to w_3.

ECM provides the credal partition in belief functions framework. We can see that w_1 and w_3 are not close and they are totally separate, but there are still many points originally from w_2 that are wrongly committed to the meta-cluster $w_1 \cup w_3$ labeled by purple plus symbol in Fig. 6.1-c,d. Even worse, that many points from w_2 labeled by green hexagon are even considered in the total ignorant cluster $w_1 \cup w_2 \cup w_3$. These unreasonable results are produced mainly because the clustering centers $\mathbf{v}_2$, $\mathbf{v}_{\{1,3\}} = \frac{\mathbf{v}_1+\mathbf{v}_3}{2}$ and $\mathbf{v}_{\{1,2,3\}} = \frac{\mathbf{v}_1+\mathbf{v}_2+\mathbf{v}_3}{3}$ are close to each other, and the belief on each cluster is determined mainly by the distance between the object and the corresponding clustering center.

In CCM, no point belongs to $w_1 \cup w_3$ or $w_1 \cup w_2 \cup w_3$. Some points in the middle of w_1 and w_2, w_2 and w_3 are respectively committed to $w_1 \cup w_2$ and $w_2 \cup w_3$ in Fig. 6.1-e,f, since these points are really difficult to classify into a particular class. This can effectively reduce the misclassification errors using meta-clusters. $w_1 \cup w_2$ can be interpreted as the transition class between w_1 and w_2, and it is similar to $w_2 \cup w_3$. It seems that CCM has much better performance than ECM and FCM for dealing with the meta-clusters. Three objects labeled by black pentagram are far form the others, and they are well detected by ECM and CCM.

If the tuning parameter α increases from $\alpha = 0.5$ to $\alpha = 1$ in ECM as Fig. 6.1-c,d, the number of points committed to the meta-clusters obviously decreases, since the distances used to determine the mass of belief of the meta-clusters are greatly penalized. The increase of γ from $\gamma = 0.5$ to $\gamma = 1$ in CCM as shown in Fig. 6.1-e,f generates few extra points committed to the meta-clusters, but the clustering results is not so sensitive to γ in CCM as ECM is to changes of its tuning parameter α.

6.3.2 Experiment 6.2 (with 4-class simulated data set)

A particular 4-class data set as in Fig. 6.2-a is applied here. We use $4 * 50 = 200$ data points generated from four 2D Gaussian distributions characterized by ($\mathbf{I}$ being the 2×2 identity matrix):

$$\begin{aligned} \mu_1 &= [-2, \quad 0]', \quad \Sigma_1 = 1.5\mathbf{I} \\ \mu_2 &= [-9, -10]', \quad \Sigma_2 = \mathbf{I} \\ \mu_3 &= [3, \quad 0]', \quad \Sigma_3 = 1.5\mathbf{I} \\ \mu_4 &= [9, \quad 8]', \quad \Sigma_4 = \mathbf{I} \end{aligned}$$

where μ is the mean value and Σ is variance value.

The number of singleton clusters is set to $c = 4$. The parameters in ECM and CCM are $\delta = 7$, $\gamma = 1$ and $\alpha = 1$, and the different value of t_c is selected here for showing its influence on the results. The clustering results obtained with FCM, ECM and CCM are shown in Fig. 6.2-b-d.

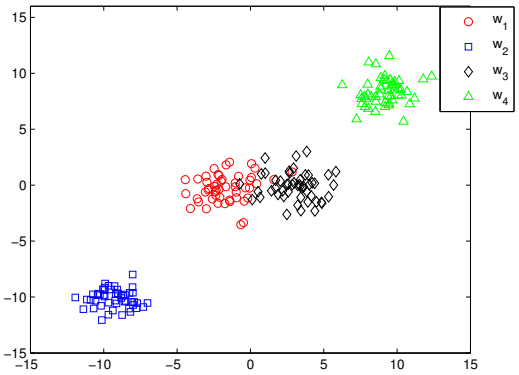
We can see from the original simulated data set in Fig. 6.2-a that w_2 and w_4 are far from each other, whereas w_1 and w_3 are close and even partly overlapped on their border. The objects on the border are really difficult to classify, and they should be prudently committed to the meta-cluster $w_1 \cup w_3$ to avoid the misclassification error. FCM produces only four singleton clusters without meta-clusters, and some objects on the border between w_1 and w_3 will be very likely wrongly classified. In ECM, a number of data points original from w_1 and w_3 are unreasonably and even wrongly committed to many meta-clusters including $w_1 \cup w_3, w_2 \cup w_4, w_1 \cup w_2 \cup w_3, w_2 \cup w_3 \cup w_4, w_1 \cup w_2 \cup w_4$ and even total ignorant cluster $w_1 \cup w_2 \cup w_3 \cup w_4$ shown by Fig. 6.2-c. The reason of this counterintuitive behavior is that these meta-cluster centers are close to the objects belonging to w_1 or w_3. This particular example explicitly show the limitations of ECM.

Once CCM is applied as shown in Fig. 6.2-d,e,f, only six points labeled by blue plus symbol in the middle of w_1 and w_3 are classified into $w_1 \cup w_3$, which consistent with what we intuitively expect. CCM is able to produce better results than ECM because the distances between the object and the involved singleton cluster centers are additionally taken into account in the determination of the belief on the meta-cluster.

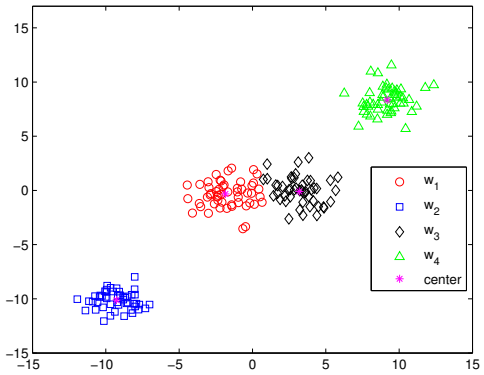
If one selects $t_c = 4$, it means that all the meta-clusters produced by the four classes are available in CCM, then the clustering results contains 16 clusters with 11 meta-clusters. If one takes $t_c = 3$, the total ignorant cluster $w_1 \cup w_2 \cup w_3 \cup w_4$ will be eliminated since its cardinality is $|w_1 \cup w_2 \cup w_3 \cup w_4| = 4 > 3$. When this threshold decreases to $t_c = 2$, it means that the meta-clusters whose cardinalities are bigger than 2 (i.e. 3 or 4) will be removed, and there will remain only 11 clusters with 6 meta-clusters. Then the computation complexity of the credal partition will become much smaller. Whereas, one can see that the clustering results with different value of t_c (i.e. $t_c = 2, 3, 4$) are almost the same as shown in Fig.6.2-d,e,f. This indicates that the small value of t_c can be used in the applications, and this can greatly reduce the computation burden in maintaining good clustering results.

