An evidential classifier based on feature selection and two-step classification strategy

Chunfeng Lian, Su Ruan, Thierry Denoeux

An evidential classifier based on feature selection and two-step classification strategy

Chunfeng Lian[a,b], Su Ruan[b,*], Thierry Denœux[a]

[a] *Sorbonne universités, Université de technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, France*
[b] *Université de Rouen, QuantIF-EA 4108 LITIS, France*

Abstract

In this paper, we investigate ways to learn efficiently from uncertain data using belief functions. In order to extract more knowledge from imperfect and insufficient information and to improve classification accuracy, we propose a supervised learning method composed of a feature selection procedure and a two-step classification strategy. Using training information, the proposed feature selection procedure automatically determines the most informative feature subset by minimizing an objective function. The proposed two-step classification strategy further improves the decision-making accuracy by using complementary information obtained during the classification process. The performance of the proposed method was evaluated on various synthetic and real datasets. A comparison with other classification methods is also presented.

Keywords: Dempster-Shafer theory, evidence theory, belief functions, uncertain data, feature selection, classification

1. Introduction

According to whether prior probabilities and class conditional densities are needed, supervised learning methods can be divided into two main categories, namely, parametric (model-based) and nonparametric (case-based) methods [1].

*Corresponding author
Email address: su.ruan@univ-rouen.fr (Su Ruan)

Because they do not need any prior knowledge other than training samples, case-based classifiers (e.g., K-nearest neighbor rule [2], multilayer perceptrons [3], support vector machines [4] and decision trees [5]) are widely used in practice, and have proved to very efficient. However, in the case of uncertain and imprecise data, many samples may be corrupted with noise or located in highly overlapping areas; consequently, it becomes difficult for these traditional methods to obtain satisfactory classification results.

Learning effectively with partial knowledge is drawing increasing attention in statical pattern recognition. Various theories from the uncertainty management community (e.g., fuzzy set theory [6, 7], possibility theory [8], rough set theory [9] and imprecise probability theory [10]) have been used to build learning methods dealing specifically with uncertain data. The theory of belief functions, also known as Dempster-Shafer theory or Evidence theory, is an extension of both probability theory and the set-membership approach [11, 12]. It has been shown to be a powerful framework for representing and reasoning with uncertain and imprecise information. A growing number of applications of belief function theory has been reported in unsupervised learning [13, 14, 15], ensemble learning [16, 17, 18], model parameter estimation [19, 20] and partially supervised learning [21, 22].

Apart from the publications mentioned above, the use of belief functions in pattern recognition has been firstly focused on supervised learning methods. In [23], an evidence-theoretic K-nearest neighbor classification (EK-NN) rule was proposed. It provided a global treatment of imperfect knowledge regarding training data, and was further optimized in [24]. In [25], a neural network classifier based on belief functions was introduced as an adaptive version of the EK-NN. Methods for building decision trees from imperfect data were presented in [26, 27]. Regression methods using belief functions were proposed in [28, 29]. Using the notion of credal partition introduced in [13], and in order to reflect the imprecision degree of the classification, a belief-based K-nearest neighbor (BK-NN) method was proposed by Liu *et al.* in [30]. To cope with the high computational complexity of the nearest-neighbors strategy, a Credal Classifi-

cation Rule (CCR) was further developed by Liu *et al.* in [31], as a simplified version of the BK-NN. The BK-NN and CCR methods assign objects not only to specific classes, but also to the disjunction of specific classes (meta-classes). This strategy allows a reduction of misclassification rate, at the cost of leaving the class of some objects unspecified. However, in many applications, a specific decision has to be made.

In this paper, we explore two complementary ways to extract more useful knowledge from the training data:

- It often happens that the dataset contains irrelevant or redundant features. So as to efficiently learn from such imperfect training information, it is essential to find the most informative feature subset;
- Additional knowledge can be gained from the testing dataset itself to help reduce the possibility of misclassification. The "easy to classify" objects in the testing dataset can provide complementary evidence to help determine the specific class of the "hard to classify" objects.

To this end, a novel supervised learning method based on belief functions is proposed in this paper. The proposed method is composed of a feature selection procedure and a two-step classification strategy, both based on a specific mass function construction method inspired by [32]. This method, called the "Dempster+Yager" combination rule, uses features of Dempster's rule, Yager's rule [33] and Shafer's discounting procedure [11] to achieve a better representation of uncertainty and imprecision in the EK-NN classifier. Through minimizing a new criterion based on belief functions, the proposed feature selection procedure searches for informative feature subsets that yield high classification accuracy and small overlap between classes. After feature selection, the proposed two-step classification strategy uses test samples that are easy to classify, as additional evidence to help classifying test samples lying in highly overlapping areas of the feature space.

The rest of this paper is organized as follows. The background on belief functions and the traditional EK-NN classification rule is recalled in the next

section. The proposed feature selection procedure and two-step classification strategy are discussed in Section 3. In Section 4, the proposed method is tested on different synthetic and real datasets, and a comparison with other methods is presented. Finally, conclusions are given in Section 5.

2. Background

2.1. Belief functions

The theory of belief functions, also known as Dempster-Shafer or Evidence theory, was introduced by Dempster and Shafer [34, 11] and further elaborated by Smets [35, 12]. As a generalization of probability theory and set-membership approaches, the theory of belief functions has proved to be an effective theoretical framework for reasoning with uncertain and imprecise information. In this section, only the basic definitions will be recalled.

Let X be a variable taking values in the *frame of discernment* $\Omega = \{\omega_1, \cdots, \omega_c\}$. Uncertain and imprecision knowledge about the actual value of X can be represented by a *mass function*, defined as a mapping m from 2^Ω to [0,1] such that $m(\emptyset) = 0$ and

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

The subsets A of Ω such that $m(A) > 0$ are called the *focal elements* of mass function m. If all focal elements are singletons, m is said to be *Bayesian*; it is then equivalent to a probability distribution. A mass function m with only one focal element is said to be *categorical* and is equivalent to a set.

For any subset $A \subseteq \Omega$, the probability that the evidence supports A can be defined as

$$Bel(A) = \sum_{B \subseteq A} m(B), \tag{2}$$

while the probability that the evidence does not contradict A is

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{3}$$

Functions Bel and Pl are called, respectively, the *belief function* and the *plausibility function* associated to m. Belief and plausibility functions are in one-to-one correspondence with mass functions. They can be regarded as providing lower and upper bounds for the degree of belief that can be attached to each subset of Ω.

Two mass functions m_1 and m_2 derived from independent items of evidence can be combined by Dempster's rule [11] to obtain a new mass function $m_1 \oplus m_2$, defined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1-Q} \sum_{B \cap C = A} m_1(B) m_2(C), \tag{4}$$

for all nonempty $A \subseteq \Omega$, where $Q = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ is the *degree of conflict* between m_1 and m_2.

