18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Classification of Incomplete Patterns Based on the Fusion of Belief Functions

Zhun-ga Liu*, Quan Pan*, Jean Dezert†, Arnaud Martin‡ and GrégoireMercier§
*Northwestern Polytechnical University, Xi'an, China.
Email: liuzhunga@nwpu.edu.cn, quanpan@nwpu.edu.cn
†ONERA - The French Aerospace Lab, F-91761 Palaiseau, France.
Email: jean.dezert@onera.fr
‡IRISA, University of Rennes 1, 22300 Lannion, France.
Email: Arnaud.Martin@univ-rennes1.fr
§Telecom Bretagne, CNRS UMR 6285 Lab-STICC/CID, 29200 Brest, France.
Email: Gregoire.Mercier@telecom-bretagne.eu

Abstract—**The influence of the missing values in the classification of incomplete pattern mainly depends on the context. In this paper, we present a fast classification method for incomplete pattern based on the fusion of belief functions where the missing values are selectively (adaptively) estimated. At first, it is assumed that the missing information is not crucial for the classification, and the object (incomplete pattern) is classified based only on the available attribute values. However, if the object cannot be clearly classified, it implies that the missing values play an important role to obtain an accurate classification. In this case, the missing values will be imputed based on the K-nearest neighbor (K-NN) and self-organizing map (SOM) techniques, and the edited pattern with the imputation is then classified. The (original or edited) pattern is respectively classified according to each training class, and the classification results represented by basic belief assignments (BBA's) are fused with proper combination rules for making the credal classification. The object is allowed to belong with different masses of belief to the specific classes and meta-classes (i.e. disjunctions of several single classes). This credal classification captures well the uncertainty and imprecision of classification, and reduces effectively the rate of misclassifications thanks to the introduction of meta-classes. The effectiveness of the proposed method with respect to other classical methods is demonstrated based on several experiments using artificial and real data sets.**

Keywords: information fusion, combination rule, belief functions, classification, incomplete pattern.

I. Introduction

In many practical classification problems, some attributes of object can be missing for various reasons (e.g. the failure of the sensors, etc). So it is crucial to develop efficient techniques to classify as best as possible the objects with missing attribute values (incomplete pattern), and the search for a solution of this problem remains an important research topic in the community [1], [2]. Many classification approaches have been proposed to deal with the incomplete patterns [1]. The simplest method consists in removing (ignoring) directly the patterns with missing values, and the classifier is designed only for the complete patterns. This method is acceptable when the incomplete data set is only a very small subset (e.g. less than 5%) of the whole data set. A widely adopted method is to fill the missing values with proper estimations [3], and then to classify the the edited patterns. There have been different works devoted to the imputation (estimation) of missing data. For example, the imputation can be done either by the statistical methods, e.g. mean imputation [4], regress imputation [2], etc, or by machine learning methods, e.g. K-nearest neighbors (K-NN) imputation [5], Fuzzy c-means (FCM) imputation [6], [7], etc. Some model-based techniques have also been developed for dealing with incomplete patterns [8]. The probability density function (PDF) of the training data (complete and incomplete cases) is estimated at first, and then the object is classified using bayesian reasoning. Other classifiers [9] have also been proposed to directly handle incomplete pattern without imputing the missing values. All these methods attempt to classify the object into a particular class with maximal probability or likelihood measure. However, the estimation of missing values is in general quite uncertain, and the different imputations of missing values can yield very different classification results, which prevent us to correctly commit the object into a particular class.

Belief function theory (BFT), also called Dempster-Shafer theory (DST) [10] and its extension [11], [12] offer a mathematical framework for modeling uncertainty and imprecise information [13]. BFT has already been applied successfully for object classification [14], [15], [17]–[19], clustering [20]–[23] and multi-source information fusion [24], etc. Some classifiers for the complete pattern based on DST have been developed by Denœux and his collaborators to come up with the evidential K-nearest neighbors [14], evidential neural network [19], etc. The extra ignorance element represented by the disjunction of all the elements in the whole frame of discernment is introduced in these classifiers to capture the totally ignorant information. However, the partial imprecision, which is very important in the classification, is not well characterized. That is why we have proposed new credal classifiers in [15]–[17], [22]. Our new classifiers take into account all the possible meta-classes (i.e. the particular disjunctions of several singleton classes) to model the partial imprecise information thanks to belief functions. The credal classification allows the objects to belong (with different masses of belief) not only to the singleton classes, but also to any set of classes corresponding to the meta-classes.

In our recent research works, a prototype-based credal classification (PCC) [25] method for the incomplete patterns has been introduced to well capture the imprecision of classification caused by the missing values. The object hard to correctly classify are committed to a suitable meta-class by PCC, which captures well the imprecision of classification caused by the missing values and also reduces the misclassification errors. In PCC, the missing values in all the incomplete patterns are imputed using the prototype of each class, and the edited pattern with each imputation is respectively classified by a standard classifier (used for the classification of complete pattern). With PCC, one obtains c pieces of classification results for one incomplete pattern in a c class problem, and the global fusion of the c results is used for the credal classification. Unfortunately, PCC classifier is computationally greedy and time-consuming, and the method of imputation of the missing values based on the prototype of each class is not so precise and accurate. That is why we propose a new innovative and more effective method for credal classification of incomplete pattern with adaptive imputation of missing values, and this method can be called Credal Classification with Adaptive Imputation (CCAI) for short.