6.3.3 Experiment 6.3 (with real remote sensing data)

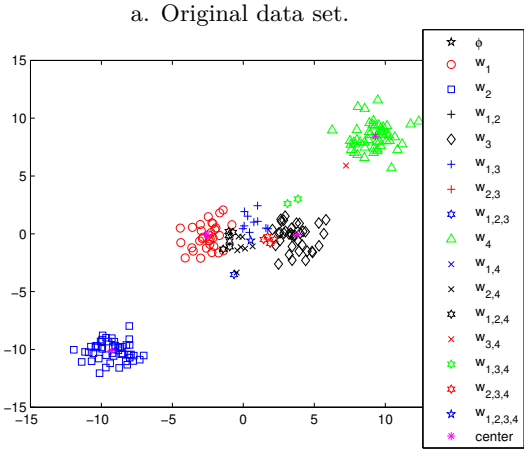
In this example, we show the potential use of CCM in the unsupervised classification of the remote sensing images with respect to FCM and ECM. A small piece QuickBird satellite image about urban region in Fig. 6.3-a is applied here, and it mainly includes wooded area, bared soil, and building area. The clustering results by different methods are shown in Fig. 6.3-b-d. The outlier



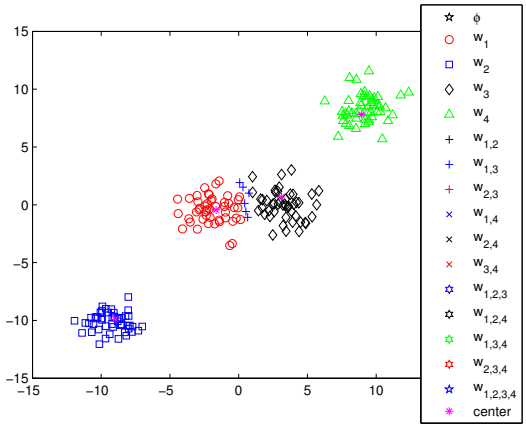
a. Original data set.



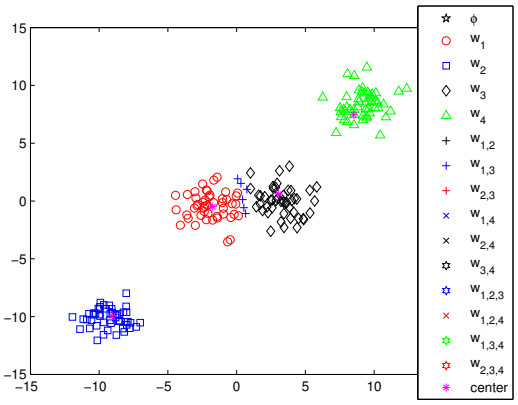
b. Clustering result by FCM.



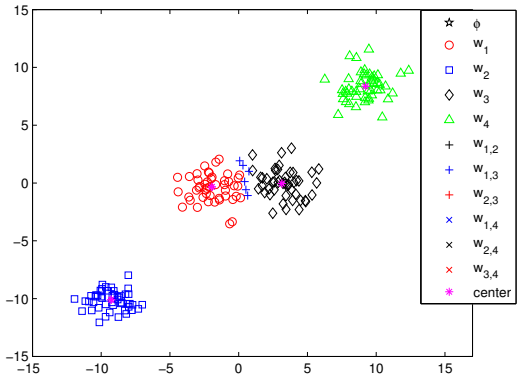
c. Clustering result by ECM.



d. Clustering result by CCM with $t_c = 4$.



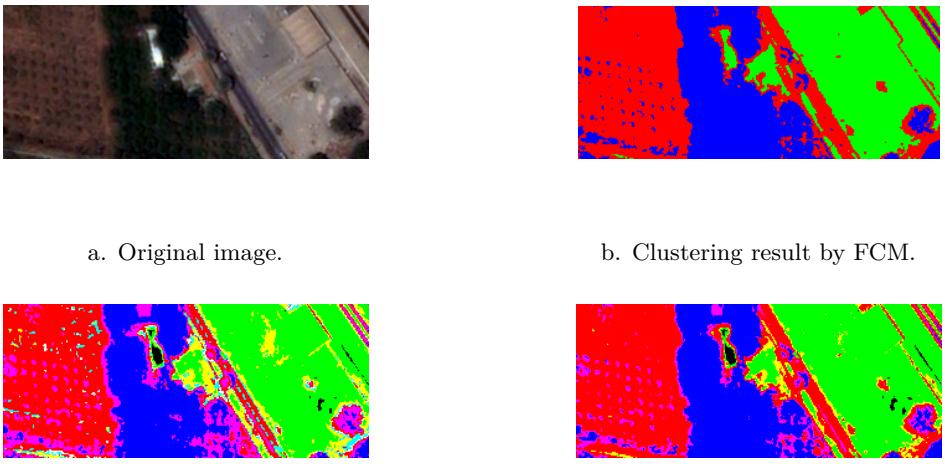
e. Clustering result by CCM with $t_c = 3$.



f. Clustering result by CCM with $t_c = 2$.

Figure 6.2 : Clustering results by different methods for 4-class data set .

threshold in CCM and ECM is $\delta = 70$. We select the default value of $\alpha = 1$ in ECM, since it provides a good result. $\gamma = 2$, and $t_c = 3$ have been used in CCM.



a. Original image. b. Clustering result by FCM.

c. Clustering result by ECM . d. Clustering result by CCM.

Figure 6.3 : Clustering results by different methods .

Table 6.3 : Class description of the classifications in image.

Class	Color in Fig. 6.3-b,c,d	description
$\emptyset$	dark area	outlier
w_1	red area	bared soil
w_2	green area	building area
w_3	blue area	wooded area
$w_1 \cup w_2$	yellow area	
$w_1 \cup w_3$	purple	
$w_2 \cup w_3$	cyan area	
$w_1 \cup w_2 \cup w_3$	white area	

The descriptions of the classifications are given in Table 6.3. FCM just produces 3 specific clusters for the image, but the border of the different classes regions are not so clear. Thus, the contents on the border correspond probably to false classifications in FCM. ECM and CCM mainly commit the contents on the border of the different classes to the meta-clusters, which is more prudent and reasonable than FCM. Nevertheless, ECM commits some points in the bared soil w_1 on the left side of the image into the incompatible meta-cluster[2] $w_2 \cup w_3$ labeled by cyan color and into the total ignorant cluster $w_1 \cup w_2 \cup w_3$ labeled by white color in Fig. 6.3-c, which is not very reasonable. With CCM, there is no point in the bared soil w_1 on the left side committed to $w_2 \cup w_3$ or $w_1 \cup w_2 \cup w_3$. Moreover, several white areas in Fig. 6.3-a are detected by outlier clustered using both ECM and CCM as shown in Fig. 6.3-c with dark color. This indicates that this special area is quite different from the other areas according to the pixel values. The clustering results by CCM show that this SPOT image is insufficient to obtain the specific classification of the observed area. Some other available sources of images, for example the radar image, etc, must be used for the fusion with this image to obtain more precise results.

[2]Clusters A and B are compatible if $A \cap B \neq \emptyset$, and they are incompatible if $A \cap B = \emptyset$.

6.3.4 Experiment 6.4 (with real data sets)

In this experiment, we use five well-known real data sets, i.e. Statlog (Heart), Iris, Ecoli, Seeds and Wine from UCI Repository [95] to test the performance of CCM with respect to ECM and FCM. In Ecoli data set, the three classes named as *im*, *pp* and *imU* are close and difficult to correctly classify, and they are selected for the evaluation of our method. The classes *im*, *pp* and *imU* contains respectively 77, 52 and 35 samples. So there are 164 samples available, and each sample contains 7 attributes. The basic attributes information of the used data sets are shown in table 6.4. The detailed information of all the used data sets can be found on UCI repository archive at `http://archive.ics.uci.edu/ml/`.

In the clustering analysis of these data sets, the instances are represented using several attributes which are all measured by real numbers here. In fact, the instance is denoted by a numerical vector, and each attribute corresponds to one dimension of the vector. For example, Iris data set can be considered as a 3-class problem with 150 samples (instances) to be clustered, and each sample is a 4-dimension numerical vector. This is similar to the other real data sets. Then, the clustering method FCM, ECM and CCM can be applied for the clustering of these real data sets as similarly done in the normal numerical data sets. The masses of belief (or fuzzy membership for FCM) of the object associated with different clusters can be obtained by optimization of the objective function described in the clustering methods (e.g. FCM, ECM and CCM).