When the degree of conflict Q between m_1 and m_2 is large, the combination result obtained by Dempster's rule may become unreliable. To cope with this problem, Yager [33] proposed to transfer the conflicting mass to the frame of discernment Ω, yielding the following combined mass function,

$$m(A) = \begin{cases} \sum_{B \cap C = A} m_1(B) m_2(C) & \text{if } A \neq \emptyset, A \subset \Omega; \\ m_1(\Omega) m_2(\Omega) + \sum_{B \cap C = \emptyset} m_1(B) m_2(C) & \text{if } A = \Omega; \\ 0, & \text{if } A = \emptyset. \end{cases} \tag{5}$$

A mass function m can be transformed into a probability function for decision-making. In Smet's *Transferable Belief Model* [12, 35], the *pignistic probability transformation* transforms a mass function into the following probability distribution:

$$BetP(\omega_q) = \sum_{A \subseteq \Omega : \omega_q \in A} \frac{m(A)}{|A|}, \tag{6}$$

for all $\omega_q \in \Omega$.

2.2. Evidential K-NN classifier

In [23], an evidence-theoretic K-nearest neighbor classification (EK-NN) rule was proposed. In this rule, each neighbor of a sample to be classified is treated as an item of evidence that supports certain hypotheses regarding the class label

of this sample. The strength of this evidence decreases with the distance to the test sample. Evidence from the K nearest neighbors is pooled using Dempster's combination rule to make the final decision.

Let $\{(X_i, Y_i), i = 1, \cdots, N\}$ be a collection of N training examples, in which $X_i = [x_1, \cdots, x_m]$ is the ith training sample with m features and $Y_i \in \{\omega_1, \cdots, \omega_c\}$ is the corresponding class label. Given an input test sample X^t, the EK-NN classifier uses the following steps to determine its class label:

- Let X_j be one of the K nearest neighbors of X^t with class label $Y_j = \omega_q$. Then the mass function induced by X_j, which supports the assertion that X^t also belongs to ω_q is

$$m_{t,j}(\{\omega_q\}) = \alpha \exp(-\gamma_q d_{t,j}^2), \tag{7a}$$

$$m_{t,j}(\Omega) = 1 - \alpha \exp(-\gamma_q d_{t,j}^2), \tag{7b}$$

  where $d_{t,j}$ is the distance between X_j and X^t. According to [23], parameter α can be heuristically set as 0.95, and $\gamma_q > 0$ ($q \in \{1, \cdots, c\}$) can be determined separately for each class as $1/d_q^2$, where d_q is the mean distance between two training samples belonging to class ω_q. The value of α and $\gamma_q > 0$ can also be optimized using the training data [24];

- Dempster's rule (4) is then used to combine all neighbors' mass functions. Test sample X^t is then assigned to the class with the maximum pignistic probability (6).

Besides Dempster's rule, some other methods were also proposed in recent publications to combine neighbors' mass functions. For instance, in the evidential classifier method [32], a new combination rule was developed specifically for outlier detection.

3. Proposed Method

Both the feature selection procedure and the two-step classification strategy proposed in this paper need proper handling of the uncertainty and imprecision

in the data. To this end, a simple and specific mass function construction procedure will first be introduced in Section 3.1. The proposed feature selection procedure and two-step classification strategy will then be presented, respectively, in Sections 3.2 and 3.3.

3.1. Construction of mass functions

We developed a specific combination rule to compute a mass function about the class label of a test sample, based on the evidence of its K-nearest neighbors. The proposed hybrid combination rule shares some features with Dempster's rule, Yager's rule [33] and Shafer's discounting procedure [11]. It will be referred to as the *"Dempster+Yager" rule* for short. In this rule, only singletons and the whole frame of discernment are considered as focal elements. Hence, all the imprecision will be succinctly represented by masses assigned to the whole frame of discernment.

As before, let $\{(X_i, Y_i), i = 1, \cdots, N\}$ be the training data. For an input instance X_t under test, the frame of discernment is $\Omega = \{\omega_1, \cdots, \omega_c\}$. Using the Dempster+Yager rule, the determination of X_t's mass function can be described as follows.

Step 1 As in the classical E-KNN method [23], the K-nearest neighbors of X_t in the training set according to the Euclidean distance measure are first found. Let X_j be the jth nearest neighbor of X_t with $Y_j = \omega_q$. The evidence regarding X_t's class label provided by X_j is quantified as described by (7).

Step 2 Nearest neighbors with the same class label ω_q are then grouped in a set Γ_q $(q = 1, \ldots, c)$. As the mass functions in the same set Γ_q have the same focal elements, there is no conflict between them. So, regardless of outliers (a particular situation that is not considered in our approach), Dempster's rule is appropriate to combine the pieces of evidences in Γ_q. As a result, the evidence provided by nonempty Γ_q is represented as a

simple mass function,

$$m_t^{\Gamma_q}(\{\omega_q\}) = 1 - \prod_{j \in \Gamma_q} m_{t,j}(\Omega), \tag{8a}$$

$$m_t^{\Gamma_q}(\Omega) = \prod_{j \in \Gamma_q} m_{t,j}(\Omega). \tag{8b}$$

If Γ_q is empty, then $m_t^{\Gamma_q}$ is defined as the vacuous mass function defined by $m_t^{\Gamma_q}(\Omega) = 1$;

Step 3 When most neighbors of a testing instance X_t belong to a specific class (e.g., ω_q), the degree belief that X_t also belongs to this class should be large. Consequently, we can postulate that the reliability of the evidence provided by each set Γ_q is increasing with its cardinality $|\Gamma_q|$. The mass functions obtained in last step should thus be further discounted as

$$dm_t^{\Gamma_q}(\{\omega_q\}) = \left(\frac{|\Gamma_q|}{|\Gamma_{max}|}\right)^{\eta} m_t^{\Gamma_q}(\omega_q), \tag{9a}$$

$$dm_t^{\Gamma_q}(\Omega) = 1 - \left(\frac{|\Gamma_q|}{|\Gamma_{max}|}\right)^{\eta} m_t^{\Gamma_q}(\omega_q), \tag{9b}$$

where $|\Gamma_{max}|$ is the maximum cardinality within $\{|\Gamma_1|, \cdots, |\Gamma_c|\}$, and $\eta \geq 0$ is a coefficient that controls the discounting level. A larger value of η results in stronger discounting. In particular, when $\eta = 0$, there is no discounting at all. The value of η can be determined by minimizing the leave-one-out cross-validation error rate. Generally, good results are obtained if we take $\eta \in [0, 2]$.

Step 4 After the discounting procedure described in the previous step, the mass functions at hand may still be partially conflicting, especially when there are similar numbers of nearest neighbors with different class labels. Since Yager's rule can have a better behavior that Dempster's rule when combining highly conflicting evidences [36, 33], it is chosen at this step to fuse the probably conflicting mass functions in sets Γ_1 to Γ_c obtained in the previous step. As the result, the global mass function regarding the

class label of object X_t is finally given by

$$m_t(\{\omega_q\}) = dm_t^{\Gamma_q}(\omega_q) \prod_{h\in\{1,\dots c\}\setminus q} dm_t^{\Gamma_h}(\Omega), \quad q = 1,\dots,c, \tag{10a}$$

$$m_t(\Omega) = 1 - \sum_{q=1}^{c} \left(dm_t^{\Gamma_q}(\{\omega_q\}) \prod_{h\in\{1,\dots c\}\setminus q} dm_t^{\Gamma_h}(\Omega) \right), \tag{10b}$$

The focal elements of m_t are singletons and the whole frame of discernment. Consequently, the credibility and plausibility criteria (i.e., Bel_t and Pl_t) will lead to the same hypotheses about X_t.