The pattern to classify usually consists of multiple attributes. Sometimes, the class of the pattern can be precisely determined using only a part (a subset) of the available attributes, which means that the other attributes are redundant and in fact unnecessary for the classification. In the classification of incomplete pattern with missing values, one can attempt at first to classify the object only using the known attributes value. If a specific classification result is obtained, it very likely means that the missing values are not very necessary for the classification, and we directly take the decision on the class of the object based on this result. However, if we the object cannot be clearly classified with the available information, it means that the missing information included in the missing attribute values is probably very crucial for making the classification. In this case, we propose a sophisticated classification strategy for the edited pattern with proper imputation of missing values obtained using K-NN and self-organizing map (SOM) techniques [26].

The information fusion technique is adopted in the classification of original incomplete pattern (without imputation of missing values) or the edited pattern (with imputation of missing values) to obtain the good results. One can respectively get the simple classification result represented by a simple basic belief assignment (BBA) according to each training class. The global fusion (ensemble) of these multiple BBA's with a proper combination rule, i.e. Dempster-Shafer (DS) rule or a new rule inspired by Dubois Prade (DP) rule depending on the actual case, is then used to determine the class of the object.

This paper is organized as follows. The basics of belief function theory is briefly recalled in section II. The new credal classification method for incomplete patterns is presented in the section III, and the proposed method is then tested and evaluated in section IV compared with several other classical methods. It is concluded in the final.

II. Basis of belief function theory

The Belief Function Theory (BFT) is also known as Dempster-Shafer Theory (DST), or the Mathematical Theory of Evidence [10]–[12]. Let us consider a frame of discernment consisting of c exclusive and exhaustive hypotheses (classes) denoted by $\Omega = \{\omega_i, i = 1, 2, \ldots, c\}$. The power-set of Ω denoted 2^Ω is the set of all the subsets of Ω, empty set included. In the classification problem, the singleton element (e.g. ω_i) represents a specific class. In this work, the disjunction (union) of several singleton elements is called a *meta-class* which characterizes the partial ignorance of classification. In BFT, the basic belief assignment (BBA) is a function $m(.)$ from 2^Ω to $[0, 1]$ satisfying $m(\emptyset) = 0$ and the normalization condition $\sum_{A\in 2^\Omega} m(A) = 1$. The subsets A of Ω such that $m(A) > 0$ are called the focal elements of the belief mass $m(.)$.

The credal classification (or partitioning) [20], [21] is defined as n-tuple $M = (\mathbf{m_1}, \cdots, \mathbf{m_n})$ of BBA's, where $\mathbf{m}_i$ is the basic belief assignment of the object $\mathbf{x}_i \in X$, $i = 1, \ldots, n$ associated with the different elements in the power-set 2^Θ. The credal classification can well model the imprecise and uncertain information thanks to the introduction of meta-class. For combining multiple sources of evidence represented by a set of BBA's, the well-known Dempster's rule [10] is still widely used. We denote it by DS (standing for Dempster-Shafer) because Dempster's rule has been widely promoted by Shafer in [10]. The combination of two BBA's $m_1(.)$ and $m_2(.)$ over 2^Ω is done with DS rule of combination defined by $m_{DS}(\emptyset) = 0$ and for $A \neq \emptyset, B, C \in 2^\Omega$ by

$$m_{DS}(A) = \frac{\sum\limits_{B\cap C=A} m_1(B)m_2(C)}{1 - \sum\limits_{B\cap C=\emptyset} m_1(B)m_2(C)} \tag{1}$$

DS rule is commutative and associative, and makes a compromise between the specificity and complexity for the combination of BBA's. However, DS rule produces unreasonable results in high conflicting cases, and as well as in some special low conflicting cases [27]. Many alternative rules have been proposed to overcome the limitations of DS rule, e.g. Dubois-Prade (DP) rule [28] and Proportional Conflict Redistributions (PCR) rules [29]. Our method is inspired by DP rule [28] defined by $m_{DP}(\emptyset) = 0$ and for $A \neq \emptyset, B, C \in 2^\Theta$ by

$$m_{DP}(A) = \sum_{B\cap C=A} m_1(B)m_2(C) + \sum_{\substack{B\cap C=\emptyset \\ B\cup C=A}} m_1(B)m_2(C) \tag{2}$$

In DP rule, the partial conflicting beliefs are all transferred to the union of the elements (i.e. meta-class) involved in the partial conflict.

III. Credal classification of incomplete pattern

Our new method consists of two main steps. In the first step, the object (incomplete pattern) is directly classified according to the known attribute values only, and the missing values are ignored. If one can get a specific classification result, the classification procedure is done because the available attribute information is sufficient for making the classification. But if the class of the object cannot be clearly identified in the first step, it means that the unavailable information included in the missing values is likely crucial for the classification. In this case, one has to enter in the second step of the method to classify the object with a proper imputation of missing values. In the classification procedure, the original or edited pattern will be respectively classified according to each class of training data. The global fusion of these classification results, which can be considered as multiple sources of evidence represented by BBA's, is then used for the credal classification of the object. The new method is referred as Credal Classification with Adaptive Imputation of missing values denoted by CCAI for conciseness.