The outlier threshold applied in ECM and CCM for both data sets is $\delta = 20$, and different values of γ, t_c and α are chosen to test their effect on the results. It is worth to note that $t_c = 3$ means that all the meta-clusters are kept, whereas $t_c = 2$ means that we just select the meta-clusters with the cardinality value of two. The clustering results of the two data sets by different methods are respectively shown in Tables 6.5–6.9 where "NA" means "Not applicable".

In this work, the common error rate Re and imprecision rate Ri_j corresponding to the meta-class are used to evaluate the performance of CCM. They are defined in the same way as already presented in Chapter 3.

Table 6.4 : Basic information of the applied data sets.

name	classes	attributes	instances
Statlog (Heart)	2	13	270
Iris	3	4	150
Seeds	3	7	210
Ecoli	3	7	164
Wine	3	13	178

Table 6.5 : Clustering results of Statlog (Heart) data set with different methods (in %).

		Re	Ri_2
FCM		40.74	NA
ECM	α=2.0	36.67	7.41
	α=1.5	34.81	12.22
	α=1	34.81	19.63
CCM ($t_c = 2$)	γ=1.0	34.44	9.26
	γ=1.5	33.70	14.07
	γ=2.0	33.33	14.81

With these real data sets, FCM produces the most misclassification errors because of the limitations of the probability framework. The number of classification errors generated by ECM

Table 6.6 : Clustering results of Iris data with different methods (in %).

		Re	Ri_2	Ri_3
FCM		10.67	NA	NA
ECM	α=2.0	8.00	4.67	0
	α=1.5	10.00	8.67	0.67
	α=1.0	10.00	15.33	6.00
CCM ($t_c = 3$)	γ=1.0	5.33	8.00	0
	γ=1.5	4.67	10.67	0
	γ=2.0	4.00	12.00	0
CCM ($t_c = 2$)	γ=1.0	5.33	8.00	NA
	γ=1.5	4.67	10.00	NA
	γ=2.0	3.33	12.00	NA

Table 6.7 : Clustering results of Seeds data with different methods (in %).

		Re	Ri_2	Ri_3
FCM		10.48	NA	NA
ECM	α=2.0	7.62	10.48	0.95
	α=1.5	5.24	14.76	2.38
	α=1.0	5.24	18.10	4.76
CCM ($t_c = 3$)	γ=1.0	5.71	10.00	0
	γ=1.5	5.24	12.86	0
	γ=2	5.24	14.29	0
CCM ($t_c = 2$)	γ=1.0	5.71	10.00	NA
	γ=1.5	5.24	12.86	NA
	γ=2	5.24	14.29	NA

Table 6.8 : Clustering results of Ecoli data with different methods (in %).

		Re	Ri_2	Ri_3
FCM		25.00	NA	NA
ECM	α=2.0	24.39	1.22	0
	α=1.5	23.17	1.83	0
	α=1.0	21.34	7.32	0.61
CCM ($t_c = 3$)	γ=1.0	20.73	4.88	0
	γ=1.5	19.51	6.71	0
	γ=2.0	18.29	8.54	0
CCM ($t_c = 2$)	γ=1.0	20.73	5.49	NA
	γ=1.5	18.90	7.93	NA
	γ=2.0	18.29	9.15	NA

is a bit less than with FCM, but it causes too many samples committed to the meta-clusters (corresponding to the high imprecision degree of the results), and even some samples are considered belonging to the total ignorant class (i.e. the frame of discernment). Moreover, the result of ECM is very sensitive to the tuning of parameter α. When α increases, the number of false classifications will increase, but the number of the objects in meta-clusters will decrease.

CCM generally provides the smallest number of errors among the different methods for the Statlog (Heart), Iris, Seeds and Ecoli data sets, and the number of samples in meta-clusters is smaller than what we obtain with ECM. For the Wine data set, ECM has lower error rate than CCM, but it is with much higher imprecision rate, and it even commits some samples in the total

Table 6.9 : Clustering results of Wine data with different methods (in %).

		Re	Ri_2	Ri_3
FCM		31.46	NA	NA
ECM	α=3.0	21.35	23.03	2.81
	α=2.0	17.98	28.65	3.37
	α=1.5	16.85	33.15	3.93
CCM ($t_c = 3$)	γ=1.5	24.72	15.73	0
	γ=2	24.72	17.98	0
	γ=3.0	23.60	22.47	0
CCM ($t_c = 2$)	γ=1.5	24.72	15.73	NA
	γ=2	24.72	17.98	NA
	γ=3.0	23.60	22.47	NA

ignorant class, which is not an efficient solution of data clustering. In CCM, there is no sample committed to the ignorant cluster for all the applied data sets which shows that this new approach provides more specific clustering results than ECM. The increasing of γ causes the decreasing of the error but the increasing of imprecision degree. So we should find a compromise between the error and imprecision, and it also depends on the imprecision rate the user is ready to accept. Moreover, the clustering results are not very sensitive to γ in CCM contrariwise to ECM in regards with the parameter α.

In CCM, if the meta-cluster threshold t_c is changed from $t_c = 3$ to $t_c = 2$ for the 3-class data sets (i.e. Iris, Seeds, Ecoli and Wine data sets), it indicates that the meta-cluster whose cardinality value is three will be eliminated, and the computation complexity will decrease. However, we find that the clustering results are almost the same with different value of t_c. So it shows again that one can choose a small value of t_c in the real applications. Consequently, CCM can still provide good clustering results with an acceptable computational burden. The CCM results reflect that the used information is in fact not sufficient for making the specific classification of the samples in the meta-clusters. Therefore, other complementary information sources are really necessary if one wants to get more precise and correct classification results.

6.4 Conclusion

The credal c-means (CCM) clustering method has been introduced in this chapter to well characterize the uncertainty and imprecision of information. CCM working with credal partition can produce three kinds of clusters: singleton clusters, meta-clusters and outlier cluster. CCM can effectively reduce the misclassification errors using the meta-clusters, and it is also robust to the outliers. If one object is very close to a singleton cluster's center, it will be committed to this singleton cluster as done with FCM and ECM. If one object is simultaneously close to several singleton clusters, it will be considered in the meta-cluster defined by the disjunction of these singleton clusters, since these singleton clusters cannot be clearly distinguished by the object. This indicates the available information is not sufficient for making the specific classification of these objects in the meta-clusters, and these objects should be treated more cautiously. If an object is too far from the others according to the given outlier threshold, it will be naturally considered as outliers. A meta-cluster threshold is introduced in CCM to eliminate the meta-clusters with big cardinalities to reduce the computational burden of CCM, while maintaining very good clustering results. The credal partition can be easily approximated to a fuzzy partition if necessary, and the transformation rule has also been given. The output of CCM is not necessarily used for making the final classification of objects, but it can serve as an interesting source of information to combine

with additional complementary information sources if one wants to get more precise results. The effectiveness of CCM has been shown through different experiments using both artificial and real data sets.

7 Conclusion and perspectives

In this chapter, we conclude this thesis and we also propose several topics to explore in future research works.

7.1 Conclusion

The credal classification of uncertain data based on belief function theory has been studied in this thesis, and it allows the objects to belong to not only single (specific) classes but also to meta-class (i.e. set of several specific classes) with different masses of belief. The credal classification can well characterize the uncertainty and imprecision of classification, and can also efficiently reduce the errors thanks to the use of belief functions.

We have proposed four credal classification methods to deal with different encountered cases. When the training samples can be used for classification, belief $c \times K$ neighbors classifier is proposed to manage the partially overlapped classes in the general and complicate cases, but it requires a high computational burden which is the necessary price one has to pay for the complicate situation. If each class can be represented by one prototype (i.e. class center) vector, a simple credal classification rule has been developed which can directly compute the mass of belief of the object belonging to each class (e.g. single class and meta-class) with quite low computational complexity. Moreover, the credal classification of incomplete data with missing values has been also developed in this work, and the imprecision of classification due to the lack of information can be well modeled using the belief functions. In fact, the intrinsic nature of the uncertainty and imprecision in the overlapped case and incomplete case are the same, and it reflects the fact that the available (known) attribute information is insufficient for making the specific classification of these patterns. When the training information becomes unavailable, the data clustering analysis must be applied, and a new efficient credal clustering method called credal c-means (CCM) has been proposed for uncertain data. This new CCM approach is an efficient extension of Fuzzy c-means clustering method in the belief functions framework.