The mass function construction procedure discussed above is summarized as a flowchart in Figure 1. It combines the advantages of Dempster's and Yager's rules. Hence, in classification applications, this specific procedure allows for a more robust representation of uncertainty than that obtained using any of the two classical combination rules. To better illustrate the performance of the proposed Dempster+Yager rule, two examples are given below.

Example 1. *To simulate a situation with conflicting pieces of evidence, we let the number of nearest neighbors be $K = 3$, and we assume that the test sample X_t lies at the same distance to all the three nearest neighbors. The first two neighbors of X_t belong to class ω_1, and the third one belongs to class ω_2. We assume that $\Omega = \{\omega_1, \omega_2\}$ and $\eta = 2$. The three mass functions and the result of their combination by Dempster's rule, Yager's rule and our Dempster+Yager rule are shown in Table 1. In this case, the Dempster+Yager rule is more conservative than Dempster's rule (it assigns a larger mass to Ω), while being more specific than Yager's rule.*

Example 2. *Table 2 illustrates an even more conflicting situation, in which two neighbors belong to ω_1 and two neighbors belong to ω_2. We still assume that the test sample X_t is at the same distance to all nearest neighbors, and we take $\eta = 2$. In this case, the Dempster+Yager rule yields the same result as Yager's rule. Both rules assign a large mass to the whole frame of discernment.*

3.2. Feature selection based on belief functions

In pattern recognition applications, the data may contain irrelevant or redundant features. Feature selection techniques are intended to cope with this issue. They aim to select a subset of features that can facilitate data interpretation while reducing storage requirements and improving prediction performance [37]. Filter, wrapper and embedded methods are three main categories of algorithms that are widely used for feature selection [38]. Filter methods such as described in [39, 40, 41], which use variable ranking as the principal selection mechanism, are simple and scalable. However, they may produce a suboptimal subset because they do not take into account the correlation between features [37]. In contrast, wrapper and embedded methods, such as sequential selection algorithms [42, 43] and direct objective optimization methods [44], use the prediction accuracy of given classifiers as the criterion for selecting feature subset. They are more likely to find optimal feature subsets than filter methods. However, up to now, none of the available wrapper or embedded methods were designed to work for imperfect data with high uncertainty and/or imprecision. Such a feature selection procedure, based on belief functions, is introduced in this section.

The proposed method tackles the feature selection issue from a novel perspective. It aims to meet the following three requirements:

1. The selected features should be informative regarding the class labels, i.e., they should not yield lower classification accuracy than the complete set of features;
2. The selected feature subset should have the ability to reduce the uncertainty of the data, i.e., it should result in a small overlap between different classes in the feature space;
3. The selected features should be as sparse as possible. A feature subset with smaller cardinality implies lower storage requirement and lower risk of overfitting.

The above three requirements can be met simultaneously by minimizing an

objective function derived from the training samples. In order to present this objective function clearly, a simple form of weighted Euclidean distance should be discussed at first. Depending on the values of a binary coefficient vector, this weighted Euclidean distance will generate different sets of K nearest neighbors for a sample under test. The weighted distance between a test sample X^t and a training sample X_i with m features is defined as

$$d_{t,i} = \sqrt{\sum_{p=1}^{m} \lambda_p (d_{t,i}^p)^2}, \tag{11}$$

where $d_{t,i}^p$ $(1 \leq p \leq m)$ is the difference between the values of the pth components of the two feature vectors and $\lambda_p \in \{0, 1\}$ is the corresponding coefficient. Obviously, the feature subset can be selected by changing the values of the coefficient vector. As the result, the pth component of the feature vector will be selected when $\lambda_p = 1$ and it will be eliminated when $\lambda_p = 0$.

Based on the weighted Euclidean distance measure (11), and using the mass function construction procedure introduced in Section 3.1, we can propose an objective function satisfying the above three requirements for a qualified feature subset. Let $\{(X_i, Y_i), i = 1, \cdots, N\}$ be a training set. The proposed three-term objective function is

$$obj = \frac{1}{n} \sum_{i=1}^{n} \sum_{q=1}^{c} (Pl_i(\omega_q) - t_{i,q})^2 + \frac{\rho}{n} \sum_{i=1}^{n} m_i(\Omega) + \delta \sum_{p=1}^{m} [1 - \exp(-\mu \lambda_p)]. \tag{12}$$

In (12), the first term is a squared error corresponding to the first requirement discussed above, Pl_i is the plausibility function of training sample X_i and $t_{i,q}$ is the qth component of a c-dimensional binary vector t_i such that $t_{i,q} = 1$ if $Y_i = \omega_q$ and $t_{i,q} = 0$ otherwise. The second term is the average mass assigned to the whole frame of discernment. It penalizes feature subsets that result in high uncertainty and imprecision, thus allowing us to meet the second requirement. The last term, which is an approximation of the l_0-norm as used in[45], forces the selected feature subset to be sparse. Here, ρ and δ are two hyper-parameters in $[0, 1]$, which influence, respectively, the number of uncertainty samples and the sparseness of resulting feature subset. Their values should be tuned to

maximize the classification accuracy. Coefficient μ is kept constant; according to [45], it is often set to 5.

Using (7)-(10), the objective function (12) can be written as

$$obj = \frac{1}{n}\sum_{i=1}^{n}\sum_{q=1}^{c}\left(1 - t_{i,q} - \sum_{h\neq q} B_h^i\right)^2 + \frac{\rho}{n}\sum_{i=1}^{n}\left(1 - \sum_{q=1}^{c} B_q^i\right) + \delta\sum_{p=1}^{m}[1 - \exp(-\mu\lambda_p)], \quad (13)$$

with

$$B_q^i = A_q^i \prod_{s\in\{1,\ldots c\}\setminus q} (1 - A_s^i) \quad (14)$$

and

$$A_q^i = \left(\frac{|\Gamma_q^i|}{|\Gamma_{max}^i|}\right)^{\eta}\left(1 - \prod_{j\in\Gamma_q^i}\left[1 - \alpha\exp(-\gamma_q \cdot d_{i,j}^2)\right]\right), \quad (15)$$

where $d_{i,j}$ is the distance between the training sample X_i and its jth nearest neighbor computed using (11) ,with coefficients $\{\lambda_1, \cdots, \lambda_c\}$ to be optimized. During the optimization process, the K nearest neighbors for each training sample (X_i, Y_i) are determined by the weighted distance measure (11) with the current weights $\{\lambda_1, \cdots, \lambda_c\}$. The mass functions m_i are computed using the construction procedure presented in Section 3.1, followed by the calculation of the plausibility value Pl_i using (3). Mass and plausibility values change with binary coefficients $\{\lambda_1, \cdots, \lambda_c\}$, which finally drives the decrease of the objective function (12)-(13).

As a global optimization method, the integer genetic algorithm [46, 47] can properly solve the integer optimization problem without gradient calculation. Hence, it is chosen in this paper to optimize $\{\lambda_1, \cdots, \lambda_c\}$, so as to find a good feature subset.