A. Step 1: Direct classification of incomplete pattern

Let us consider a set of test patterns (samples) $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ to be classified based on a set of labeled training patterns $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_s\}$ over the frame of discernment $\Omega = \{\omega_1, \ldots, \omega_c\}$. In this work, we focus on the classification of incomplete pattern in which some attribute values are absent. So we consider all the test patterns (e.g. $\mathbf{x}_i, i = 1, \ldots, n$) with several missing values. The training data set Y may also have incomplete patterns in some applications. However, if the incomplete patterns take a very small amount say less than 5% in the training data set, they can be ignored in the classification. If the percentage of incomplete patterns is big, the missing values must usually be estimated at first, and the classifier will be trained using the edited (complete) patterns. In the real applications, one can also just choose the complete labeled patterns to include in the training data set when the training information is sufficient. So for simplicity and convenience, we consider that the labeled samples (e.g. $\mathbf{y}_j, j = 1, \ldots, s$) of the training set Y are all complete patterns in the sequel.

In the first step of classification, the incomplete pattern say $\mathbf{x}_i$ will be respectively classified according to each training class by a normal classifier (for dealing with the complete pattern) at first, and all the missing values are ignored here. In this work, we adopt a very simple classification method for the convenience of computation, and $\mathbf{x}_i$ is directly classified based on the distance to the prototype of each class.

The prototype of each class $\{\mathbf{o}_1, \ldots, \mathbf{o}_c\}$ corresponding to $\{\omega_1, \ldots, \omega_c\}$ is given by the arithmetic average vector of the training patterns in the same class. Mathematically, the prototype is computed for $g = 1, \ldots, c$ by

$$\mathbf{o}_g = \frac{1}{N_g} \sum_{\mathbf{y}_j \in \omega_g} \mathbf{y}_j \tag{3}$$

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

In a c-class problem, one can get c pieces of simple classification result for $\mathbf{x}_i$ according to each class of training data, and each result is represented by a simple BBA's including two focal elements, i.e. the singleton class and the ignorant class (Ω) to characterize the full ignorance. The belief of $\mathbf{x}_i$ belonging to class w_g is computed based on the distance between $\mathbf{x}_i$ and the corresponding prototype $\mathbf{o}_g$. Mahalanobis distance is adopted here to deal with the anisotropic class, and the missing values are ignored in the calculation of this distance. The other mass of belief is assigned to the ignorant class Ω. Therefore, the BBA's construction is done by

$$\begin{cases} m_i^{\mathbf{o}_g}(w_g) = e^{-\eta d_{ig}} \\ m_i^{\mathbf{o}_g}(\Omega) = 1 - e^{-\eta d_{ig}} \end{cases} \tag{4}$$

with

$$d_{ig} = \sqrt{\frac{1}{p} \sum_{j=1}^{p} \left(\frac{x_{ij} - o_{gj}}{\delta_{gj}} \right)^2} \tag{5}$$

and

$$\delta_{gj} = \sqrt{\frac{1}{N_g} \sum_{\mathbf{y}_i \in \omega_g} (y_{ij} - o_{gj})^2} \tag{6}$$

where x_{ij} is value of $\mathbf{x}_i$ in j-th dimension, and y_{ij} is value of $\mathbf{y}_i$ in j-th dimension. p is the number of available attribute values in the object $\mathbf{x}_i$. The coefficient $1/p$ is necessary to normalize the distance value because each test sample can have a different number of missing values. δ_{gj} is the average distance of all training samples in class ω_g to the prototype o_g in j-th dimension. N_g is the number of training samples in ω_g. η is a tuning parameter, and the bigger η generally yields smaller mass of belief on the specific class w_g.

Obviously, the smaller distance measure, the bigger mass of belief on the singleton class. This particular structure of BBA's indicates that we can just confirm the degree of the object $\mathbf{x}_i$ associated with the specific class w_g only according to training data in w_g. The other mass of belief reflects the level of belief one has on full ignorance, and it is committed to the ignorant class Ω. Similarly, one calculates c independent BBA's $m_i^{\mathbf{o}_g}(w_g), g = 1, \ldots, c$ based on the different training classes.

Before combining these c BBA's, we examine whether a specific classification result can be derived from these c BBA's. This is done as follows: if it holds that $m_i^{\mathbf{o}_{1st}}(w_{1st}) = argmax_g(m_i^{\mathbf{o}_g}(w_g))$, then the object will be considered to belong very likely to the class w_{1st}, which obtains the biggest mass of belief in the c BBA's. The class with the second biggest mass of belief is denoted w_{2nd}.

The distinguishability degree $\chi_i \in (0, 1]$ of an object $\mathbf{x}_i$ associated with different classes is defined by:

$$\chi_i = \frac{m_i^{\mathbf{o}_{2nd}}(w_{2nd})}{m_i^{\mathbf{o}_{max}}(w_{max})} \tag{7}$$

Let ϵ be a chosen small positive distinguishability threshold value in $(0, 1]$. If the condition $\chi_i \leq \epsilon$ is satisfied, it means that all the classes involved in the computation of χ_i can be clearly distinguished of $\mathbf{x}_i$. In this case, it is very likely to obtain a specific classification result from the fusion of the c BBA's. The condition $\chi_i \leq \epsilon$ also indicates that the available attribute information is sufficient for making the classification of the object, and the imputation of the missing values is not necessary. If $\chi_i \leq \epsilon$ condition holds, he c BBA's are directly combined with DS rule (1) to obtain the final classification results of the object because DS rule usually produces specific combination result with acceptable computation burden in the low conflicting case. In such case, the meta-class is not included in the fusion result, because these different classes are considered distinguishable based on the condition of distinguishability. Moreover, the mass of belief of the full ignorance class Ω, which represents the noisy data (outliers), can be proportionally redistributed to other singleton classes for more specific results if one knows a priori that the noisy data is not involved.

