Neurocomputing 133 (2014) 459–470

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Classification of uncertain and imprecise data based on evidence theory

Zhun-ga Liu [a,*], Quan Pan [a], Jean Dezert [b]

[a] School of Automation, Northwestern Polytechnical University, Xi'an, China
[b] ONERA – The French Aerospace Lab, F-91761 Palaiseau, France

ARTICLE INFO

Article history:
Received 1 March 2013
Received in revised form
3 October 2013
Accepted 6 December 2013
Communicated by N.T. Nguyen
Available online 11 January 2014

Keywords:
Evidence theory
Credal classification
Belief functions
K-NN
Data classification

ABSTRACT

In this paper, we present a new belief $c \times K$ neighbor (BCKN) classifier based on evidence theory for data classification when the available attribute information appears insufficient to correctly classify objects in specific classes. In BCKN, the query object is classified according to its K nearest neighbors in each class, and $c \times K$ neighbors are involved in the BCKN approach (c being the number of classes). BCKN works with the credal classification introduced in the belief function framework. 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 an 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 credal classification is interesting to explore the imprecision of class, and it can also provide a deeper insight into the data structure. The results of several experiments are given and analyzed to illustrate the potential of this new BCKN approach.

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. The well known case-based classification methods, like K-nearest neighbor (K-NN) [1,2], decision trees [3,4], support vector machine (SVM) [5–7], artificial neural networks (ANNs) [8], 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).

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 probabilistic framework cannot well model and manage the imprecision of data. In fact, 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 [9,10].

The belief functions [11,13,12,14], introduced originally in the mathematical theory of evidence theory by Shafer in 1976, also known as Dempster-Shafer Theory (DST), offer a rigorous mathematical formalism to model uncertain and imprecise information produced by a source of evidence. This formalism has been already applied in many fields, including classification [15–20], clustering [21–23] and information fusion [24,25]. A recent concept called credal partition [22] has been introduced by Denœux and Masson based on the belief function for data clustering (unsupervised classification). The credal partition is an extension of the probabilistic partition based on a frame of discernment $\Omega = \{w_1, \ldots, w_c\}$ that allows the samples to belong not only to the specific classes (e.g. w_i) but also to a set of classes called a meta-class (e.g. $w_i \cup w_j$) with different masses of belief. The credal partition provides a deeper insight into the data structure as already reported in [21]. An evidential version of fuzzy c-means (FCM) clustering method [21] inspired by FCM [26] and Dave's Noise-Clustering algorithm [27] has been developed using credal partition to deal with imprecise data and outliers.

* Corresponding author.
E-mail addresses: liuzhunga@gmail.com (Z.-g. Liu), jean.dezert@onera.fr (J. Dezert).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.009

Some data classifiers have already been developed based on belief functions in the past. For instance, Smets [28] and Appriou [29] have proposed the model-based classifier based on the Generalized Bayes Theorem (GBT) [17]. GBT is an extension of Bayes theorem in Smets transferable belief model (TBM) [12,13]. Some case-based classifiers have also been proposed by Denœux [15,16]. Particularly, the evidential version of K-nearest neighbors (EK-NNs) method has been proposed in [15] based on DST, for working only with the specific classes and the extra ignorant class defined by the union of all the specific classes. An ensemble technique for the combination of evidential K-NN classifier based on DST has been proposed in [30] to improve the accuracy. A neural network classifier has also been developed in [16] under the belief functions framework that allows one extra ignorant class as possible output of this classifier.

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. In this work, we propose a new case-based belief classifier working with credal classification corresponding to the credal partition in data clustering, 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. This is the innovation of this paper.

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 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* (BCKNs) 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 detail in Section 3. Such a credal classification (a classification based on soft assignments represented by belief functions) is very interesting in many applications, especially 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 a 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, the 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 paper is organized as follows. The background on the belief functions is briefly introduced in the next section. The details of BCKN are presented in Section 3. Several experiments are given in Section 4 to show how BCKN performs with respect to other classical methods. Concluding remarks are given in the last section of this paper.

2. Background on belief functions

The belief functions have been introduced in 1976 by Shafer in his mathematical theory of evidence, known also as Dempster–Shafer theory (DST) [11,13,12,14] 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. 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 $\theta_i \cup \theta_j = \{\theta_i, \theta_j\}$ is interpreted as the proposition "the truth value of unknown solution of the problem under concern is either in θ_i or in θ_j". So that Ω represents the full ignorance (uncertainty).

Shafer [11] considers the subsets as propositions in the case we are concerned with the true value of some quantity w 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 $\mathcal{P}_w(A)$ is thus in one-to-one correspondence with the subset A of Ω. Such a 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 $c(\cdot)$. 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 $m(\cdot)$ from 2^Ω to $[0, 1]$ satisfying

$$\begin{cases} \sum_{A \in 2^\Omega} m(A) = 1 \\ m(\emptyset) = 0 \end{cases} \tag{1}$$

The subsets A of Ω such that $m(A) > 0$ are called the *focal elements* of $m(\cdot)$, and the set of all its focal elements is called the *core* of $m(\cdot)$. 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

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

ambiguity) degree between the class A and B for the classification of the object.

A *Credal partition* [22,21] 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}.