3.3. Two-step classification

After selecting features using the procedure described in the previous section, a two-step classification strategy allows us to classify unknown test samples

based on belief functions. For a test dataset $\mathbf{T} = \{S_j, j = 1, \ldots, n_t\}$, this two-step classification strategy can be described as follows:

1. Using the Dempster+Yager combination rule, the mass function m_j of each test sample S_j is first derived from training pairs (X_i, Y_i), $i = 1, \ldots, N$. Based on m_j, the collection $\mathbf{T}$ is divided into two groups T^1 and T^2, where $\mathrm{T}^1 = \{S_j : \max_{A \subseteq \Omega} m_j(A) \neq m_j(\Omega)\}$ and $\mathrm{T}^2 = \{S_j : \max_{A \subseteq \Omega} m_j(A) = m_j(\Omega)\}$;
2. Then, test samples in T^1 are classified into the classes with highest masses. For instance, if $m_j(\{\omega_1\}) > m_j(\{\omega_q\})$ for all $q \neq 1$, we label S_j as ω_1;
3. After classifying the test samples in T^1, we add these labeled test samples to the training set $\{(X_i, Y_i), i = 1, \ldots, N\}$, and therefore obtain a larger training set $\{(X_i', Y_i'), i = 1, \ldots, N'\}$. The center (or prototype) p_j of each class ω_j is then defined by averaging the training samples corresponding to this class,
$$p_j = \frac{1}{c_j} \sum_{Y_i' = \omega_j} X_i', \tag{16}$$
where c_j is the cardinality of the set $\{X_i' | Y_i' = \omega_j\}$ of training patterns in class ω_j, and $j = 1, \ldots, c$.
4. To each test pattern in group T^2 (i.e., uncertain samples with the largest mass of belief on Ω), and taking into account the correlations of the given dataset, the Mahalanobis distance measure is used to compute the distances of this test pattern to each class center. Let S_0 be a test sample within T^2, the distance from it to center p_j is
$$md(S_0, p_j) = \sqrt{\sum_{q=1}^{m} \frac{(S_0^q - p_j^q)^2}{(\delta_j^q)^2}}, \tag{17}$$
where S_0^q and p_j^q are, respectively, the qth dimension of S_0 and p_j, and δ_j^q is the standard deviation of the qth feature among training samples belonging to class ω_j. Based on the distances $\{md(S_0, p_1), \cdots, md(S_0, p_1)\}$, S_0 is finally allocated to the nearest class.

Using the procedure discussed above, test samples that are easy to classify provide additional evidence to help classifying highly uncertainty test samples. As will be shown in the next section, this strategy enhances the classification accuracy of the EK-NN rule, especially in highly overlapping regions of the feature space.

4. Experimental Results

The presented experiments are composed of two parts. In the first part, the feasibility of the proposed feature selection procedure was evaluated on two synthetic datasets. In each synthetic dataset, the numbers of relevant, redundant and irrelevant features were varied to assess the robustness of the method under different situations. In addition, to show the validity of the two-step classification strategy, we compared it in detail with the EK-NN classifier [23, 24, 1] on another synthetic dataset.

In the second part, we first compared the performance of the proposed feature selection procedure with some classical wrapper selection methods on seven real datasets. Then, on the same real datasets, the classification accuracy of the proposed two-step classification strategy was compared with other well-known classifiers after selecting features using different methods. Finally, we tried to determine whether the proposed feature selection procedure can help to improve classification performance of other classifiers. The classification performance of the proposed two-step procedure was further compared with other methods using the same feature subsets selected by the proposed procedure.

4.1. Performance on synthetic datasets

4.1.1. Feature selection

The feasibility of the proposed feature selection procedure was assessed on two different kinds of synthetic datasets. The generating mechanisms for the two different datasets are described below.

Synthetic Data 1: These data were generated using the procedure described in [48]. The feature space contains n_r informative features uniformly distributed between -1 and +1. The output label for a given sample is defined as

$$y = \begin{cases} \omega_1 & \text{if } \max_i(x_i) > 2^{1-\frac{1}{n_r}} - 1, \\ \omega_2 & \text{otherwise,} \end{cases} \tag{18}$$

were x_i is the ith feature. Besides the relevant features, there are also n_i irrelevant features uniformly distributed between -1 and +1, without any relation with the class label, and n_c redundant features copied from the relevant features. The optimal discriminating surface for this synthetic data is highly non-linear.

Synthetic Data 2: To generate these data, two informative features were first obtained from four different two-dimensional normal distributions, $N(m_1, I)$ and $N(m_2, I)$ for class 1; $N(m_3, I)$ and $N(m_4, I)$ for class 2. Here, $m_1 = [3, 3]$, $m_2 = [6, 6]$, $m_3 = [3, 6]$ and $m_4 = [6, 3]$. In addition, there are n_i irrelevant features, all randomly generated from the normal distribution $N(4.5, 2)$, and n_c redundant features copied from relevant features.

For both synthetic datasets, we set $n_r = 2$, $n_i \in \{6, 16, 26, 36, 46\}$ and $n_c = 2$ to simulate five different situations. In each case, we generated 150 training instances, and used the proposed procedure to search for the most informative feature subset. Then, 150 test instances were generated. We used the EK-NN classifier to classify these test instances with all features, and simultaneously used the proposed two-step classification strategy to classify them with all features and with the selected feature subset. In the five situations, we always set $\eta = 0.5$, $\rho = 0.5$, $\delta = 0.05$ and $K = 5$. The results are shown in Tables 3 and 4. For both datasets, the selection procedure always found the two relevant features. The two-step classification strategy resulted in higher accuracy than the EK-NN classifier. The feature selection procedure brought further improvement of classification performance, especially when the dimension of the initial feature space was large. These results show the feasibility of the proposed feature

selection procedure.

4.1.2. Two-step classification

In addition to the previous experiment, the performance of the proposed two-step classification strategy was tested solely on another synthetic dataset constructed from four normal distributions with means $m_1 = [3, 3]$, $m_2 = [3, 6.5]$, $m_3 = [6.5, 3]$, $m_4 = [6.5, 6.5]$ and variance matrix $\Sigma = 2I$. Instances generated from $N(m_1, \Sigma)$ and $N(m_2, \Sigma)$ with equal probabilities were labeled as ω_1, while other instances generated from $N(m_3, \Sigma)$ and $N(m_4, \Sigma)$ with equal probabilities were labeled as ω_2. Classes ω_1 and ω_2 had the same number of instances, and the sizes of training and testing datasets were both 500.

The classification results of the two-step classification strategy were compared with those of the EK-NN classifier with $K = 5$ and $\eta = 0.5$. Figure 2(a) shows the training samples and the corresponding test samples. Figures 2(b) and (c) display the credal partitions (i.e., the mass functions for each of the test samples [13, 14]) obtained, respectively, using the EK-NN classifier and the proposed method. The blue, green and black points represent instances with highest mass function on $\{\omega_1\}$, $\{\omega_2\}$ and Ω, respectively. When comparing Figures 2(b)-(c) with Figure 2(a), we can see that the proposed method results in more imprecise mass functions for the test samples in overlapping regions. This is mainly because the proposed Dempster+Yager rule has better ability than Dempster's rule to deal with highly imprecise instances (such as the boundary samples shown in Figure 2(c)).