If the distinguishability condition $\chi_i \leq \epsilon$ is not satisfied, it means that the classes w_{1st} and w_{2nd} cannot be clearly distinguished for the object with respect to the chosen threshold value ϵ, indicating that missing attribute values play almost surely a crucial role in the classification. In this case, the missing values must be properly imputed to recover the unavailable attribute information before entering the classification procedure. This is the Step 2 of our method which is explained in the next subsection.

B. Step 2: Classification of incomplete pattern with imputation of missing values

1) Multiple estimation of missing values: In the estimation of the missing attribute values, there exist various methods. Particularly, the K-NN imputation method generally provides good performance. However, the main drawback of KNN method is its big computational burden, since one needs to calculate the distances of the object with all the training samples. Inspired by [30], we propose to use the Self Organized Map (SOM) technique [26], [30] to reduce the computational complexity. SOM can be applied in each class of training data, and then $M \times N$ weighting vectors will be obtained after the optimization procedure. These optimized weighting vectors allow to characterize well the topological features of the whole class, and they will be used to represent the corresponding data class. The number of the weighting vectors is usually small (e.g. 5×6). So the K nearest neighbors of the test pattern associated with these weighting vectors in the SOM can be easily found with low computational complexity[1]. The selected weighting vector no. k in the class w_g, $g = 1, \ldots, c$ is denoted $\sigma_k^{w_g}$, for $k = 1, \ldots, K$.

In each class, the K selected close weighting vectors provide different contributions (weights) in the estimation of missing values. The weight $p_{ik}^{w_g}$ of each vector is defined based on the distance between the object $\mathbf{x}_i$ and weighting vector $\sigma_k^{w_g}$ as follows

$$p_{ik}^{w_g} = e^{(-\lambda d_{ik}^{w_g})} \tag{8}$$

with

$$\lambda = \frac{cNM(cNM-1)}{2\sum\limits_{i,j} d(\sigma_i, \sigma_j)} \tag{9}$$

where $d_{ik}^{w_g}$ is the Euclidean distance between $\mathbf{x}_i$ and the neighbor $\mathbf{o}_k^{w_g}$ ignoring the missing values, and $\frac{1}{\lambda}$ is the average distance between each pair of weighting vectors produced by SOM in all the classes; c is the number of classes; $M \times N$ is the number of weighting vectors obtained by SOM in each class; and $d(\sigma_i, \sigma_j)$ is the Euclidean distance between any two weighting vectors σ_i and σ_j.

The weighted mean value $\hat{\mathbf{y}}_i^{w_g}$ of the selected K weighting vectors in class training class w_g will be used for the imputation of missing values. It is calculated by

$$\hat{\mathbf{y}}_i^{w_g} = (\sum_{k=1}^{K} p_{ik}^{w_g} \sigma_k^{w_g}) / (\sum_{k=1}^{K} p_{ik}^{w_g}) \tag{10}$$

The missing values in $\mathbf{x}_i$ will be filled by the values of $\hat{\mathbf{y}}_i^{w_g}$ in the same dimensions. By doing this, we get the edited pattern $\mathbf{x}_i^{w_g}$ according to the training class w_g. Then $\mathbf{x}_i^{w_g}$ will be simply classified only based on the training data in w_g as similarly done in the direct classification of incomplete pattern using eq. (4) of Step 1 for convenience[2].

The classification of $\mathbf{x}_i$ with the estimation of missing values is also respectively done based on the other training classes according to this procedure. For a c-class problem, there are c training classes, and therefore one can get c pieces of classification results with respect to one object.

2) Ensemble classifier for credal classification: These c pieces of results obtained by each class of training data in a c-class problem are considered with different weights, since the estimations of the missing values according to different classes have different reliabilities. The weighting factor of the classification result associated with the class w_g can be defined by the sum of the weights of the K selected SOM weighting vectors for the contributions to the missing values imputation in w_g, which is given by

$$\rho_i^{w_g} = \sum_{k=1}^{K} p_{ik}^{w_g} \tag{11}$$

The result with the biggest weighting factor $\rho_i^{w_{max}}$ is considered as the most reliable, because one assumes that the object must belong to one of the labeled classes (i.e. w_g, $g = 1, \ldots, c$). So the biggest weighting factor will be normalized as one. The other relative weighting factors are defined by:

$$\hat{\alpha}_i^{w_g} = \frac{\rho_i^{w_g}}{\rho_i^{w_{max}}} \tag{12}$$

If the condition[3] $\hat{\alpha}_i^{w_g} < \epsilon$ is satisfied, the corresponding estimation of the missing values and the classification result

[1] The training of SOM using the labeled patterns becomes time consuming when the number of labeled patterns is big, but fortunately it can be done off-line. In our experiments, the running time performance shown in the results does not include the computational time spent for the off-line procedures.

[2] Of course, some other sophisticated classifiers can also be applied here according to the selection of user, but the choice of classifier is not the main purpose of this work.