The belief function $Bel(\cdot)$ and the plausibility function $Pl(\cdot)$ [11] 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 $m(\cdot)$ by

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

$$Pl(A)=\sum_{B\in 2^{\Omega}|A\cap B\neq\varnothing} m(B) \tag{3}$$

Shafer [11] 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 the DS (Dempster–Shafer) rule for short in the sequel. Mathematically, the DS rule of combination of two bba's $m_1(\cdot)$ and $m_2(\cdot)$ defined on 2^{Ω} is defined by $m_{DS}(\varnothing)=0$ and for $A\neq\varnothing\in 2^{\Omega}$ by

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

$$=\frac{\sum_{B,C\in 2^{\Omega}|B\cap C=A} m_1(B)m_2(C)}{\sum_{B,C\in 2^{\Omega}|B\cap C\neq\varnothing} m_1(B)m_2(C)} \tag{4}$$

In the DS rule, the total conflicting belief mass $\sum_{B,C\in 2^{\Omega}|B\cap C=\varnothing} m_1(B)m_2(C)$ is redistributed back to all the focal elements through the normalization. In the combination of high conflicting sources of evidence, the DS rule will produce very unreasonable results, and it is better to not use it in such cases [31–33]. It has also been proved recently that the DS rule also suffers of a very serious flaw even in low conflicting cases for very specific belief structures [33,34].

In order to palliate the drawbacks of the DS rule, many alternative rules of combinations have been proposed [14]. Among them, Dubois and Prade have proposed another rule, denoted the DP rule in [35] that will be used in BCKN. The basic idea behind the DP rule comprises to transfer each partial conflicting mass on the union of the elements involved in the partial conflict. Mathematically, the DP rule is defined by $m_{DP}(\varnothing)=0$ and for $A\in 2^{\Omega}\backslash\{\varnothing\}$ by

$$m_{DP}(A)=\sum_{B,C\in 2^{\Omega}|B\cap C=A} m_1(B)m_2(C)+\sum_{\substack{B,C\in 2^{\Omega}\\ B\cap C=\varnothing\\ B\cup C=A}} m_1(B)m_2(C) \tag{5}$$

The pignistic (betting) probability transformation $BetP(\cdot)$ introduced by Smets in his transferable belief model [12,13] is commonly used to approximate any bba $m(\cdot)$ into a probability measure for hard decision-making support. $BetP(A)$ is defined for $A\in 2^{\Omega}\backslash\{\varnothing\}$ by

$$BetP(A)=\sum_{B\in 2^{\Omega},A\subseteq B}\frac{|A\cap B|}{|B|}m(B) \tag{6}$$

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 the 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 $BetP(\cdot)$ for decision-making support if the computational burden is a strong constraint (like in real-time military classification applications).

3. Belief $C\times K$ neighbors classifier

3.1. Principle of BCKN

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 [11,36] which is out of the scope of this paper. For the classification of an input sample (the object), we choose K nearest neighbors 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 the 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. 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 the BCKN approach are described in detail in the next subsections.

3.2. 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 $\Omega=\{w_0,w_1,\ldots,w_c\}$ 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 4.

The K nearest neighbors 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

$$d_{si}=\|\mathbf{y}_s-\mathbf{x}_i\| \tag{7}$$

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.[2] The bba's about $\mathbf{y}_s$ are defined for $i=1,\ldots,c\times K$ and $\mathbf{x}_i\in w_g$ by

$$\begin{cases} m_{si}(w_g)=e^{-\gamma d_{si}} \\ m_{si}(\Omega)=1-e^{-\gamma d_{si}} \end{cases} \tag{8}$$

where $\gamma>0$ in Eq. (8) 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

[2] There exist other methods of construction of bba's [37,38], but they need more tuning parameters and have a higher computation complexity which makes them not easy to use.

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

$$\gamma = \frac{1}{\overline{d}} \tag{9}$$

with

$$\overline{d} = \frac{1}{cn_i(n_i-1)} \sum_{i=1}^{c} \sum_{j=1}^{n_i} \sum_{l=1}^{n_i} \| \mathbf{x}_j - \mathbf{x}_l \| \tag{10}$$

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. (8), 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 a 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.

3.3. 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:

- Step 1 (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 detail in the next subsections.

3.3.1. Step 1: The sub-combination of bba′s in the same group

The DS rule is usually considered as acceptable in most situations where the bba′s are not too conflicting. However, the DS rule has several serious limitations as reported [33]. It is not appropriate to use the 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:

Lemma 1. *Let us consider a group of K bba′s defined on 2^{Ω} having the following structure $m_{si}(w_g) = \varepsilon_i$ and $m_{si}(\Omega) = 1 - \varepsilon_i$, where ε_i are small positive values, $i = 1, \ldots, K$. Let us denote $\varepsilon = \min[\varepsilon_1, \ldots, \varepsilon_K]$. The combined mass of belief obtained by the DS fusion of the K bba′s $m_{si}(\cdot)$ will be focused on w_g because one always gets $m_{DS}(w_g) > 0.5 > m_{DS}(\Omega)$ as soon as $K > -\ln 2/\ln(1-\varepsilon)$.*

Proof. In applying the DS rule for combining the K bba′s, we get

$$\begin{cases} m_{DS}(w_g) = 1 - \prod_{i=1}^{K} (1-\varepsilon_i) \\ m_{DS}(\Omega) = \prod_{i=1}^{K} (1-\varepsilon_i) \end{cases} \tag{11}$$

Whence 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 $\varepsilon = \min[\varepsilon_1, \ldots, \varepsilon_K]$, one always has

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

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

The Lemma 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 $\varepsilon = \varepsilon_i$, $i = 1, \ldots, K$ is a small value (say $\varepsilon = 0.2$) 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 a 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 $\mathbf{y}_s$ does not 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 the 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

$$\begin{cases} m_s^g(w_g) = \frac{1}{K} \sum_{i=1}^{K} m_{si}(w_g) \\ m_s^g(\Omega) = \frac{1}{K} \sum_{i=1}^{K} m_{si}(\Omega) \end{cases} \tag{12}$$

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.