7.1.1 Belief c $\times$ K neighbors classifier (BCKN)

In the classification of uncertain data, the different classes can be partially overlapped in some cases, and the objects in the overlapped zones are quite difficult to correctly classify, since these overlapped classes appear undistinguishable for these objects according to the available attribute information. In order to well model such uncertainty and imprecision and to reduce the errors, a new belief $c \times K$ neighbors (BCKN) classifier has been developed working with credal classification based on the belief function theory. In BCKN, the query object is classified using its K nearest neighbors in each class, and there are total $c \times K$ neighbors involved in a c-class problem. Then, $c \times K$ BBA's are constructed corresponding to the $c \times K$ neighbors based on the distance between the

object and the neighbors, and the global fusion of these BBA's is used for the credal classification of the object. The object can be committed, with different masses of belief, to a particular class or to the proper meta-class, and the outlier represented by ignorant class can also been detected. The objects lying in the overlapped zone of different classes cannot be reasonably committed to a particular class, and they will be classified to the associated meta-class defined by the union of these different classes. The objects too far from the others will be naturally considered as outliers. This approach can reduce the misclassification errors by introducing the meta-class to characterize the partial imprecision of classification, Of course, this is achieved at the detriment of the overall classification precision, which is usually preferable in some applications. The output of the BCKN classifier can be used as a primary source of information to ask for other complementary means of analysis when more precise results on the ambiguous objects are necessary. The performance of BCKN method has been tested and evaluated with respect to other classical methods through several experiments using both synthetic data sets and real data sets. Our comparative analysis shows that BCKN is able to efficiently reduce the classification errors by increasing judiciously the imprecision rate in the applications, and a suitable compromise between the error rate and imprecision rate must always be found by optimizing the threshold parameter for the selection of the meta-classes in practice.

7.1.2 Credal classification rule (CCR)

BCKN is an efficient classifier of uncertain data for dealing with the complicate case, but it has the high computational complexity as other K-NN alike classifiers. In some simple cases, the data classes can be well characterized using the prototypes vectors, and we have proposed a new prototype-based credal classification rule (CCR), which provides a fast solution to directly calculate the mass of belief of the object associated with the single classes and meta-classes. In CCR, each class center (i.e. prototype) should be obtained at first, and the specific class center is simply defined by the arithmetic mean value of the training data in the corresponding class. The meta-class center is considered with the same (and as small as possible) distance to the centers of all the involved specific classes, and it can be found by optimizing these constraints. Once obtained, the effective meta-class will be selected according to the distances between the meta-class center and each specific class center. The useful meta-class center should be closer to the involved classes' centers than to other centers. The meta-class will be ignored if it doesn't satisfy the proposed conditions. The mass of belief of a given object with a specific class is determined from the Mahalanobis distance between the object and the center of the corresponding class. The mass of belief on the meta-class mainly depends on the distance between the object and the center of meta-class taking into account the associated indistinguishability degree defined using the distances of the object to centers of all the involved specific classes. A tuning threshold is used to detect the noises and outliers. The specific class consists of all the objects that are sufficiently close to its center. The meta-classes are used to capture the imprecision in the classification of the objects when they are difficult to correctly classify because of the poor quality of available attributes. The objects assigned to a meta-class should be close to the center of this meta-class and meanwhile its distances to all centers of the involved specific classes should be similar. The objects too far from the others will be considered as outliers (noise). The experiments using both artificial and real data sets were presented to evaluate and compare the performances of the new CCR method with respect to other classification methods. Our comparative analysis shows that CCR provides efficient credal classification results with a relatively low computational burden.

7.1.3 Credal classification of incomplete data with missing values

It can happen that the useful attribute values are missing in many applications. The edited incomplete pattern with different possible estimations of missing values may yield distinct classification results. A prototype-based credal classification (PCC) method for incomplete patterns has been proposed using belief functions to model the uncertainty (imprecision) of classification caused by the lack of information of the missing data. In PCC, the class prototypes obtained using training samples are used to estimate the missing values. For example, if one has to deal with c prototypes in a c-class problem, it yields c estimations of the missing values. Each estimation has a different weighting factor determined by the distance between the object and the corresponding prototype ignoring the missing values. The edited patterns with different estimations will be respectively classified using a normal classifier dealing with the complete pattern. Then, one can get multiple (e.g. c) classification results for the incomplete pattern. Since all these distinct classification results are potentially admissible, the final credal classification of the object depends on the global fusion of these classification results. However, these results cannot be equally treated, because the estimations of the missing values are obtained with different weights. The discounting technique will be applied in the classification results using the weighting factors of the associated estimations before the fusion process. The 2-step fusion strategy is used for the global fusion, and a new combination rule has been introduced. It transfers the conflicting beliefs to the proper selected meta-classes, since the conflicting beliefs can capture the imprecision degree of the classification delivered by different estimations of the missing values. The incomplete patterns that are very difficult to classify in a specific class will be reasonably committed to some proper meta-classes by the PCC method in order to reduce errors. Several experiments using real data sets have been presented to illustrate the effectiveness of PCC method. Our analysis indicates that PCC is able to efficiently reduce the error rate and well characterize the imprecision of classification thanks to the meta-class.

7.1.4 Credal c-means (CCM) clustering method

When no training information is available for the classification, data clustering analysis can be applied. Evidential C-means is an extension of FCM in the belief function framework, but it produces very unreasonable results on the clustering of close data sets. The credal c-means (CCM) clustering method working with credal partition has been developed in this thesis to overcome the limitations of ECM. The CCM allows the objects to lie in both the singleton clusters and the sets of clusters (i.e. meta-cluster) with different masses of belief. CCM can reduce the errors using the meta-cluster that is able to characterize the imprecision of classification for the partially overlapped clusters. In CCM, the mass of belief on each specific (single) cluster is proportional to the distance of the object to the corresponding clustering center. Whereas, the mass of belief on the meta-cluster is determined by both the distances of the object to the meta-cluster's center and to all the involved specific clusters' centers. If one object is too far from the other data points with respect to the given threshold, it will be considered as an outlier. According to this basic principle, an objective function has been proposed, and the clustering centers and the mass of belief of object belonging to each specific cluster and meta-cluster can be obtained by optimization of this objective function. For the convenience and simplicity of the linear optimization, the center of meta-cluster is defined by the mean value of the involved specific classes' centers. If the object is quite close to only one clustering center, and it is naturally committed to this cluster as done in the classical methods. If the object is simultaneously close to several clusters, and it is difficult to classify it correctly into a particular cluster, because these clusters are not easily distinguishable for this object. In such case, this object will be cautiously committed to the meta-cluster defined by the union of these several clusters. CCM is robust to the noisy data thanks to the outlier

cluster. The credal partition can be simply reduced to fuzzy partition as in FCM if necessary, and transformation way has been given using Pignistic probability transformation $BetP(.)$. The experimental evaluation based on synthetic and real data shows that CCM can well deal with the special cases where ECM does not work well, and CCM can also efficiently reduce the errors by capturing the imprecision of the classification.