[3] The threshold ϵ is the same as in section III-A, because it is also used to measure the distinguishability degree here.

are not very reliable. Very likely, the object does not belong to this class. It is implicitly assumed that the object can belong to only one class in reality. If this result whose relative weighting factor is very small (w.r.t. ϵ) is still considered useful, it will be (more or less) harmful for the final classification of the object. So if the condition $\hat{\alpha}_i^{w_g} < \epsilon$ holds, then the relative weighting factor is set to zero. More precisely, we will take

$$\alpha_i^{w_g} = \begin{cases} 0, & \text{if} \quad \hat{\alpha}_i^{w_g} < \epsilon \\ \frac{\rho_i^{w_g}}{\rho_i^{w_{max}}}, & \text{otherwise.} \end{cases} \tag{13}$$

After the estimation of weighting (discounting) factors $\alpha_i^{w_g}$, the c classification results (the BBA's $m_i^{\mathbf{o}_g}(.)$) are classically discounted [10] by

$$\begin{cases} \hat{m}_i^{\mathbf{o}_g}(w_g) = \alpha_i^{w_g} m_i^{\mathbf{o}_g}(w_g) \\ \hat{m}_i^{\mathbf{o}_g}(\Omega) = 1 - \alpha_i^{w_g} + \alpha_i^{w_g} m_i^{\mathbf{o}_g}(\Omega) \end{cases} \tag{14}$$

These discounted BBA's will be globally combined to get the credal classification result. If $\alpha_i^{w_g} = 0$, one gets $\hat{m}_i^{\mathbf{o}_g}(\Omega) = 1$, and this fully ignorant (vacuous) BBA plays a neutral role in the global fusion process for the final classification of the object.

Although we have done our best to estimate the missing values, the estimation can be quite imprecise when the estimations are obtained from different class with the similar weighting factors, and the different estimations probably lead to distinct classification results. In such case, we prefer to cautiously keep (rather to ignore) the uncertainty, and maintain the uncertainty in the classification result. Such uncertainty can be well reflected by the conflict of these classification results represented by the BBA's. DS rule is not suitable here, because all the conflicting beliefs are distributed to other focal elements. A particular combination rule inspired by DP rule is introduced here to fuse these BBA's according to the current context. In our new rule, the partial conflicting beliefs are prudently transferred to the proper meta-class to reveal the imprecision degree of the classification caused by the missing values. This new rule of combination is defined by:

$$\begin{cases} m_i(w_g) = \hat{m}_i^{\mathbf{o}_g}(w_g) \prod_{j \neq g} \hat{m}_i^{\mathbf{o}_j}(\Omega) \\ m_i(A) = \prod_{\bigcup_j w_j = A} \hat{m}_i^{\mathbf{o}_j}(w_j) \prod_{k \neq j} \hat{m}_i^{\mathbf{o}_k}(\Omega) \end{cases} \tag{15}$$

The global fusion formula (15) consists of two parts. In the first part, we use the conjunctive combination to commit the mass of belief to the specific (singleton) class, whereas the disjunctive combination is used to transfer the conflicting beliefs to the proper meta-class in the second part.

The test pattern can be classified according to the fusion results, and the object is considered belonging to the class (singleton class or meta-class) with the maximum mass of belief. This is called hard credal classification. If one object is classified to a particular class, it means that this object has been correctly classified with the proper imputation of missing values. If one object is committed to a meta-class (e.g. $A \cup B$), it means that we just know that this object belongs to one of the specific classes (e.g. A or B) included in the meta-class, but we cannot specify which one. This case can happen when the missing values are essential for the accurate classification of this object, but the missing values cannot be estimated very well according to the context, and different estimations will induce the classification of the object into distinct classes (e.g. A or B).

With traditional classifiers, the missing values in each object are usually estimated before making the classification. In our CCAI approach, many objects can be directly classified based on the distances to each class prototype, and the imputation of missing values is ignored according to the context. So the computation complexity of CCAI is generally relatively low with respect to other methods like KNNI, PCC, etc.

Guideline for tuning of the parameters ϵ and η: η in eq. (4) is associated with the calculation of mass of belief on the specific class, and the bigger η value will lead to smaller mass of belief committed to the specific class. We advise to take $\eta \in [0.5, 0.8]$, and the value $\eta = 0.7$ can be taken as the default value. The parameter ϵ is the threshold for changing the classification strategy. It is also used in Eq. (13) for the calculation of the discounting factor. The bigger ϵ will makes fewer objects committed to the meta-classes (corresponding to the low imprecision of classification), but it increases the risk of misclassification error. ϵ should be tuned according to the compromise one can accept between the misclassification error and imprecision.

IV. Experiments

Two experiments with artificial and real data sets have been used to test the performance of this new CCAI method compared with the K-NN imputation (KNNI) method [5], FCM imputation (FCMI) method [6], [7] and our previous credal classification PCC method [25]. The evidential neural network classifier (ENN) [19] is adopted here to classify the edited pattern with the estimated values in PCC, KNNI and FCMI, since ENN produce generally good results in the classification. The parameters of ENN can be automatically optimized as explained in [19]. In the applications of PCC, the tuning parameter ϵ can be tuned according to the imprecision rate one can accept. In CCAI, a small number of the nodes in the 2-dimensional grid of SOM is given by $M \times N = 3 \times 4$, and we take the value of $K = N = 4$ in K-NN for the imputation of missing values. This seems to provide good performance in the sequel experiments. In order to show the ability of CCAI and PCC to deal with the meta-classes, the hard credal classification is applied, and the class of each object is decided according to the criterion of the maximal mass of belief.