3.3.2. 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 $w_i \cap w_j = \varnothing$) 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 the DS rule.

If all the conflicting beliefs are kept and committed to the associated meta-classes (as done classically in the 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 specificities 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 $m_s^2(w_2)=0.5+\varepsilon$, $m_s^2(\Omega)=0.5-\varepsilon$. If $\varepsilon=0$, 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 the DP rule. If $\varepsilon>0$, the meta-class $w_1\cup w_2$ will get the most belief after the combination of the two bba's by the 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 a 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 $(m_s^{\max}(w_{\max})-m_s^i(w_i))\le t$, $i=1,\ldots,c$ (t being a given threshold), then the class w_i will also be 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 $\psi_{\max}\triangleq\{w_i|m_s^{\max}(w_{\max})-m_s^i(w_i)\le t\}$ with big mass of belief. It says that $\mathbf{y}_s$ likely belongs to one of the 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 1. Let us consider $\Omega=\{w_0,w_1,w_2,w_3\}$. 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[3] $\Psi=\{w_1\cup w_2,w_1\cup w_3,w_2\cup w_3\}$. 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 (4) and DP rule (5). It is mathematically defined by the formulas (13) and (14). The sub-combination results associated with the $\mathbf{y}_s$ and different classes can be fused sequentially by

$$m_s^{1,g}(A)=\begin{cases}\sum\limits_{B_1,B_2\in 2^\Omega|B_1\cap B_2=A} m_s^{1,g-1}(B_1)m_s^g(B_2) & \text{for } A\notin\Psi\\ \sum\limits_{B_1,B_2\in 2^\Omega|B_1\cup B_2=A} m_s^{1,g-1}(B_1)m_s^g(B_2) & \text{for } A\in\Psi\end{cases}\tag{13}$$

$m_s^{1,g}(\cdot)$ is the unnormalized combination results of $m_s^1(\cdot),\ldots,m_s^g(\cdot),g=1,\ldots,c$. By convention, one takes $m_s^{1,1}(\cdot)=m_s^1(\cdot)$. It is worth to note that this combination rule is close to the DP rule (5) in its principle, but the summation of the combined bba is not one (i.e. here one can have $\sum m_s^{1,g}(\cdot)\le 1$) if some partial conflicting beliefs are not preserved. In the 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 associative because of the very particular structure of bba's $m_s^i(\cdot)$ satisfying the BCKN model. This will be shown in Example 2.

The unnormalized fusion results obtained by (13) 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

$$m_s(A)=\frac{m_s^{1,c}(A)}{\sum_j m_s^{1,c}(B_j)}\tag{14}$$

where $m_s^{1,c}(\cdot)$ is the unnormalized bba obtained after combining sequentially all the bba's $m_s^g(\cdot)$ for $g=1,2,\ldots,c$ with the formula (13).

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. (14). If none meta-class is selected, the global fusion rule reduces to the 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 the DP rule, and all the partial conflicting belief masses are transferred onto these meta-classes. This global fusion rule can be considered as a compromise between the DS rule and the DP rule, since we only select a subset of all possible meta-classes on which the conflicting belief masses will be redistributed. The selection of the admissible meta-classes used in the BCKN method depends on the current context. The global fusion results can be used for the classification making support. The belief function $Bel(\cdot)$, the plausibility function $Pl(\cdot)$ and pignistic probability $BetP(\cdot)$ introduced in Section 2 can be used for final hard (binary) assignment of the objects to a specific class when it is really necessary. Such a 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.

The pseudo-code of the BCKN is given in Table 1 for convenience.

3.4. 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

Table 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 (8);
	Combination of bba's from neighbors with the same label in (12);
	Selection of meta-classes according to sub-combination results;
	Global fusion of these sub-combination results in (13) and (14);
	Credal classification of $\mathbf{y}_s$ based on the global fusion results.
	end

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

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 also be 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 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[4] 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 us suppose in this example that these bba's are given by

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

Thus $m_{s1}(\cdot)$ and $m_{s2}(\cdot)$ strongly support w_1, $m_{s3}(\cdot)$ and $m_{s4}(\cdot)$ support w_2, and $m_{s5}(\cdot)$ and $m_{s6}(\cdot)$ support moderately the class w_3. The combination results of the bba's in the same group using the averaging rule (Eq. (12)) gives

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

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[5] structure of the bba's, the unnormalized combination rule (13) 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. (13)

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

If A is a selected meta-class, say $A=w_1\cup w_2$, then one gets from Eq. (13)

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

Such a 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 (13) of Step 2 is

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

These masses are then normalized according to (14), and we get

$$m_s(w_1)=0.1495,\quad m_s(w_2)=0.1055,\quad m_s(w_3)=0.0113$$
$$m_s(w_1\cup w_2)=\mathbf{0.5980},\quad m_s(w_1\cup w_3)=0.0641$$
$$m_s(w_2\cup w_3)=0.0452,\quad m_s(\Omega)=0.0264$$

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 is 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.