Figures 2(d)-(f) show the classification results obtained, respectively, by EK-NN, the Dempster+Yager rule and the two-step classification strategy; the magenta stars represent misclassified instances. These results show that the proposed Dempster+Yager combination rule yields higher classification accuracy than EK-NN on these imprecise data and the two-step classification strategy further improves the performance. The calculated error rates for EK-NN, Dempster+Yager combination rule and two-step classification strategy are, respectively, 9.80%, 8.80% and 7.80%.

In addition, we also estimated the influence of parameter η on our two-step classification procedure, using this synthetic dataset. The value of η was chosen in$\{0, 0.5, 1, 1.5, 2\}$, K was set to 5, and we evaluated the performance 50 times with each η. The average misclassification error rates are reported in Table 5. As can be seen, the value of η had some limited influence on the classification accuracy, although the procedure appears not to be very sensitive to this coefficient. The best performance was obtained with $\eta = 0.5$.

4.2. Performance on real datasets

In this section, the proposed feature selection procedure and two-step classification strategy are compared with some classical wrapper selection methods and usual classifiers. The comparison was performed on seven real datasets. Six of them were downloaded from the UCI Machine Learning Repository [49], and one (the lung cancer dataset) was obtained from real patients[1]. Some characteristics of these datasets are summarized in Table 6. As in [31], "in the yeast dataset, three classes named as CYT, NUC and ME3 were selected, since these three classes are close and difficult to discriminate".

4.2.1. Feature selection performance

The proposed feature selection procedure was compared with three classical wrapper methods: sequential forward selection (SFS), sequential backward selection (SBS) and sequential floating forward selection (SFFS) [42, 38]. We used ten-fold cross validation for the six UCI datasets and the leave-one-out strategy for the lung cancer data (since it has only 25 instances). For all datasets, we iteratively chose one subset of the data as the test set, and treated the other subsets of data as training samples. At each iteration, we used SFS, SBS, SFFS and the proposed procedure to select features from the training data, and then executed the proposed two-step classification strategy to classify test instances with the selected feature subsets. The average misclassification rates obtained

[1]This lung tumor dataset was provided by laboratory LITIS and Centre Henri Becquerel, 76038 Rouen, France.

by different methods were calculated. In addition, based on feature frequency statistics, the robustness of selected feature subsets was evaluated using the method introduced in [50].

The misclassification rate, robustness and average feature subset size for all methods are summarized in Table 7. As can be seen, the proposed feature selection procedure performed uniformly well on all datasets. It resulted in more robust feature subsets than the other three classical wrapper methods, and simultaneously yielded higher classification accuracy.

4.2.2. Classification performance

Using the same seven real datasets as in the previous experiment, the classification performance of the proposed two-step classification was compared with that of six other classifiers: Artificial Neural Networks (ANN) [51], Classification And Regression Tree (CART) [5], Support Vector Machine (SVM) [4], EK-NN, Belief-based K-Nearest neighbor classifier (BK-NN) [30] and CCR [31]. The first three methods are classical classifiers, while the last three are either well-known or recent evidential classifiers based on belief functions. We can remark that, in BK-NN and CCR, the classification performance is assessed using two measures: the error rate $R_e = (N_e/T) \times 100\%$, where N_e is the number of misclassified samples assigned to wrong meta-classes, and T is the number of test samples; and the imprecision rate $R_I = (N_I/T) \times 100\%$, where N_I is the number of test samples with highest mass functions on non-singletons (i.e., on meta-classes). The BK-NN and CCR methods do not make any direct decision for highly imprecise samples, but transfer them to the meta-classes. Hence, the error rate R_e of BK-NN and CCR is decreased.

Since the proposed method includes feature selection, a classical wrapper selection method, sequential floating forward selection (SFFS), was used with all the other classifiers, to make the classification results comparable. As in the previous experiment, we used ten-fold cross-validation for the six UCI datasets and leave-one-out for the lung cancer data. The average misclassification rates obtained by different classifiers are reported in Table 8. As can be seen, the

proposed method has higher classification accuracy than those of ANN, CART, SVM and EK-NN on all datasets, especially on the lung cancer data. BK-NN and CCR resulted in the lowest error rate on the Seeds and Wine data. However, due to the fact that a nonspecific decision has been made for uncertain objects, they also have large imprecision rates. Therefore, we can conclude that the proposed classification method performed well on these real datasets.

4.2.3. Generality of the proposed method

To evaluate the generality of the proposed feature selection method, we tried to determine whether feature subsets selected by it can improve the classification performance of other classifiers. To this end, the above classifiers were used again to classify the same real datasets, using all the features and feature subsets selected by the proposed method. We used the same protocol as in the previous experiment (ten-fold cross validation for the six UCI datasets and leave-one-our for the cancer data). The average classification error rates are reported in Table 9. In this experiment, a selected feature subset was regarded as feasible for a testing classifier, if it results in no less classification accuracy than the whole set of features. The notations to show whether selected feature subsets are feasible for given classifiers are also presented in Table 9.

Based on obtained results, we can see that the feature subsets selected by the proposed method were feasible for testing classifiers in most cases. Especially, on the Iris and Lung Cancer data, the selected feature subsets resulted in higher accuracy for all classifiers; on the WDBC and Parkinsons data, they were not feasible only for ANN. To sum up, among the 49 classifier-dataset configurations, the proposed feature selection procedure failed eight times, including three times for ANN, twice for CART and CCR, and once for EK-NN. These results show that the proposed feature selection procedure is, in some sense, general as it can be used with other classifiers. However, it works better if it is used for the proposed two-step classification (it always resulted in large improvement of classification accuracy), and other evidential classifiers based on belief functions and K-nearest neighbor strategy (such as EK-NN and BK-NN). As shown in

Table 9, the proposed two-step classification resulted in the lowest classification error on most datasets using the selected feature subsets.

Since the proposed feature selection procedure seems to be applicable to other classifiers, using the same feature subsets selected by it, we further compared the classification performance of the proposed two-step classification with that of other classifiers. In order to make the comparison more comprehensive, we used two-fold cross-validation for the six UCI datasets, so as to simulate a situation in which there are more test data but less training data. The comparison was executed 200 times. The average error rates for the different classifiers are reported in Table 10. As can be seen, all classifiers performed poorly on the Yeast data. This dataset is actually very difficult to classify. The BK-NN and CCR methods yielded lower error rates than did our method on these data. However, due to the fact that nonspecific decisions can be made for uncertain objects, they also yielded large imprecision rates. Similar results can be found on the Iris and Seeds data when comparing BK-NN with our method. On the WDBC and Parkinsons data, EK-NN and the proposed two-step classification had similar performance. On the Lung Cancer data, both SVM and our two-step classification lead to perfect prediction with the selected feature subset.

In summary, it appears from these results that the proposed two-step classification generally outperformed the other classifiers on the real datasets considered in these experiments. The proposed feature selection procedure has also been found to yield better results when used jointly with the proposed two-step classification strategy.