In our simulations, the misclassification is declared (counted) for one object truly originated from w_i if it is classified into A with $w_i \cap A = \emptyset$. If $w_i \cap A \neq \emptyset$ and $A \neq w_i$ then it will be considered as an imprecise classification. The error rate denoted by Re is calculated by $Re = N_e/T$, where N_e is number of misclassification errors, and T is the number of objects under test. The imprecision rate denoted by Ri_j is calculated by $Ri_j = Ni_j/T$, where Ni_j is number of objects committed to the meta-classes with the cardinality value j. In our 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.

A. Experiment 1 (artificial data set)

In the first experiment, we show the interest of credal classification based on belief functions with respect to the traditional classification working with probability framework. A 3-class data set $\Omega = \{\omega_1, \omega_2, \omega_3\}$ obtained from three 2-D uniform distributions is considered here. Each class has 200 training samples and 200 test samples, and there are 600 training samples and 600 test samples in total as shown in Fig.1.

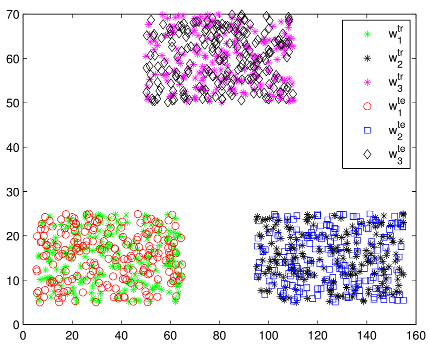


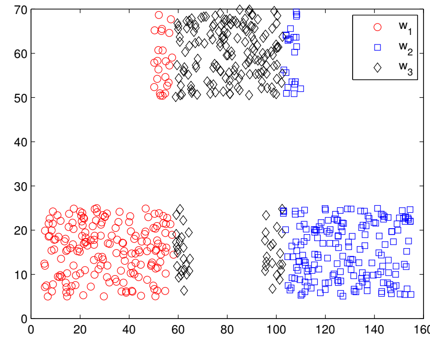
Figure 1. Training data and test data.

The uniform distributions of the three classes are characterized by the following interval bounds:

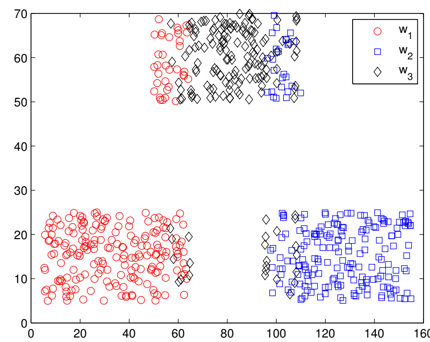
	x-label interval	y-label interval
w_1	(5, 65)	(5, 25)
w_2	(95, 155)	(5, 25)
w_3	(50, 110)	(50, 70)

The values in the second dimension corresponding to y-coordinate of test samples are all missing. So test samples are classified according to the only one available value in the first dimension corresponding to x-coordinate. A particular value of $K = 9$ is selected in the classifier K-NN imputation method[4]. The classification results of the test objects by different methods are given in Fig. 2 (a)–(c). 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 %) are specified in the caption of each subfigure.

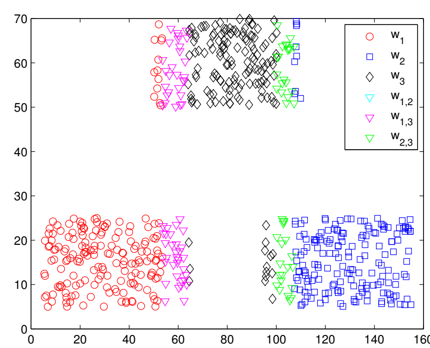
[4]In fact, the choice of K ranking from 7 to 15 does not affect seriously the results.



(a). Classification result by FCMI
($Re = 14.67, time = 0.0469s$).



(b). Classification result by KNNI
($Re = 14.17, time = 7.9531s$).



(c). Classification result by CCAI
($Re = 5.83, Ri_2 = 16.83, time = 0.0469s$).

Figure 2. Classification results of a 3-class artificial data set by different methods.

Because the y value in the test sample is missing, the class w_3 appears partially overlapped with the classes w_1 and w_2 on their margins according to the value of x-coordinate as shown in Fig. 1. The missing value of the samples in the overlapped parts can be filled by quite different estimations obtained from different classes with the almost same reliabilities. For example, the estimation of the missing values of the objects in the right margin of w_1 and the left margin of w_3 can be obtained according to the training class w_1 or w_3. The edited pattern with the estimation from w_1 will be classified into class w_1, whereas it will be committed to class w_3 if the estimation is drawn from w_3. It is similar to the test samples in the left margin of w_2 and the right margin of w_3. This indicates that the missing value play a crucial rule in the classification of these objects, but unfortunately the estimation of these involved missing values are quite uncertain according to context. So these objects are prudently classified into the proper meta-class (e.g. $w_1 \cup w_3$ and $w_2 \cup w_3$) by CCAI. The CCAI results indicate that these objects belong to one of the specific classes included in the meta-classes, but these specific classes cannot be clearly distinguished by the object based only on the available values. If one wants to get more precise and accurate classification results, one needs to request for additional resources for gathering more useful information. The other objects in the left margin of w_1, right margin of w_2 and middle of w_3 can be correctly classified based on the only known value in x-coordinate, and it is not necessary to estimate the missing value for the classification of these objects in CCAI. However, all the test samples are classified into specific classes by the traditional methods KNNI and FCMI, and this causes many errors due to the limitation of probability framework. Thus, CCAI produces less error rate than KNNI and FCMI thanks to the use of meta-classes. Meanwhile, the computational time of CCAI is similar to that of FCMI, and is much shorter than KNNI because of the introduction of SOM technique in the estimation of missing values. It shows that the computational complexity of CCAI is relatively low. This simple example shows the interest and the potential of the credal classification obtained with CCAI method.