3.5. 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 [15], 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.

[4] 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 of the training samples belongs to w_0.

[5] The focal elements of each bba are nested.

Example 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-combinations 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.

4. 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 (ANNs) and Support Vector Machine (SVM) methods. In the following three experiments (i.e. Experiment 1–3), the tuning of parameters in these different methods is introduced as follows. The different methods have been programmed and tested with Matlab™ software. The parameters of EK-NN were automatically optimized using the method introduced in [39]. 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 σ=0.125. The tuning threshold t in BCKN is optimized using the training data. The optimized value corresponds to a suitable compromise between the error rate and the imprecision rate (for example, the 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 offline.

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 to 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 R_e is calculated by $R_e = N_e/T$, where N_e is the number of objects wrongly classified, and T is the total number of the objects tested. The imprecision rate denoted by R_{l_j} is calculated by $R_{l_j} = N_{l_j}/T$, where N_{l_j} is the number of objects committed to the meta-classes with the cardinality value j.

4.1. Experiment 1

This experiment consists of two particular tests (numerical simulations) and shows how BCKN works and its difference with

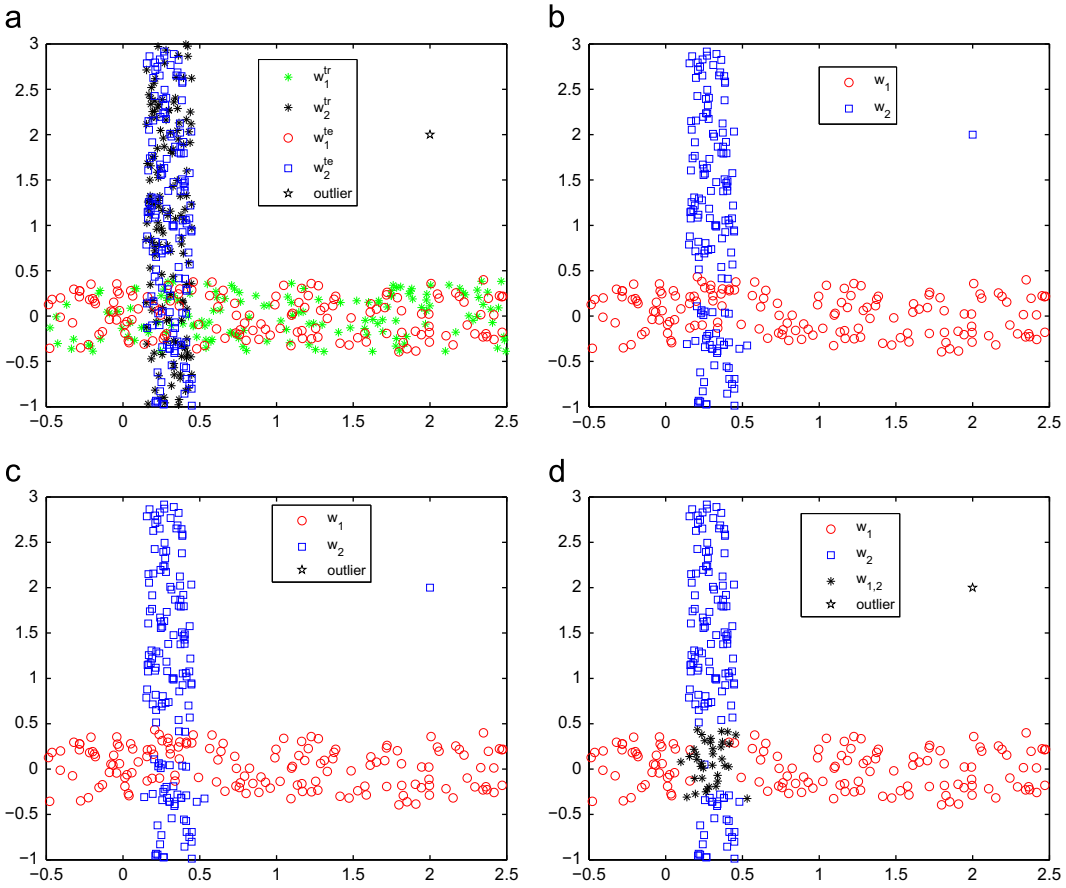

Fig. 1. Classification results by *K*-NN, EK-NN and BCKN. (a) Training data and test data, (b) classification result by *K*-NN, (c) classification result by EK-NN and (d) classification result by BCKN (t=0.005).

respect to EK-NN and K-NN methods. The tuning of parameters of these methods have been presented in the beginning of Section 4.

4.1.1. 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 in Fig. 1a. 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 produces 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 in Fig. 1b–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 in Fig. 1a, 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 in Fig. 1b and 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 meta-class $w_1 \cup w_2$ thanks to the belief functions framework as shown in Fig. 1d. 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 an 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.

4.1.2. Test 2

A 3-class 2-D data set composed by three rings shown in Fig. 2 (a) is used in this example. Each class contains 303 training samples and 303 objects for testing. The radii 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. 2(b)–(d).