5. Conclusions

In this paper, we addressed the problem of learning effectively from insufficient and uncertain data. The contribution of this paper is threefold. First, we proposed a variant of the EK-NN method based on a hybrid Dempster+Yager rule, which transfers part of the conflicting mass to the frame of discernment. This new mass construction method results in less specific mass functions than

those obtained using the orignal EK-NN method introduced in [23]. The second contribution is a feature selection method that finds informative feature subsets by minimizing a special objective function using mixed integer genetic algorithm. This objective function is designed to minimize the imprecision of the mass functions, so as to obtain feature subspaces that maximize the separation between classes. Finally, the third contribution is a two-step classification strategy, which was shown to further improve classification accuracy by using already classified objects as additional pieces of evidence. These three improvements of the EK-NN method were assessed separately and jointly using several synthetic and real datasets. The proposed procedures were shown to have excellent performance as compared to other state-of-art feature selection and classification algorithms.

Acknowledgements

This research was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02). It was also supported by a scholarship from the Chinese Research Council.

References

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

[2] E. Fix, J. L. Hodges Jr, Discriminatory analysis-nonparametric discrimination: consistency properties, Tech. rep., DTIC Document (1951).

[3] H. White, Learning in artificial neural networks: A statistical perspective, Neural computation 1 (4) (1989) 425–464.

[4] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[5] L. Olshen, C. J. Stone, Classification and regression trees, Wadsworth International Group.

[6] L. A. Zadeh, Fuzzy sets, Information and control 8 (3) (1965) 338–353.

[7] G. J. Klir, B. Yuan, Fuzzy sets and fuzzy logic, Vol. 4, Prentice Hall New Jersey, 1995.

[8] D. Dubois, H. Prade, Ranking fuzzy numbers in the setting of possibility theory, Information sciences 30 (3) (1983) 183–224.

[9] Z. Pawlak, Imprecise Categories, Approximations and Rough Sets, Springer, 1991.

[10] P. Walley, Towards a unified theory of imprecise probability, International Journal of Approximate Reasoning 24 (2) (2000) 125–148.

[11] G. Shafer, A mathematical theory of evidence, Vol. 1, Princeton university press Princeton, 1976.

[12] P. Smets, R. Kennes, The transferable belief model, Artificial intelligence 66 (2) (1994) 191–234.

[13] T. Denœux, M.-H. Masson, Evclus: evidential clustering of proximity data, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 34 (1) (2004) 95–109.

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

[15] S. Le Hegarat-Mascle, I. Bloch, D. Vidal-Madjar, Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing, Geoscience and Remote Sensing, IEEE Transactions on 35 (4) (1997) 1018–1031.

[16] B. Quost, M.-H. Masson, T. Denœux, Classifier fusion in the Dempster–Shafer framework using optimized t-norm based combination rules, International Journal of Approximate Reasoning 52 (3) (2011) 353–374.

[17] H. Altınçay, Ensembling evidential k-nearest neighbor classifiers through multi-modal perturbation, Applied Soft Computing 7 (3) (2007) 1072–1083.

[18] M.-H. Masson, T. Denœux, Ensemble clustering in the belief functions framework, International Journal of Approximate Reasoning 52 (1) (2011) 92–109.

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

[20] T. Denœux, Maximum likelihood from evidential data: an extension of the em algorithm, in: Combining Soft Computing and Statistical Methods in Data Analysis, Springer, 2010, pp. 181–188.

[21] E. Côme, L. Oukhellou, T. Denœux, P. Aknin, Learning from partially supervised data using mixture models and belief functions, Pattern recognition 42 (3) (2009) 334–348.

[22] E. Ramasso, T. Denœux, Making use of partial knowledge about hidden states in hmms: an approach based on belief functions., IEEE Transactions on Fuzzy Systems (2013) 1–12.

[23] T. Denœux, A k-nearest neighbor classification rule based on Dempster-Shafer theory, Systems, Man and Cybernetics, IEEE Transactions on 25 (5) (1995) 804–813.

[24] L. M. Zouhal, T. Denœux, An evidence-theoretic k-nn rule with parameter optimization, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 28 (2) (1998) 263–271.

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

[26] T. Denœux, M. S. Bjanger, Induction of decision trees from partially classified data using belief functions, in: Systems, Man, and Cybernetics, 2000 IEEE International Conference on, Vol. 4, IEEE, 2000, pp. 2923–2928.

[27] Z. Elouedi, K. Mellouli, P. Smets, Belief decision trees: theoretical foundations, International Journal of Approximate Reasoning 28 (2) (2001) 91–124.

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

[29] H. Laanaya, A. Martin, D. Aboutajdine, A. Khenchaf, Support vector regression of membership functions and belief functions–application for pattern recognition, Information Fusion 11 (4) (2010) 338–350.

[30] Z.-G. Liu, Q. Pan, J. Dezert, A new belief-based k-nearest neighbor classification method, Pattern Recognition 46 (3) (2013) 834–844.

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

[32] Z.-G. Liu, Q. Pan, J. Dezert, Evidential classifier for imprecise data based on belief functions, Knowledge-Based Systems 52 (2013) 246–257.

[33] R. R. Yager, On the Dempster-Shafer framework and new combination rules, Information sciences 41 (2) (1987) 93–137.

[34] A. P. Dempster, Upper and lower probabilities induced by a multivalued mapping, The annals of mathematical statistics (1967) 325–339.

[35] P. Smets, The combination of evidence in the transferable belief model, Pattern Analysis and Machine Intelligence, IEEE Transactions on 12 (5) (1990) 447–458.

[36] J.-B. Yang, D.-L. Xu, Evidential reasoning rule for evidence combination, Artificial Intelligence 205 (2013) 1–29.

[37] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, The Journal of Machine Learning Research 3 (2003) 1157–1182.

[38] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial intelligence 97 (1) (1997) 273–324.

[39] A. L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artificial intelligence 97 (1) (1997) 245–271.

[40] G. Forman, An extensive empirical study of feature selection metrics for text classification, The Journal of machine learning research 3 (2003) 1289–1305.

[41] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (8) (2005) 1226–1238.

[42] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern recognition letters 15 (11) (1994) 1119–1125.

[43] S. Nakariyakul, D. P. Casasent, An improvement on floating search algorithms for feature subset selection, Pattern Recognition 42 (9) (2009) 1932–1940.

[44] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine learning 46 (1-3) (2002) 389–422.

[45] J. Weston, A. Elisseeff, B. Schölkopf, M. Tipping, Use of the zero norm with linear models and kernel methods, The Journal of Machine Learning Research 3 (2003) 1439–1461.

[46] K. Deb, An efficient constraint handling method for genetic algorithms, Computer methods in applied mechanics and engineering 186 (2) (2000) 311–338.

[47] K. Deep, K. P. Singh, M. Kansal, C. Mohan, A real coded genetic algorithm for solving integer and mixed integer optimization problems, Applied Mathematics and Computation 212 (2) (2009) 505–518.

[48] S. Perkins, K. Lacker, J. Theiler, Grafting: Fast, incremental feature selection by gradient descent in function space, The Journal of Machine Learning Research 3 (2003) 1333–1356.

[49] A. Frank, A. Asuncion, Uci machine learning repository, 2010, URL http://archive. ics. uci. edu/ml 15 (2011) 22.