B. Experiment 2 (real data set)

Four well known real data sets (Breast cancer, Iris, Seeds and Wine data sets) available from UCI Machine Learning Repository [32] are used in this experiment to evaluate the performance of CCAI with respect to KNNI, FCMI and PCC. ENN is also used here as standard classifier. The basic information of these four real data sets is given in Table I.

The cross validation is performed on all the data sets, and we use the simplest 2-fold cross validation[5] 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. Each test sample has n missing (unknown) values, and they are missing completely at random in every dimension. The average error rate Re and imprecision rate Ri (for PCC and CCAI) of the different methods are given in Table II. Particularly, the reported classification result of KNNI is the average with K value ranging from 5 to 15.

Table I
BASIC INFORMATION OF THE USED DATA SETS.

name	classes	attributes	instances
Breast (B)	2	9	699
Iris (I)	3	4	150
Seeds (S)	3	7	210
Wine (W)	3	13	178

Table II
CLASSIFICATION RESULTS FOR DIFFERENT REAL DATA SETS (IN %).

data n	FCMI	KNNI	PCC	CCAI
	Re	Re	$\{Re, Ri_2\}$	$\{Re, Ri_2\}$
B 3	3.81	3.95	{3.81, 2.34}	{3.66, 0}
B 6	7.32	8.20	{5.42,1.32}	{4.83, 1.61}
B 7	11.42	11.54	{10.10, 2.64}	{9.00, 0.66}
I 1	7.33	4.89	{5.33, 2.67}	{4.00, 1.33}
I 2	14.11	11.33	{8.67,4.00}	{8.00, 4.67}
I 3	17.33	18.44	{12.67, 9.33}	{11.33, 12}
S 2	15.24	11.19	{9.52, 4.76 }	{9.52, 0}
S 4	17.14	11.98	{10.48, 4.29}	{10.00, 0.48}
S 6	20.95	25.71	{16.19, 14.76}	{16.19, 13.81}
W 3	26.97	26.97	{26.97, 1.69}	{6.74, 1.12}
W 7	33.24	30.43	{29.78, 2.25}	{7.30, 3.93}
W 11	33.43	30.90	{30.34, 2.81}	{12.36, 3.93}

One can see that the credal classification of PCC and CCAI always produce the lower error rate than the traditional FCMI and KNNI methods, since some objects that cannot be correctly classified using only the available attribute values have been properly committed to the meta-classes, which can well reveal the imprecision of classification. In CCAI, some objects with the imputation of missing values are still classified into the meta-class. It indicates that these missing values play a crucial role in the classification, but the estimation of these missing values is no very good. In other words, the missing values can be filled with the similar reliabilities by different estimated data, which lead to distinct classification results. So we have to cautiously assign them to the meta-class to reduce the risk of misclassification. Compared with our previous method PCC, this new method CCAI generally provide better performance with lower error rate and imprecision rate, and it is mainly because more accurate estimation method (i.e. $SOM + KNN$) for missing values is adopted in CCAI. This third experiment using real data sets for different applications shows the effectiveness and interest of this new CCAI method with respect to other methods.

V. CONCLUSION

A fast credal classification method with adaptive imputation of missing values (called CCAI) for dealing with incomplete pattern has been presented. In step 1 of CCAI method, some objects (incomplete pattern) are directly classified ignoring the missing values if the specific classification result can be obtained, which effectively reduces the computation complexity because it avoids the imputation of the missing values. However, if the available information is not sufficient to achieve a specific classification of the object, we estimate (recover) the missing values before entering the classification

[5]More precisely, the samples in each class 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.

procedure in the second step. The SOM and K-NN approaches are applied to make the estimation of missing attributes with a good compromise between the estimation accuracy and computation burden. Information fusion technique is employed to combine the multiple simple classification results respectively obtained from each training class for the final credal classification of object. The credal classification in this work allows the object to belong to different singleton classes and meta-class with different masses of belief. Once the object is committed to a meta-class (e.g. $A \cup B$), it means that the missing values cannot be accurately recovered according to the context, and the estimation is not very good. Different estimations will lead the object to distinct classes (e.g. A or B) involved in the meta-class. So some other sources of information will be required to achieve more precise classification of the object if necessary. Two experiments have been applied to test the performance of CCAI method with artificial and real data sets. The results show that the credal classification is able to well capture the imprecision of classification and effectively reduces the misclassification errors as well.

Acknowledgements

This work has been partially supported by National Natural Science Foundation of China (Nos. 61135001, 61403310, 61473230), the Fundamental Research Funds for the Central Universities (No. 3102014JCQ01067) and Natural Science Foundation of Shaanxi (No. 2015JQ6265).

References

[1] P. Garcia-Laencina, J. Sancho-Gomez, A. Figueiras-Vidal, *Pattern classification with missing data: a review*, Neural Comput Appl. Vol.19:263–282, 2010.

[2] R.J. Little, D.B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, New York, 1987 (Second edition published in 2002).