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 in Fig. 2(b) and (c). K-NN and EK-NN generate both 109 misclassifications. In BCKN, the objects

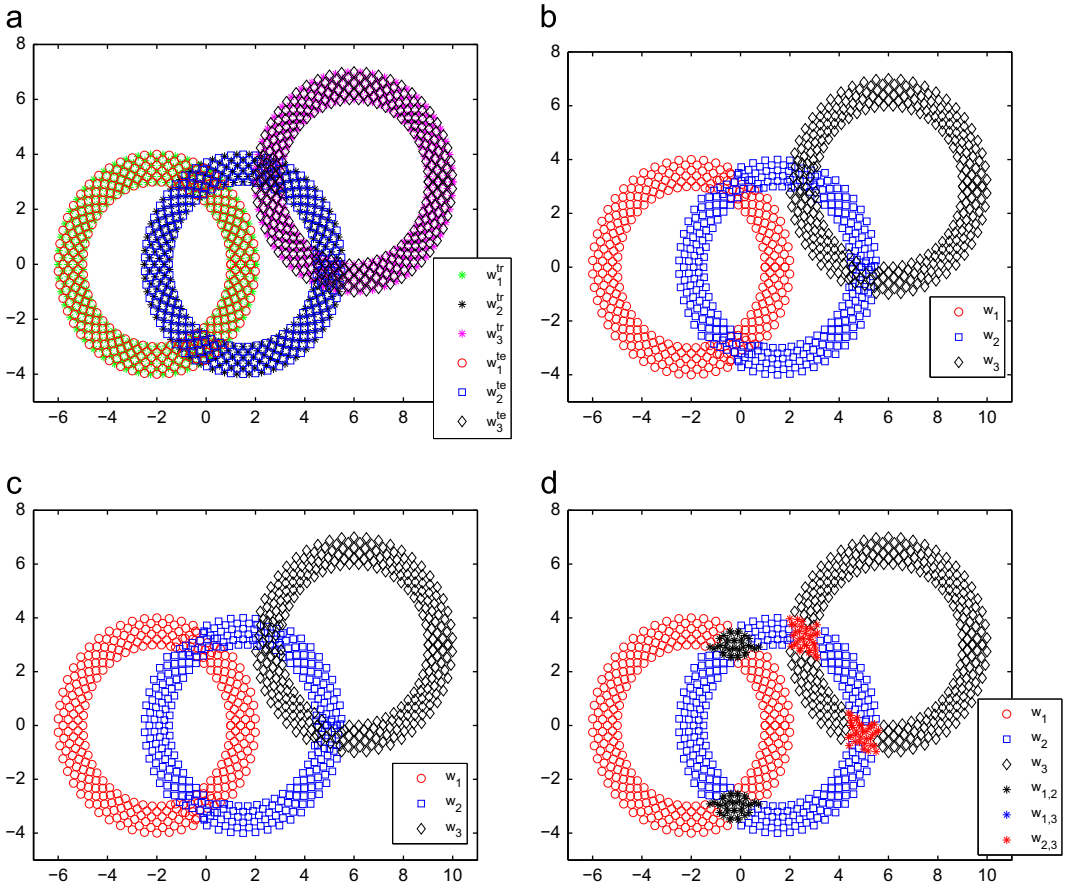

Fig. 2. Classification results of a 3-class data set by K-NN, EK-NN and BCKN. (a) Training data and test data, (b) classification result by K-NN, (c) classification result by EK-NN and (d) classification result by BCKN ($t=0.002$).

in the overlapping zones are reasonably automatically associated to meta-classes as shown in Fig. 2(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.

4.2. Experiment 2

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-class 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 mean vectors and covariance matrices ($\mathbf{I}$ being the 2×2 identity matrix):

$$\mu_1=(-5,0),\quad \Sigma_1=[1,0;0,6]$$
$$\mu_2=(5,0),\quad \Sigma_2=[1,0;0,6]$$
$$\mu_3=(0,5),\quad \Sigma_3=[6,0;0,1]$$
$$\mu_4=(0,-5),\quad \Sigma_4=[6,0;0,1]$$

There are 4×100 test samples, and the training sets contain $3\times N$ 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 in Fig. 3. 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 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, no object is committed to the meta-class with a cardinality value of three or four, and that is why we have only considered Ri_{s2} and Ri_{a2}. The tuning of parameters in these different methods has been introduced in the beginning of Section 4.

Table 2
The statistics of classification results by different methods (in %).