[50] P. Somol, J. Novovicova, Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality, Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (11) (2010) 1921–1939.

[51] C. M. Bishop, Neural networks for pattern recognition, Oxford University Press, 1995.

List of Figures

1 Flowchart of mass function construction. Mass functions m^{Γ_q}, dm^{Γ_q} for $q = 1, \ldots, c$ and m_t are calculated by Equations 8 to 10. 32

2 Test of the two-step classification strategy on a synthetic dataset; (a) shows training and test samples; (b) and (c) are credal partition obtained, respectively, by the EK-NN classifier and the two-step classification rule. The blue, green and black points represent instances with highest mass function on $\{\omega_1\}$, $\{\omega_2\}$ and Ω respectively; (d)-(f) are classification results obtained, respectively, by EK-NN, the proposed Dempster+Yager combination and the two-step classification strategy; the magenta stars represent misclassification instances. The calculated error rates for (d)-(f) are, respectively, 9.80%, 8.80% and 7.80% (color version is suggested). 33

Table 1: Combination result with different rules in Example 1.

	#1	#2	#3	Dempster's rule	Yager's rule	Dempster+Yager rule
$m(\{\omega_1\})$	0.8	0.8	0	0.8276	0.1920	**0.7680**
$m(\{\omega_1\})$	0	0	0.8	0.1379	0.0320	**0.0080**
$m(\Omega)$	0.2	0.2	0.2	0.0345	0.7760	**0.2240**

Table 2: Combination result with different rules in Example 2.

	#1	#2	#3	#4	Dempster's rule	Yager's rule	Dempster+Yager rule
$m(\{\omega_1\})$	0.8	0.8	0	0	0.4898	**0.0384**	**0.0384**
$m(\{\omega_1\})$	0	0	0.8	0.8	0.4898	**0.0384**	**0.0384**
$m(\Omega)$	0.2	0.2	0.2	0.2	0.0204	**0.9232**	**0.9232**

Table 3: Cardinality of selected feature subsets for synthetic data 1, and comparison of classification error (in %) between selected feature subset (with fs) and all features (without fs). Here n_r, n_c and n_i represent the number of relevant, redundant and irrelevant features, respectively.

n_r	n_c	n_i	subset cardinality	EK-NN error	two-step classification error	
					without fs	with fs
2	2	6	**2**	14.67	12.67	**2.67**
2	2	16	**2**	17.33	12.00	**1.33**
2	2	26	**2**	23.33	18.67	**4.00**
2	2	36	**2**	28.67	26.67	**5.33**
2	2	46	**2**	29.33	23.33	**4.67**

Table 4: Cardinality of selected feature subsets for synthetic data 2, and comparison of classification error (in %) between selected feature subset (with fs) and all features (without fs). Here n_r, n_c and n_i represent the number of relevant, redundant and irrelevant features, respectively. The number of relevant features here is two (i.e., $n_r = 2$).

n_c	n_i	subset cardinality	EK-NN error	two-step classification error	
				without fs	with fs
2	6	**2**	21.33	12.00	**8.67**
2	16	**2**	34.67	26.00	**14.67**
2	26	**2**	31.33	27.33	**16.00**
2	36	**2**	52.67	37.33	**11.33**
2	46	**2**	50.00	39.33	**8.00**

Table 5: Influence of parameter η on the proposed method.

η	0	0.5	1	1.5	2
error rate (%)	11.03	**10.94**	11.26	11.27	11.27

Table 6: Briefly description of the seven real datasets used in our experiments.

data set	number of classes	number of features	number of instances
Iris	3	4	150
Seeds	3	7	210
Wine	3	13	178
Yeast	3	8	1055
WDBC	2	30	569
Parkinsons	2	22	195
Lung Cancer	2	52	25

Table 7: Comparison of the proposed feature selection method with classical wrapper methods on seven real datasets. The proposed two-step classification was used to obtain average misclassify ratio. The robustness of selected feature subset is evaluated by the way proposed in [50].

	Iris			**Seeds**		
	Error(%)	Robustness(%)	Subset Size	Error(%)	Robustness(%)	Subset Size
All	2.67	n/a	4	7.62	n/a	7
SFS	4.67	54.55	1	11.90	57.97	2
SBS	5.33	21.05	2	10.95	23.88	3
SFFS	5.33	21.62	3	5.24	54.93	2
EFS*	***2.00***	***100***	3	***4.76***	***81.18***	3
	Wine			**Yeast**		
	Error(%)	Robustness(%)	Subset Size	Error(%)	Robustness(%)	Subset Size
All	13.04	n/a	13	38.87	n/a	8
SFS	30.50	75	1	61.99	***100***	1
SBS	6.24	42.47	5	48.35	***100***	1
SFFS	7.29	57.58	4	36.21	40	5
EFS*	***5.13***	***91.89***	3	***32.51***	***100***	2
	WDBC			**Parkinsons**		
	Error(%)	Robustness(%)	Subset Size	Error(%)	Robustness(%)	Subset Size
All	7.20	n/a	30	13.37	n/a	22
SFS	14.44	80	1	15.82	33.33	1
SBS	19.67	22.22	2	19.03	23.91	2
SFFS	9.87	25	4	13.79	43.65	3
EFS*	***5.80***	***92.37***	3	***8.63***	***100***	3
	Lung Cancer+					
	Error(%)	Robustness(%)	Subset Size			
All	32.00	n/a	52			
SFS	16.00	78.64	2			
SBS	36.00	32.76	9			
SFFS	28.00	94.27	2			
EFS*	***0***	***97.92***	4			

Table 8: Misclassification rates (in %) of the proposed method and six other classifiers with sequential floating forward feature selection (SFFS). For BK-NN and CCR, R_e and R_i represent, respectively, the error and imprecision rates.

		Iris	Seeds	Wine	Yeast	WDBC	Parkinson	Lung Cancer$^+$
SFFS +	ANN	8.00	7.62	9.64	32.57	9.15	9.63	16.00
	CART	8.00	7.14	9.09	37.55	10.04	11.21	16.00
	SVM	6.00	7.14	6.83	36.14	8.28	13.26	16.00
	EK-NN	5.33	6.67	6.18	35.07	9.70	16.39	24.00
	BK-NN	4.00	2.38	6.74	16.31	7.22	9.18	24.00
	(R_e, R_i)	4.67	11.90	5.13	40.84	8.44	13.37	0
	CCR	4.00	3.81	3.99	19.53	5.99	16.42	24.00
	(R_e, R_i)	4.67	18.57	15.33	36.11	15.83	12.26	4.00
our method		**2.00**	**4.76**	**5.13**	**32.51**	**5.80**	**8.63**	**0**

Table 9: Evaluating the feasibility of proposed feature selection procedure (EFS) for different classifiers. The classification error rate obtained by all features is compared with that obtained by selected feature subsets.