7.2 Perspectives

7.2.1 Credal classification of sequential data with few training samples

In some applications, like those related with defense, the training samples are quite difficult to obtain for making the target classification (identification). In such case, the training information is insufficient for the learning phase of classifier. We plan to develop a new classifier, in which the sequential test samples will be conditionally included in the training data set. In the classification of sequential data, if the test sample can be clearly classified with high confidence, it will be included in the training data set to make the training information more and more substantial. The selection of the proper test samples will be carefully studied. The class label of the training data selected from the test samples will be not binary (0 or 1), but either characterized by a probabilistic measure (for the classifier working with probability framework) or belief functions (for the classifier working with belief function theory), etc. For this, efficient management techniques will have to be developed to deal with this uncertain class label information for the classification of coming samples. Moreover, we will also globally take into account how to well characterize the uncertainty and imprecision of classification due to the insufficient attribute information.

7.2.2 Classification of incomplete pattern using imprecise probability

In the classification of incomplete pattern, the estimation of the missing values can be represented using an interval rather than several single values. The interval seems more suitable for the characterization of uncertainty of the estimations. Thus, the classification of the object (incomplete pattern) could be also represented using probability (fuzzy membership) interval, which could help to reveal the ambiguity degree of the classification caused by the missing data. At first, the interval of the estimation value for the missing attributes should be determined. Then, the classification of the object will be done. It is crucial to find a method to determine how to find the lower and upper boundary of the probability (fuzzy membership) of the object belonging to each class based on the estimated interval value of missing attributes. Then, the class of the object will be decided according to a strategy based on the probability interval. For example, the object could be assigned to the class with the maximum expected probability value. However, if the intervals associated with different classes are (partially) overlapped, then the object could be committed to the proper meta-class defined by the union of these several classes according to the context.

7.2.3 Credal c-means clustering with some constraints

In the data clustering analysis, there may exist extra knowledge on some constraints on clusters. For example, we may know beforehand that some samples belong to a same cluster. In the CCM clustering method developed in this thesis, no constraint information has been taken into account. Therefore, it will be necessary to include constraints in the objective function to improve the clustering performance of CCM. The main challenging issue will be the mathematical modeling of the constraints in the objective function, and its minimization. For example, one can use the

conflicting beliefs to model the constraint, and the conflicting belief should be as small as possible if the associated objects are in the same cluster. It will be better if one can use linear equations to capture the constraints, since this will be facilitate for the optimization of the objective function. If the modeling equation is nonlinear, it should be as simple as possible to make the optimization tractable by existing non-linear optimization techniques. Once the objective function is determined, it will be optimized to obtain the clustering centers and BBA's of each object belonging to the different classes.

7.2.4 Unified evaluation criteria for performance of credal classifier

The performance evaluation with uncertainty is an important and still open problem. Some new evaluation measures has been proposed by Arnaud, et al, for image classification and segmentation taking into account the uncertain labels [125]. In the credal classifier, the meta-class (i.e. a set of classes) has been introduced to capture the imprecision of classification, and the imprecision rate is defined for the analysis of performance of a credal classifier. In the traditional classifiers, the object is usually committed to a particular class rather than to meta-class. So there is no imprecision rate in the classification results, and the accuracy rate is the common criteria to evaluate the performance of the traditional classifier. We want to develop an efficient unified evaluation criteria taking into account both the accuracy rate and imprecision rate, and this unified criteria can be compared with the common accuracy rate of traditional classifiers. Moreover, the confusion matrix of credal classification will be designed, which will be helpful to compute various quantitative metrics, like recall rate, precision rate, error rate, imprecision rate, and so on. An extended criterion based on the confusion matrix could also be developed to facilitate the comparison of performances of different credal classifiers.

Bibliography

[1] A.-L. Jousselme, P. Maupin, and E. Bossé. Uncertainty in a situation analysis perspective. In *Proc. of 6th int. conf. on information fusion*, Australia, July 2003. 21

[2] A.-L Jousselme, C. Liu, D. Grenier, and E. Bossé. Measuring ambiguity in the evidence theory. *IEEE Trans. Systems, Man and Cybernics-Part A*, 36(5):890–903, 2006. 21, 73, 91

[3] G. Shafer. *A Mathematical Theory of Evidence*. Princeton Univ. Press, 1976. 21, 25, 26, 27, 28, 41, 42, 73, 77, 91, 97

[4] F. Smarandache and J. Dezert. Advances and applications of DSmT for information fusion. *American Research Press, Rehoboth*, Vol. 1-3, 2004–2009. 21, 25, 27, 28, 41, 73, 91

[5] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–243, 1994. 21, 25, 28, 29, 30, 41, 73, 91, 97

[6] P. Smets. The combination of evidence in the transferable belief model. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990. 21, 25, 29, 30, 41, 73, 91, 97

[7] P. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Joural of Approximate Reasoning*, 38(2):133–147, 2005. 21, 25, 29, 30, 41, 73

[8] R. Yager. On the Dempster-Shafer framework and new combination rules. *Information Sciences*, 41(2):93–138, 1987. 21, 27

[9] D. Dubois and H. Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4(4):244–264, 1988. 21, 27

[10] P. Smets. Analyzing the combination of conflicting belief functions. *Information Fusion*, 8(4):387–412, 2007. 21, 27, 28

[11] E. Lefevre, O. Colot, and P. Vannoorenberghe. Belief functions combination and conflict management. *Information Fusion Journal*, 3(2):149–162, 2002. 21, 27

[12] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Systems, Man and Cybernetics*, 25(5):804–813, 1995. 21, 22, 25, 30, 31, 41, 50, 64, 65, 76, 83

[13] T. Denœux and P. Smets. Classification using belief functions: relationship between case-based and model-based approaches. *IEEE Trans. Systems, Man and Cybernetics-Part B*, 36(6):1395–1406, 2006. 21, 29, 30

[14] S. Kanj, F. Abdallah, and T. Denœux. Evidential multi-label classification using the random k-label sets approach. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 21, 30

[15] J. Ma, W. Liu, and P. Miller. An evidential improvement for gender profiling. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 21, 30

[16] A. Antonucci. An interval-valued dissimilarity measure for belief functions based on credal semantics. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 21, 30

[17] P. Lucas. Certainty-factor-like structures in bayesian belief networks. *Knowledge-based systems*, 14(7):327–335, 2001. 21, 30

[18] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997. 21, 30

[19] T. Denœux and L. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 122(3):47–62, 2001. 21, 30

[20] J. Francois, Y. Grandvalet, T. Denœux, and J.M. Roger. Resample and combine: An approach to improving uncertainty representation in evidential pattern classification. *Information Fusion*, (4):75–85, 2003. 21, 30

[21] S. Petit-Renaud and T. Denœux. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning*, 35(1):1–28, 2004. 21, 30

[22] B. Quost, T. Denœux, and M.-H. Masson. Pairwise classifier combination using belief functions. *Pattern Recognition Letters*, 28(5):644–653, 2007. 21, 30

[23] M.-H. Masson and T. Denœux. ECM: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384–1397, 2008. 21, 22, 25, 26, 29, 30, 34, 35, 60, 91, 94, 97

[24] T. Denœux and M.-H. Masson. EVCLUS: EVidential CLUStering of proximity data. *IEEE Trans. Systems, Man and Cybernetics-Part B*, 34(1):95–109, 2004. 21, 26, 30, 34

[25] T. Denœux and M.-H. Masson. Clustering of proximity data using belief functions. In *Proc. of 9th inter. conf. on information processing and management uncertainty in Knowledge-Based Systems*, France, July 2002. 21, 34, 91

[26] M.-H. Masson and T. Denœux. Clustering interval-valued data using belief functions. *Pattern Recognition Letters*, 25(2):163–171, 2004. 21, 34

[27] M.-H. Masson and T. Denœux. RECM: Relational evidential c-means algorithm. *Pattern Recognition Letters*, 30:1015–1026, 2009. 21, 36, 91