Method		$N=100$	$N=200$	$N=300$
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

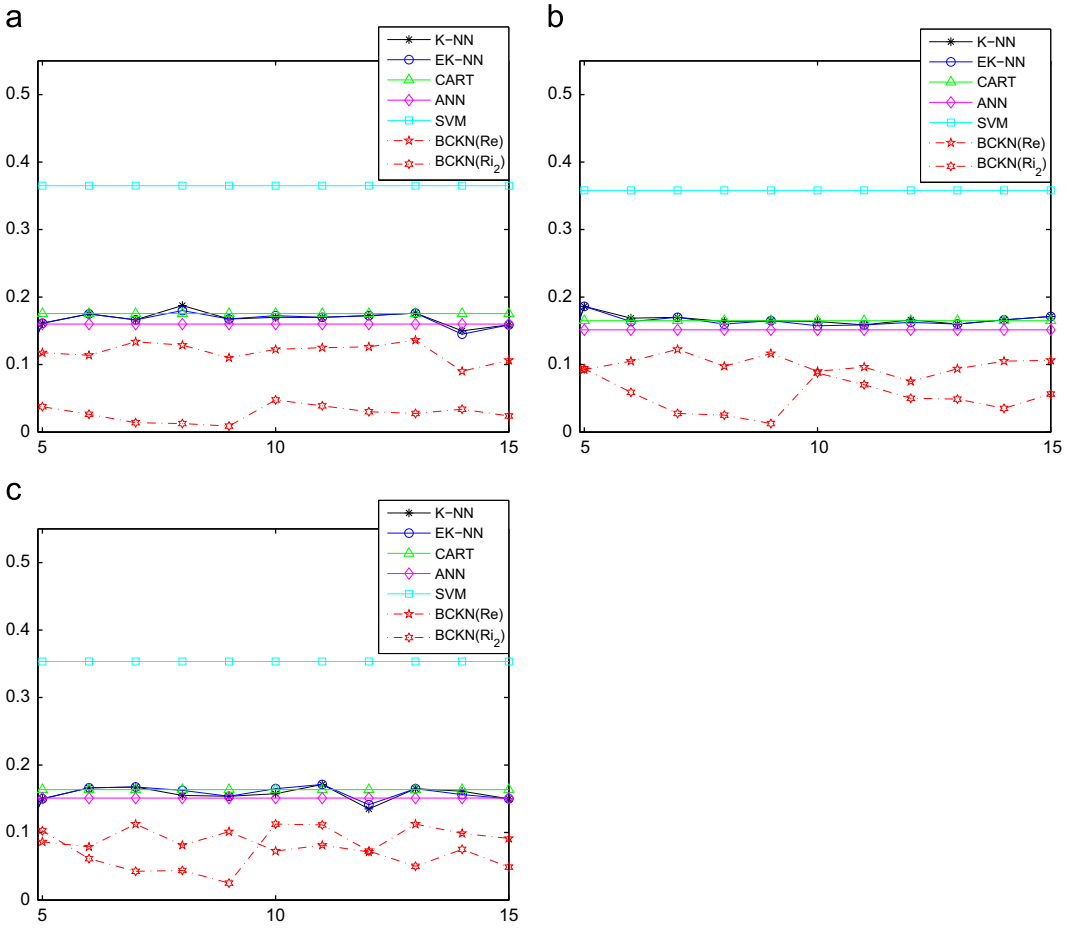

Fig. 3. Classification results of the 4-class problem by different methods. (a) Classification results with $N=100$, (b) classification results with $N=200$ and (c) classification results with $N=300$.