	Iris			**Seeds**		
	err without fs(%)	err with fs(%)	EFS feasibility	err without fs(%)	err with fs(%)	EFS feasibility
ANN	6.00	5.33	√	***4.76***	***4.76***	√
CART	7.33	7.33	√	7.14	7.62	**X**
SVM	6.00	4.67	√	6.19	5.24	√
EK-NN	4.67	4.00	√	10.00	5.71	√
BK-NN	(2.67,5.33)	(2.00,4.67)	√	(4.76,13.33)	(3.33,10.00)	√
CCR	(4.67,2.00)	(2.67,3.33)	√	(6.67,8.10)	(10.48,6.19)	**X**
two-step	2.67	***2.00***	√	7.62	***4.76***	√
	Wine			**Yeast**		
	err without fs(%)	err with fs(%)	EFS feasibility	err without fs(%)	err with fs(%)	EFS feasibility
ANN	***3.34***	6.18	**X**	36.42	33.84	√
CART	9.52	6.71	√	36.32	36.78	**X**
SVM	12.68	5.60	√	34.33	***32.71***	√
EK-NN	25.75	4.45	√	36.02	37.05	**X**
BK-NN	(26.30,17.61)	(2.19,6.22)	√	(15.71,42.09)	(16.95,40.77)	√
CCR	(3.47,0)	(3.93,5.07)	**X**	(21.68,32.29)	(31.66,8.82)	√
two-step	10.15	3.41	√	37.34	33.08	√
	WDBC			**Parkinsons**		
	err without fs(%)	err with fs(%)	EFS feasibility	err without fs(%)	err with fs(%)	EFS feasibility
ANN	***4.75***	6.32	**X**	11.27	12.35	**X**
CART	7.90	7.56	√	15.07	11.82	√
SVM	10.03	6.33	√	19.54	11.38	√
EK-NN	6.50	5.98	√	14.37	10.69	√
BK-NN	(10.70,20.22)	(3.69,7.73)	√	(17.88,18.95)	(5.58,15.35)	√
CCR	(8.09,1.40)	(4.19,15.01)	√	(22.15,0)	(17.49,8.58)	√
two-step	7.01	5.28	√	12.85	***9.11***	√
	Lung Cancer$^+$					
	err without fs(%)	err with fs(%)	EFS feasibility			
ANN	32.00	8.00	√			
CART	24.00	12.00	√			
SVM	24.00	***0***	√			
EK-NN	28.00	4.00	√			
BK-NN	(50.00,28.00)	(4.00,0)	√			
CCR	(16.00,12.00)	(4.00,20.00)	√			
two-step	32.00	***0***	√			

Table 10: Classification error rates of different methods using the same feature subsets selected by the proposed selection procedure.

	Iris	Seeds	Wine	Yeast	WDBC	Parkinsons	Lung Cancer^{+}
ANN	6.23	8.62	7.07	35.09	6.52	13.92	8.00
CART	5,50	11.70	8.22	37.76	8.07	16.75	12.00
SVM	3.78	9.72	5.71	33.68	6.47	13.53	**0**
EK-NN	4.04	6.19	5.96	38.20	**5.71**	12.43	4.00
BK-NN	2.03	3.96	4.57	18.92	5.97	9.29	4.00
(R_e, R_i)	5.67	7.44	6.67	40.03	7.19	16.03	0
CCR	3.49	5.79	5.01	20.88	6.83	19.28	4.00
(R_e, R_i)	2.90	16.73	3.72	38.52	5.39	5.55	20.00
two-step	**2.52**	**4.94**	**4.42**	**32.97**	5.86	**12.37**	**0**

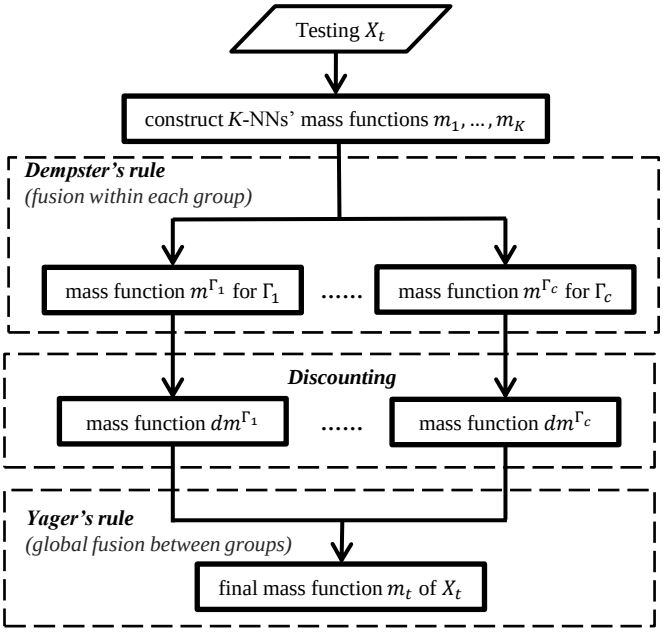

Figure 1: Flowchart of mass function construction. Mass functions m^{Γ_q}, dm^{Γ_q} for $q = 1, \ldots, c$ and m_t are calculated by Equations 8 to 10.

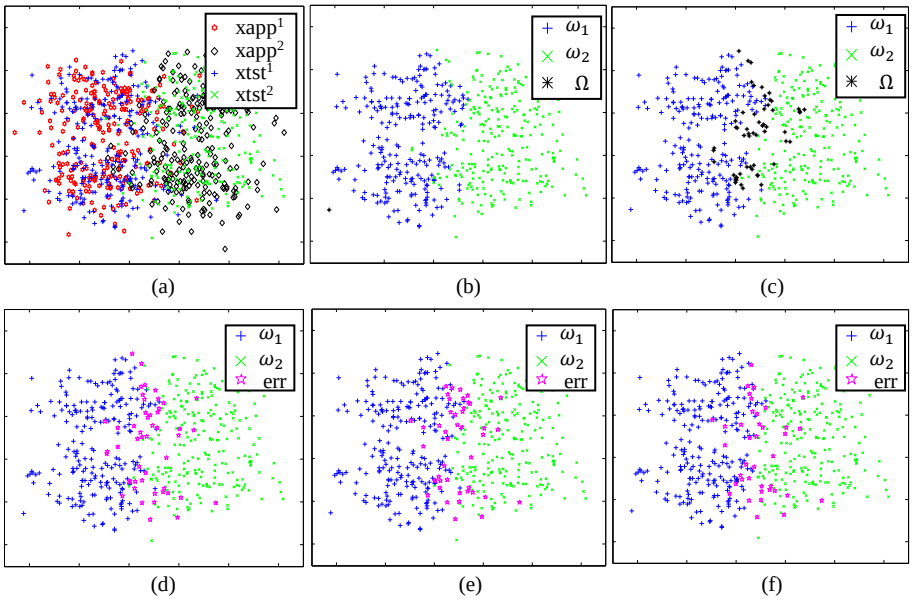

Figure 2: Test of the two-step classification strategy on a synthetic dataset; (a) shows training and test samples; (b) and (c) are credal partition obtained, respectively, by the EK-NN classifier and the two-step classification rule. The blue, green and black points represent instances with highest mass function on $\{\omega_1\}$, $\{\omega_2\}$ and Ω respectively; (d)-(f) are classification results obtained, respectively, by EK-NN, the proposed Dempster+Yager combination and the two-step classification strategy; the magenta stars represent misclassification instances. The calculated error rates for (d)-(f) are, respectively, 9.80%, 8.80% and 7.80% (color version is suggested).