[28] V. Antoine, B. Quost, M.-H. Masson, and T. Denœux. CECM: Constrained evidential c-means algorithm. *Computational Statistics and Data Analysis*, 56(4):894–914, 2012. 21, 36, 91

[29] J. Dezert and J.-M. Tacnet. Evidential reasoning for multi-criteria analysis based on DSmT-AHP. In *Proc. of inter. Symposium on Analytic HierarchyNetwork Process*, Italy, 2011. 21

[30] D. Ludmila and S. Pavel. An interpretation of intuitionistic fuzzy sets in terms of evidence theory: Decision making aspect. *Knowledge-based systems*, 23(8):772–782, 2010. 21, 30

[31] L. Dymova and P. Sevastjanov. An interpretation of intuitionistic fuzzy sets in terms of evidence theory: Decision making aspect. *Knowledge-Based Systems*, 23:772–782, 2010. 21

[32] J. Klein and O. Colot. A belief function model for pixel data. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 21

[33] B. Lelandais, I. Gardin, L. Mouchard, P. Vera, and S. Ruan. Using belief function theory to deal with uncertainties and imprecisions in image processing. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 21

[34] A. Znaidia, H.L. Borgne, and C. Hudelot. Belief theory for large-scale multi-label image classification. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 21

BIBLIOGRAPHY

[35] M. Shoyaib, M. Abdullah-Al-Wadud, S.M. Zahid Ishraque, and O. Chae. Facial expression classification based on Dempster-Shafer theory of evidence. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 21

[36] Z.-g. Liu, J. Dezert, G. Mercier, and Q. Pan. Dynamic evidential reasoning for change detection in remote sensing images. *IEEE Trans. Geoscience and Remote Sensing*, 50(5):1955–1967, 2012. 21

[37] Z.-g. Liu, G. Mercier, J. Dezert, and Q. Pan. Change detection in heterogeneous remote sensing images based on multidimensional evidential reasoning. *IEEE Geoscience and Remote Sensing Letters*, 11(1):168–172, 2014. 21

[38] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. Systems, Man and Cybernetics-Part A*, 30(2):131–150, 2000. 21, 30, 76, 83

[39] J. Bezdek. *Pattern Recognition with fuzzy objective function algorithms*. New-York:Plenum Press, 1981. 22, 25, 33, 34, 91, 94

[40] L. Zadeh. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *AI magzine*, 7(2):85–90, 1986. 26, 78

[41] L. Zadeh. *On the validity of Dempster's rule of combination, Memo M 79/24*. Univ. of California,Berkeley, 1979. 26

[42] L. Zadeh. Review of mathematical theory of evidence, by Glenn Shafer. *AI Magazine*, 5(3):81–83, 1984. 26

[43] L. Zadeh. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *AI Magazine*, 7(2):85–90, 1986. 26

[44] P. Wang. A defect in Dempster-Shafer theory. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 560–566. Morgan Kaufmann Publishers, Seattle, WA, 1994. 26, 78

[45] A. Tchamova and J. Dezert. On the behavior of Dempster's rule of combination and the foundations of Dempster-Shafer theory. In *Proc. of 6th IEEE Int. Conf. on Intelligent Systems IS 12*, Sofia, Bulgaria, Sept. 2012. 26, 44, 78

[46] J. Dezert and A. Tchamova. On the validity of Dempster's fusion rule and its interpretation as a generalization of bayesian fusion rule. *International Journal of Intelligent Systems*, 29(3), 2014. 26, 44, 78

[47] D. Mercier, B. Quost, and T. Denœux. Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion*, 9(2):246ÍC258, 2008. 27

[48] C. Murphy. Combining belief functions when evidence conflicts. *Decision Support Systems*, 29(1):1–9, 2000. 27

[49] Y. Deng, W. Shi, Z. Zhu, and Q. Liu. Combining belief functions based on distance of evidence. *Decision Support Systems*, 38(3):489–493, 2004. 27

[50] A. Martin, A.-L. Jousselme, and C. Osswald. Conflict measure for the discounting operation on belief functions. In *Proc. of 11th inter. conf. on information fusion*, Germany, 2008. 27

[51] T. Denœux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, 172, 2008. 28

[52] J. Dezert. Foundations for a new theory of plausible and paradoxical reasoning. *Information Security*, 9:13–57, 2002. 28

[53] J. Dezert and F. Smarandache. A new probabilistic transformation of belief mass assignment. In *Proc. of the 11th inter. conf. on information fusion*, Germany, July 2008. 28, 29

[54] J. Dezert and F. Smarandache. On the generation of hyper-power sets for DSmT. In *Proceedings of Fusion*, Cairns, Australia, 2003. 28

[55] A. Martin and C. Osswald. Human experts fusion for image classification. *Information & Security : An International Journal, Special issue on Fusing Uncertain, Imprecise and Conflicting Information*, 20:122–141, 2006. 29

[56] A. Martin and C. Osswald. A new generalization of the proportional conflict redistribution rule stable in terms of decision. *in Advances and Applications of DSmT for Information Fusion, (Collected Works, Vol. 2), F. Smarandache and J. Dezert (Editors), American Research Press*, 2006. 29

[57] A. Martin and I. Quidu. Decision support with belief functions theory for seabed characterization. In *Proceedings of 11th International Conference on information fusion*, Cologne, Germany, 2008. 29, 30

[58] B. Cobb and P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3), 2006. 29

[59] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 30, 41

[60] N. Barakat and A.P. Bradley. Rule extraction from support vector machines: A review. *Neurocomputing*, 74(1-3), 2010. 30

[61] C.-M. Bishop. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press, 1995. 30, 41, 65, 76

[62] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and regression trees*. Chapman and Hall (Wadsworth, Inc.), New York, 1984. 30, 41, 65

[63] A. Agrawala. *Machine Recognition of Patterns*. IEEE Press, New York, 1977. 30

[64] H. Laanaya, A. Martin, D. Aboutajdine, and A. Khenchaf. Support vector regression of membership functions and belief functions - application for pattern recognition. *Information Fusion*, 11(4), 2010. 30

[65] T. Denœux and M.-H. Masson. Evidential reasoning in large partially ordered sets. application to multi-label classification, ensemble clustering and preference aggregation. *Annals of Operations Research*, 195(1), 2012. 30

[66] A.-L. Jousselme and P. Maupin. An evidential pattern matching approach for vehicle identification. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 30

[67] A. Fiche, A. Martin, J. Cexus, and A. Khenchaf. A comparison between a bayesian approach and a method based on continuous belief functions for pattern recognition. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 30

[68] E. Ramasso, M. Rombaut, and N. Zerhouni. Prognostic by classification of predictions combining similarity-based estimation and belief functions. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 30

BIBLIOGRAPHY

[69] S. Tim, M. Rombaut, and D. Pellerin. Adaptive initialization of a EvKNN classification algorithm. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 30

[70] N. Sutton-Charani, S. Destercke, and T. Denœux. Classification trees based on belief functions. In *Proc. of the 2nd inter. conf. on belief functions*, France, May 2012. 30

[71] T. Denœux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Trans. Knowledge and Data Engineering*, 25(1):119–130, 2013. 30, 34

[72] V. Antoine, B. Quost, M.-H. Masson, and T. Denœux. CEVCLUS: Evidential clustering with instance-level constraints for relational data. *Soft Computing*, 18(7), 2014. 30, 34

[73] J. Han, Y. Deng, and C. Han. Sequential weighted combination for unreliable evidence based on evidence variance. *Decision Support Systems*, 56:387–393, 2013. 30

[74] Z.-g. Liu, J. Dezert, Q. Pan, and G. Mercier. Combination of sources of evidence with different discounting factors based on a new dissimilarity measure. *Decision Support Systems*, 52(1), 2011. 30