[3] A. Farhangfar, L. Kurgan, J. Dy, *Impact of imputation of missing values on classification error for discrete data*, Pattern Recognition Vol.41:3692–3705, 2008.

[4] D.J. Mundfrom, A. Whitcomb, *Imputing missing values: The effect on the accuracy of classification*, Multiple Linear Regression Viewpoints. Vol. 25(1):13–19, 1998.

[5] G. Batista, M.C. Monard, *A Study of K-Nearest Neighbour as an Imputation Method*, in Proc. of Second International Conference on Hybrid Intelligent Systems (IOS Press, v. 87), pp. 251–260, 2002.

[6] J. Luengo, J.A. Saez, F.Herrera, *Missing data imputation for fuzzy rule-based classification systems*, Soft Computing, Vol.16(5):863–881, 2012.

[7] D. Li, J.Deogun, W. Spaulding, B. Shuart, *Towards missing data imputation: a study of fuzzy k-means clustering method*, In: 4th international conference of rough sets and current trends in computing (RSCTC04), pp 573–579, 2004.

[8] Z. Ghahramani, M.I. Jordan, *Supervised learning from incomplete data via an EM approach*, In: Cowan JD et al. (Eds) Adv. Neural Inf. Process., Morgan Kaufmann Publishers Inc., Vol.6:120–127, 1994.

[9] K. Pelckmans, J.D. Brabanter, J.A.K. Suykens, B.D. Moor, *Handling missing values in support vector machine classifiers*, Neural Networks, Vol. 18(5-6):684–692, 2005.

[10] G. Shafer, *A mathematical theory of evidence*, Princeton Univ. Press, 1976.

[11] F. Smarandache, J. Dezert (Editors), *Advances and applications of DSmT for information fusion*, American Research Press, Rehoboth, Vol. 1-4, 2004-2015. http://fs.gallup.unm.edu/DSmT.htm

[12] P. Smets, *The combination of evidence in the transferable belief model*, IEEE Trans. on Pattern Anal. and Mach. Intell., Vol. 12(5):447–458,1990.

[13] A.-L. Jousselme, C. Liu, D. Grenier, E. Bossé, *Measuring ambiguity in the evidence theory*, IEEE Trans. on SMC, Part A: 36(5):890–903, Sept. 2006.

[14] T. Denœux, *A k-nearest neighbor classification rule based on Dempster-Shafer Theory*, IEEE Trans. on Systems, Man and Cybernetics, Vol. 25(5):804–813, 1995.

[15] Z.-g. Liu, Q. Pan, J. Dezert, *A new belief-based K-nearest neighbor classification method*, Pattern Recognition, Vol.46(3): 834–844, 2013.

[16] Z.-g. Liu, Q. Pan, J. Dezert, G. Mercier, *Fuzzy-belief K-nearest neighbor classifier for uncertain data*, in Proc. of 17th Int. Conference on Information Fusion (Fusion 2014), Spain, July 7-10, 2014.

[17] Z.-g. Liu, Q. Pan, J. Dezert, G. Mercier, *Credal classification rule for uncertain data based on belief functions*, Pattern Recognition, Vol. 47(7): 2532–2541,2014.

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

[19] T. Denœux, *A neural network classifier based on Dempster-Shafer theory*, IEEE Trans. Systems, Man and Cybernetics: Part A, Vol.30(2): 131–150, 2000.

[20] M.-H. Masson, T. Denœux, *ECM: An evidential version of the fuzzy c-means algorithm*, Pattern Recognition, Vol.41(4):1384–1397, 2008.

[21] T. Denœux, M.-H. Masson, *EVCLUS: EVidential CLUStering of proximity data*, IEEE Trans. on Systems, Man and Cybernetics: Part B, Vol.34(1):95–109, 2004.

[22] Z.-g. Liu, Q. Pan, J. Dezert, G. Mercier, *Credal c-means clustering method based on belief functions*, Knowledge-based systems, Vol.74: 119-132, 2015.

[23] T. Denœux, *Maximum likelihood estimation from uncertain data in the belief function framework*, IEEE Transactions on Knowledge and Data Engineering, Vol. 25(1):119–130, 2013.

[24] Z.-g. Liu, J. Dezert, Q. Pan, G. Mercier, *Combination of sources of evidence with different discounting factors based on a new dissimilarity measure*, Decision Support Systems, Vol.52:133–141, 2011.

[25] Z.-g. Liu, Q. Pan, G. Mercier, J. Dezert, *A new incomplete pattern classification method based on evidential reasoning*, IEEE Trans. Cybernetics, Vol.45(4) :635-646, 2015.

[26] T. Kohonen, *The Self-Organizing Map*, Proceedings of the IEEE, Vol.78(9):1464–1480, 1990.

[27] J. Dezert, 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, Vol. 29(3):223–252, 2014.

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

[29] F. Smarandache, J. Dezert, *On the consistency of PCR6 with the averaging rule and its application to probability estimation*, in Proc. of 16th Int. Conference on Information Fusion (Fusion 2013), Istanbul, Turkey, July 9-12, 2013.

[30] I. Hammami, J. Dezert, G. Mercier, A. Hamouda, On the estimation of mass functions using Self Organizing Maps, in Proc. of Belief 2014 Conf. Oxford, UK, Sept. 26–29, 2014.

[31] S. Geisser, *Predictive inference: an introduction*, New York, NY: Chapman and Hall, 1993.

[32] A. Frank, 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