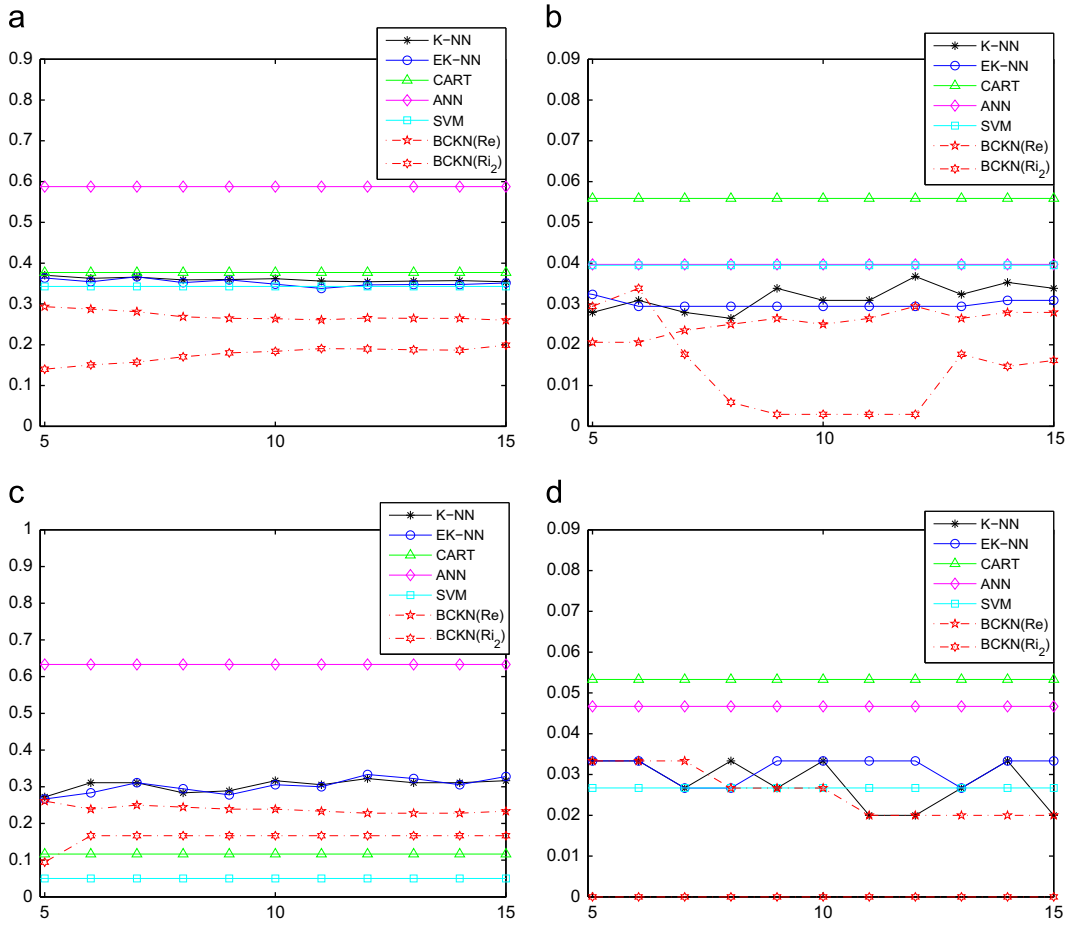

Fig. 4. Classification results of real data sets by different methods. (a) Classification results of Yeast data, (b) classification results of Breast cancer data, (c) classification results of Wine data and (d) classification results of Iris data.

In Figs. 3 and 4, the *X*-axis corresponds to the *K* values, and the *Y*-axis corresponds to the classification error rate *Re* expressed in $[0,1]$ (and also the imprecision rate Ri_2 for BCKN) of the classification methods.

We can observe in Table 2 and Fig. 3 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 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 in Table 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.

4.3. Experiment 3

Four well-known data sets available from UCI [40] (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. Three classes (*CYT*, *NUC* and *ME*3) 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 [41]. 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 in Fig. 4(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 4. The outlier class is absent in these 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 4.

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_{a_2} 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 no 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

Table 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

Table 4
The statistics of the classification results for different real data sets (in %).

Method		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
	Re_a	27.05	2.54	23.84	2.55
BCKN	Ri_{a2}	17.59	1.34	16.01	0
	Time	0.7484	0.2714	0.0211	0.0156

and ANN methods, but SVM and CART obtain better results than the other neighbor-based methods.[6] BCKN requires a bit more time-consuming than *K*-NN and EK-NN in Table 4, which says 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.

5. Conclusion

A new belief $c \times K$ neighbors (BCKNs) classifier has been developed 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 classification allows to reduce the error rate by introducing the meta-class, which characterizes the partial imprecision of classification, and it also allows 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 the 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. The computation complexity of BCKN is higher than with *K*-NN and EK-NN due to the enlarged credal classification done with BCKN. The precise evaluation and the reduction of the computational burden of the BCKN will be investigated in our future works.

[6] 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.

Acknowledgments

This work has been partially supported by National Natural Science Foundation of China (Nos. 61075029 and 61135001).