[75] S. Hegarat-Mascle, I. Bloch, and D. Vidal-Madjar. Application of dempster shafer evidence theory to unsupervised classification in multisource remote sensing. *IEEE Trans. Geosci. Remote Sensing*, 35(4):1018–1031, 1997. 30

[76] P. Smets. Belief functions: the disjunctive rule of combination and the generalized bayesian theorem. *International Journal of Approximate reasoning*, 9:1–35, 1993. 30, 73

[77] A. Appriou. Uncertain data aggregation in classification and tracking processes. In *B. Bouchon-Meunier, editor, Aggregation and Fusion of imperfect information*, pages 231–260, Physica-Verlag, Heidelberg, 1998. 30

[78] P. Vannoorenbergue and T. Denoeux. Handling uncertain labels in multiclass problems using belief decision trees. In *Proceedings of 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 1919–1926, France, 2002. 30

[79] L. Zouhal and T. Denœux. Generalizing the evidence theoretic K-NN rule to fuzzy pattern recognition. In *Proc. of 2nd Int. ICSC symposium on fuzzy logic and applications,*, Zurich, Switzerland, 1997. 30

[80] H. Altincay. Ensembling evidential k-nearest neighbor classifiers through multi-modal perturbation. *Applied Soft Computing*, 7(3):1072–1083, 2007. 30

[81] R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1), 1973. 33

[82] W. Zhanga, D. Zhao, and X. Wang. Agglomerative clustering via maximum incremental path integral. *Pattern Recognition*, 46(11), 2013. 33

[83] Y. Ma, H. Derksen, H. Wei, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(9), 2007. 33

[84] S. Lloyd. Least squares quantization in PCM. *IEEE. Trans. Information Theory*, 28(2):129–137, 1982. 33, 34

[85] R. Krishnapuram and J. Keller. The possibilistic c-means algorithm: insights and recommendations. *IEEE Trans. Fuzzy Systems*, 4(3):385–393, 1996. 33, 34

[86] R. Krishnapuram and J. Keller. A possibilistic approach to clustering. *IEEE Trans. Fuzzy Systems*, 1(2):98–110, 1993. 33

[87] J. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3), 1973. 33

[88] R. Dave. Clustering of relational data containing noise and outliers. *Pattern Recognition Letters*, (12):657–664, 1991. 34

[89] S. Sen and R. Dave. Clustering of relational data containing noise and outliers. In *Proc. of 7th Inter. conf. on fuzzy systems*, Alaska, USA, May 1998. 34

[90] R. Krishnapuram and R. Dave. Robust clustering methods: A unified view. *IEEE Trans. Fuzzy Systems*, 5(2):270–293, 1997. 34

[91] Z.-g. Liu, J. Dezert, G. Mercier, and Q. Pan. Belief c-means: An extension of fuzzy c-means algorithm in belief functions framework. *Pattern Recognition Letters*, 33(3), 2012. 36

[92] P. Garcia-Laencina, J. Sancho-Gomez, and A. Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Comput Appl.*, 19:263–282, 2010. 37, 73

[93] R. Little and D. Rubin. *Statistical Analysis with Missing Data.* John Wiley &Sons, New York, 1987. 37, 73

[94] A. Farhangfar, L. Kurgan, and J. Dy. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41:3692–3705, 2008. 37, 73

[95] A. Frank and A. Asuncion. *UCI Machine Learning Repository.* University of California, School of Information and Computer Science, Irvine, CA, USA, 2010, http://archive.ics.uci.edu/ml/. 37, 55, 65, 70, 88, 104

[96] R. Little and D. Rubin. *Statistical analysis with missing data (2nd Edition).* Wiley, 2002. 37, 38, 73

[97] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society*, B39:1–38, 1977. 37

[98] O. Troyanskaya, M. Cantor, O. Alter, G. Sherlock, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. 38, 73

[99] G. Batista and M. Monard. A study of k-nearest neighbour as an imputation method. *in Proc. of Second Inter. Conf. on Hybrid Intelligent Systems (IOS Press, v. 87)*, pages 251–260, 2002. 38, 73

[100] P. Chan and O. Dunn. The treatment of missing values in discriminant analysis. *Journal of the American Statistical Association*, 6:473–477, 1972. 38, 73

[101] J. Schafer. *Analysis of incomplete multivariate data.* Chapman & Hall, Florida, 1997. 38, 73

[102] D. Rubin. *Multiple imputation for nonresponse in surveys.* Wiley, 1987. 38, 73

[103] F. Fessant and S. Midenet. Self-organizing map for data estimation and correction in surveys. *Neural Comput. Appl.*, 10(4), 2002. 38, 39, 73

[104] K. Jian, H. Chen, and S. Yuan. Classification for incomplete data using classifier ensembles. In *International Conference on Neural Networks and Brain.*, pages 559–563, Beijing, China, 2005. 38

BIBLIOGRAPHY

[105] P. Juszczak and R. Duin. Combining one-class classifiers to classify missing data. In *Roli F. et al (eds) Mult. Classif. Syst., Lect Notes Comput. Sci.*, pages 92–101. Springer, 2004. 38

[106] J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1), 1986. 38

[107] R. Hathaway and J. Bezdek. Fuzzy c-means clustering of incomplete data. *IEEE Trans. Syst Man Cybern B: Cybern*, 31(5), 2001. 38

[108] H. Ichihashi and K. Honda. Fuzzy c-means classifier for incomplete data sets with outliers and missing values. In *Proc. of Intl Conf Comput Intell Modell Control Autom*, pages 457–464, Washington, DC, USA, 2005. 38

[109] C. Lim, J. Leong, and M. Kuan. A hybrid neural network system for pattern classification tasks with missing features. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(4), 2005. 38

[110] K. Pelckmans, J. Brabanter, J. Suykens, and D. Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5–6), 2005. 38

[111] T. Kohonen. *Self-organizing maps.* 3rd edn. Springer, 2006. 39

[112] P. Sharpe and Solly R. Dealing with missing values in neural network-based diagnostic systems. *Neural Computing and Applications*, 3(2), 1995. 39

[113] E. Fix and J. Hodges. Discriminatory analysis. nonparametric discrimination: consistency properties. In *Techn. Rep. 4, Project number 21-49-004, USAF School of Aviation Medicine,USA*, pages 261–279, 1951. 41, 59, 76

[114] S. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Trans. Systems, Man and Cybernics*, 6:325–327, 1976. 41, 59

[115] J. Keller and J. Givens. A fuzzy k-nearest neighbor algorithm. *IEEE Trans. Systems, Man and Cybernetics*, 15(4):580–585, 1985. 41, 59

[116] F. Smarandache, J. Dezert, and J.-M. Tacnet. Fusion of sources of evidence with different importances and reliabilities. In *Proceedings of the 13rd international conference on information fusion*, Scotland, UK, 2010. 42

[117] J. Dezert, Z. Liu, and G. Mercier. Edge detection in color images based on DSmT. In *Proc. of 14th int. conf. on information fusion*, Chicago, USA, July 2011. 43, 64

[118] L. Zouhal and T. Denœux. An evidence-theoretic k-nn rule with parameter optimization. *IEEE Trans. Systems, Man and Cybernetics - Part C*, 28(2):263–271, 1998. 50, 65, 83

[119] S. Geisser. *Predictive Inference: An introduction.* New York, NY: Chapman and Hall, 1993. 56

[120] T. Coleman. *Large sparse numerical optimization.* New York: Springer-Verlag, 1984. 62

[121] D. Li, J. Deogun, W. Spaulding, and B. Shuart. Towards missing data imputation: a study of fuzzy k-means clustering method. In *Proceedings of the 4th international conference of rough sets and current trends in computing (RSCTC04)*, pages 573–579. Springer-Verlag Berlin Heidelberg, Uppsala, Sweden, 2004. 73, 83

[122] J. Luengo, J. Saez, and F. Herrera. Missing data imputation for fuzzy rule-based classification systems. *Soft Computing*, 16(5), 2012. 73, 83

[123] J. Lemmer. Confidence factors, empiricism and the Dempster-Shafer theory of evidence. In *Proceedings of the 1st conference on Uncertainty in Artificial Intelligence*, pages 160–176, Los Angeles, CA, 1985. 78

[124] G. Provan. The validity of Dempster-Shafer belief functions. *International Journal of Approximate Reasoning*, 6(3), 1992. 78

[125] A. Martin and H. Laanaya. Evaluation for uncertain image classification and segmentation. *Pattern Recognition*, 39(11), 2006. 113

Publications

List of contributions related to this thesis

1. Zhunga Liu, Quan Pan, Grégoire Mercier, Jean Dezert, *A new incomplete pattern classification method based on evidential reasoning*, IEEE Transactions on Cybernetics, DOI: 10.1109/TCYB.2014.2332037, 2014.

2. Zhunga Liu, Quan Pan, Jean Dezert, Grégoire Mercier, *Credal classification rule for uncertain data based on belief functions*, Pattern Recognition, Vol.47(7): 2532–2541, 2014.

3. Zhunga Liu, Quan Pan, Jean Dezert, *Classification of uncertain and imprecise data based on evidence theory*, Neurocomputing, Vol. 133:459–470, 2014.

4. Zhunga Liu, Quan Pan, Jean Dezert, Gregoire Mercier, *Credal c-means clustering method based on belief functions*, Knowledge-based systems, DOI: 10.1016/j.knosys.2014.11.013, 2014.

5. Zhun-ga Liu, Quan Pan, Jean Dezert, *A new belief-based K-nearest neighbor classification method*, Pattern Recognition, Vol. 46(3): 834–844, 2013.

6. Zhun-ga Liu, Quan Pan, Jean Dezert, *Evidential classifier for imprecise data based on belief functions*, Knowledge-based systems, Vol.52:246–257, 2013.

7. Zhun-ga Liu, Jean Dezert, Grégoire Mercier, Quan Pan, *Belief C-Means: An Extension of Fuzzy C-Means Algorithm in Belief Functions Framework*, Pattern Recognition Letters, Vol.33(3): 291–300, 2012.

8. Zhun-ga Liu, Quan Pan, Jean Dezert, Grégoire Mercier, Yong Liu, *Fuzzy-belief K-nearest neighbor classifier for uncertain data*, Proceedings of the 17th International Conference on Information Fusion (FUSION 2014), Salamanca, Spain, Jul. 2014.

9. Zhun-ga Liu, Quan Pan, Jean Dezert, Grégoire Mercier, *Credal classification of uncertain data using belief functions*, Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2013), Manchester, U.K, Oct. 2013.

10. Zhunga Liu, Jean Dezert, Quan Pan, Yongmei Cheng, *A new evidential c-means clustering method*, Proceedings of the 15th International Conference on Information Fusion (FUSION 2012), Singapore, July 2012.

List of extra contributions

1. Zhunga Liu, Grégoire Mercier, Jean Dezert, Quan Pan, *Change Detection in Heterogeneous Remote Sensing Images Based on Multidimensional Evidential Reasoning*, IEEE Geoscience Remote Sensing Letters, Vol. 11(1): 168–172, 2014.

2. Zhun-ga Liu, Jean Dezert, Grégoire Mercier, Quan Pan, *Dynamic Evidential Reasoning For Change Detection In Remote Sensing Images*, IEEE Transactions on Geoscience and Remote Sensing, Vol. 50(5): 1955–1967, 2012.

3. Zhun-ga Liu, Jean Dezert, Quan Pan, Grégoire Mercier, *Combination of sources of evidence with different discounting factors based on a new dissimilarity measure*, Decision Support Systems, Vol.52: 133–141, 2011.

4. Zhun-ga Liu, Jean Dezert, Grégoire Mercier, Quan Pan, Yong-mei Cheng, *Change detection from remote sensing images based on evidential reasoning*, Proceedings of the 14th International Conference on Information Fusion (FUSION 2011), Chicago, USA, Jul. 2011.

5. Jean Dezert, Deqiang Han, Zhunga Liu, Jean-Marc Tacnet, *Hierarchical Proportional Redistribution for bba Approximation*, Proceedings of the 2nd International Conference on Belief Functions, Compiegne, France, May 2012.

6. Jean Dezert, Zhun-ga Liu, and Grégoire Mercier, *Edge detection in color images based on DSmT*, 2011 Proceedings of the 14th International Conference on Information Fusion (FUSION 2011), Chicago, USA, Jul. 2011.

Technopole Brest-Iroise - CS 83818
29238 Brest Cedex 3
France
Tél : + 33 (0)2 29 00 11 11
www.telecom-bretagne.eu

Résumé

Cette thèse s'intéresse à la classification crédibiliste de données fondée sur la théorie des fonctions de masse. Lorsque des échantillons labellisés sont disponibles en nombre suffisant, une classification supervisée peut être appliquée. Certains classifieurs ont été développés sur la base de la théorie de Demspter-Shafer, et l'ignorance totale est caractérisée en utilisant une pondération des fonctions de masse. Or, l'information imprécise partielle n'est pas prise en compte dans ces méthodes, et la classification est souvent partiellement imprécise entre un très petit nombre de classes.

Dans cette thèse, nous avons étudié la classification crédibiliste de données incertaines sur la base des fonctions de croyance, et deux classifieurs crédibilistes ont été proposés. La classification crédibiliste permet à un objet d'appartenir à des classes simples mais aussi à des méta-classes définies par l'union de plusieurs classes simples. Ces méta-classes modélisent l'imprécision partielle de classification et réduisent le taux d'erreur de classification. Une méthode de classification crédibiliste appelée c × K plus proches voisins crédibilistes a été introduite. Lorsque chaque classe peut être représentée par son centre de classe, nous avons également proposé une règle simple de classification crédibiliste (CCR), qui calcule directement la masse de croyance de l'échantillon appartenant à chaque classe et une méta-classe avec une faible complexité calculatoire. En outre, une méthode de classification crédibiliste de données incomplètes a été également développée, et elle est capable de modéliser de telles informations incertaines et imprécises provenant de valeurs manquantes.

Mots-clés : Théorie crédibiliste, Données manquantes, Classification

Abstract

Modeling and managing uncertainty in the classification problem remains an important and interesting research topic. Credal classification of uncertain data based on belief function theory has been studied in this thesis, and it allows the object to belong not only to the single classes, but also to any set of classes (called meta-class) with different masses of belief. The credal classification is then of interest to explore the imprecision of classes.

Classification methods can be mainly identified by supervised, unsupervised and semi-supervised ones according to the availability of training information. We focus on the supervised and unsupervised classifications. When there are a lot of training samples available in the classification, two credal classifiers for uncertain data are proposed for dealing with different cases. A belief c × K neighbors (BCKN) classifier has been proposed based on belief function theory. In BCKN, the query object is classified according to its K nearest neighbors in each class, and c × K basic belief assignments (BBA's) are determined according to the distances between the object and these neighbors, and the global fusion of them is used for the credal classification of object. When each class of data can be represented by the prototype vector, a simple credal classification rule (CCR) has been developed using belief functions. Moreover, the missing attribute data is often encountered in classification problem. The different estimations of the missing values can lead to distinct classification results sometimes, and this yields high imprecision and uncertainty of classification due to the lack of information in the missing values.

Keywords : Credal theory, Missing data, Classification

TELECOM Bretagne

n° d'ordre : 2014telb0333

Télécom Bretagne

Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3

Tél : + 33(0) 29 00 11 11 - Fax : + 33(0) 29 00 10 00