References

[1] B.W. Silveman, M.C. Jones, E. Fix, J.L. Hodges, An important contribution to nonparametric discriminant analysis and density estimation, Int. Stat. Rev. 57 (3) (1989) 233–247 1951.
[2] P. Piro, R. Nock, F. Nielsen, M. Barlaud, Leveraging *k*-NN for generic classification boosting, Neurocomputing 80 (2012) 3–9.
[3] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, Chapman and Hall (Wadsworth, Inc.), New York, 1984.
[4] A. Kumar, M. Hanmandlu, H.M. Gupta, Fuzzy binary decision tree for biometric based personal authentication, Neurocomputing 99 (2013) 87–97.
[5] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[6] N. Barakat, Andrew P. Bradley, Rule extraction from support vector machines: a review, Neurocomputing 74 (1–3) (2010) 178–190.
[7] W. An, M. Liang, Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises, Neurocomputing 110 (2013) 101–110.
[8] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.
[9] A.-L. Jousselme, P. Maupin, E. Bossé, Uncertainty in a situation analysis perspective, in: Proceedings of International Conference on Fusion, July 2003, pp. 1207–1213.
[10] A.-L. Jousselme, C. Liu, D. Grenier, E. Bossé, Measuring ambiguity in the evidence theory, IEEE Trans. SMC, Part A 36 (September (5)) (2006) 890–903.
[11] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.
[12] P. Smets, The combination of evidence in the transferable belief model, IEEE Trans. Pattern Anal. Mach. Intell. 12 (5) (1990) 447–458.
[13] P. Smets, R. Kennes, The transferable belief model, Artif. Intell. 66 (2) (1994) 191–243.
[14] F. Smarandache, J. Dezert (Eds.), Advances and Applications of DSmT for Information Fusion, American Research Press, Rehoboth, 2004–2009, pp. 1–3, ⟨http://www.gallup.unm.edu/~smarandache/DSmT-book1-3.pdf⟩.
[15] T. Denœux, A *k*-nearest neighbor classification rule based on Dempster–Shafer theory, IEEE Trans. Syst. Man Cybern. 25 (05) (1995) 804–813.
[16] T. Denœux, A neural network classifier based on Dempster–Shafer theory, IEEE Trans. Syst. Man Cybern. A 30 (2) (2000) 131–150.
[17] T. Denœux, P. Smets, Classification using belief functions: relationship between case-based and model-based approaches, IEEE Trans. Syst. Man Cybern. Part B 36 (6) (2006) 1395–1406.
[18] N. Sutton-Charani, S. Destercke, T. Denœux, Classification trees based on belief functions, in: T. Denoeux, M.-H. Masson (Eds.), Belief Functions: Theory and Applications, AISC, vol. 164, 2012, pp. 77–84.
[19] F. Karem, M. Dhibi, A. Martin, Combination of supervised unsupervised classification using the theory of belief functions, in: T. Denoeux, M.-H. Masson (Eds.), Belief Functions: Theory and Applications, AISC, vol. 164, 2012, pp. 85–92.
[20] S. Kanj, F. Abdallah, T. Denœux, Evidential multi-label classification using the random *k*-label sets approach, in: T. Denoeux, M.-H. Masson (Eds.), Belief Functions: Theory and Applications, AISC, vol. 164, 2012, pp. 21–28.
[21] M.-H. Masson, T. Denœux, ECM: an evidential version of the fuzzy c-means algorithm, Pattern Recognit. 41 (4) (2008) 1384–1397.
[22] T. Denœux, M.-H. Masson, EVCLUS: eVidential CLUStering of proximity data, IEEE Trans. Syst. Man Cybern. Part B 34 (1) (2004) 95–109.
[23] Z.-g. Liu, J. Dezert, G. Mercier, Q. Pan, Belief C-means: an extension of fuzzy C-Means algorithm in belief functions framework, Pattern Recognit. Lett. 33 (3) (2012) 291–300.
[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, Decis. Support Syst. 52 (2011) 133–141.
[25] A. Sinha, H.M. Chen, D.G. Danu, T. Kirubarajan, M. Farooq, Estimation and decision fusion: a survey, Neurocomputing 71 (13–15) (2008) 2650–2656.
[26] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[27] R.N. Dave, Clustering relational data containing noise and outliers, Pattern Recognit. Lett. 12 (1991) 657–664.
[28] Ph. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem, Int. J. Approx. Reason. 9 (1993) 1–35.
[29] A. Appriou, Uncertain data aggregation in classification and tracking processes, in: B. Bouchon-Meunier (Ed.), Aggregation and Fusion of Imperfect Information, Physica-Verlag, Heidelberg, 1998, pp. 231–260.
[30] H. Altincay, Ensembling evidential *k*-nearest neighbor classifiers through multi-modal perturbation, Appl. Soft Comput. 7 (3) (2007) 1072–1083.
[31] L.A. Zadeh, A simple view of the Dempster–Shafer theory of evidence and its implication for the rule of combination, AI Mag. 7 (2) (1986) 85–90.
[32] P. Wang, A defect in Dempster–Shafer theory, in: Proceedings of 10th Conference on Uncertainty in AI, 1994, pp. 560–566.
[33] J. Dezert, P. Wang, A. Tchamova, On the validity of Dempster–Dhafer theory, in: Proceedings of 15th International Conference on Information Fusion (Fusion 2012), Singapore, July 2012.
[34] A. Tchamova, J. Dezert, On the behavior of Dempster's rule of combination and the foundations of Dempster–Shafer Theory, best paper awards, in: Proceedings of the 6th IEEE International Conference on Intelligent Systems, Sofia, Bulgaria, September 2012.
[35] D. Dubois, H. Prade, Representation and combination of uncertainty with belief functions and possibility measures, Comput. Intell. 4 (4) (1988) 244–264.
[36] F. Smarandache, J. Dezert J., J.-M. Tacnet, Fusion of sources of evidence with different importances and reliabilities, in: Proceedings of International Conference on Fusion, Edinburgh, Scotland, UK, July 2010, pp. 26–29.
[37] J. Dezert, Z.-g. Liu, G. Mercier, Edge detection in color images based on DSmT, in: Proceedings of International Conference on Fusion, Chicago, USA, July 2011.
[38] J. Klein, O. Colot, A belief function model for pixel data, in: T. Denoeux, M.-H. Masson (Eds.), Belief Functions: Theory and Applications, AISC, vol. 164, 2012, pp. 189–196.
[39] L.M. Zouhal, T. Denœux, An evidence-theoretic *k*-NN rule with parameter optimization, IEEE Trans. Syst. Man Cybern., Part C 28 (2) (1998) 263–271.
[40] A. Frank, A. Asuncion, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2010, ⟨http://archive.ics.uci.edu/ml⟩.
[41] S. Geisser, Predictive Inference: An Introduction, Chapman and Hall, New York, NY, 1993.

Zhun-ga Liu was born in China, 1984. He received the Bachelor and Master degree from Northwestern Polytechnical University (NPU)in 2007 and 2010. He started the Ph.D. course in NPU since September 2009. He has been studied in Telecom Bretagne as a joint Ph.D. student since September 2010. His research work focuses on belief functions theory and its application in pattern recognition.

Quan Pan was born in China 1961. He received the Bachelor degree in Huazhong University of Science and Technology, and he received the Master and Doctor degree in Northwestern Polytechnical University (NPU) in 1991 and 1997. He has been a professor since 1998 in NPU. His current research interests are information fusion and pattern recognition.

Jean Dezert was born in France, 1962. He received the electrical engineering degree from Ecole Française de Radioélectricité Electronique and Informatique (EFREI), Paris, in 1985, the D.E.A. degree in 1986 from the University Paris VII and his Ph.D. from the University Paris XI, Orsay, in 1990. Since 1993, he is a senior research scientist in the Information Modeling and Processing Department (DTIM) at ONERA. His current research interest focuses on belief functions theory, especially for DSmT which has been developed by him and Prof. Smarandache.