Evidence Combination Based on Credal Belief Redistribution for Pattern Classification

Zhun-ga Liu, Yu Liu, Jean Dezert, Fabio Cuzzolin

***Abstract*—Evidence theory, also called belief functions theory, provides an efficient tool to represent and combine uncertain information for pattern classification. Evidence combination can be interpreted, in some applications, as classifier fusion. The sources of evidence corresponding to multiple classifiers usually exhibit different classification qualities, and they are often discounted using different weights before combination. In order to achieve the best possible fusion performance, a new Credal Belief Redistribution (CBR) method is proposed to revise such evidence. The rationale of CBR consists in transferring belief from one class not just to other classes but also to the associated disjunctions of classes (i.e., meta-classes). As classification accuracy for different objects in a given classifier can also vary, the evidence is revised according to prior knowledge mined from its training neighbors. If the selected neighbors are relatively close to the evidence, a large amount of belief will be discounted for redistribution. Otherwise, only a small fraction of belief will enter the redistribution procedure. An imprecision matrix estimated based on these neighbors is employed to specifically redistribute the discounted beliefs. This matrix expresses the likelihood of misclassification (i.e., the probability of a test pattern belonging to a class different from the one assigned to it by the classifier). In CBR, the discounted beliefs are divided into two parts. One part is transferred between singleton classes, whereas the other is cautiously committed to the associated meta-classes. By doing this, one can efficiently reduce the chance of misclassification by modeling partial imprecision. The multiple revised pieces of evidence are finally combined by Dempster-Shafer rule to reduce uncertainty and further improve classification accuracy. The effectiveness of CBR is extensively validated on several real datasets from the UCI repository, and critically compared with that of other related fusion methods.**

***Index terms*-** evidence theory, belief functions, pattern classification, discounting, classifier fusion.

I. Introduction

In a multi-source information fusion system, the fusion generally exploits the complementary knowledge provided by different sources to reduce uncertainty and improve accuracy for decision making support. Information fusion techniques can be broadly divided into three levels [1]: signal level, feature-level and decision-level fusion. In this work, we mainly focus on *decision-level fusion* methods [2], which can efficiently deal with the heterogeneous information sources probed by different sensors. At decision level such sources of information are typically mapped to probabilities, belief functions, fuzzy membership functions, or other uncertainty measures. Decision level information fusion can then be interpreted as classifier fusion, in which each information source corresponds to a classifier.

In this context, belief functions theory [3]–[6], also called evidence theory or Dempster-Shafer Theory (DST), provides an efficient tool to characterize and combine uncertain information [7], and has been widely applied to pattern classification [8]–[14], clustering [15]–[17] and information fusion [18]–[22], among others. In [9], for instance, an Evidential K-Nearest Neighbors (EK-NN) classifier is presented. Each neighbor provides a piece of classification evidence represented by a Basic Belief Assignment (BBA), and the resulting K BBAs are combined by DS rule for making the final class decision. In order to reduce the computational burden of EK-NN, an Evidential Neural Network (ENN) is further developed in [8], generally delivering good performance with relatively lower complexity. For combining classifiers in the DS framework, a general t-norm based combination rule is developed in [18] to deal with non-independent information. Such a rule ranges in behaviour between the DS and the cautious rules as a function of a parameter, which can be optimized based on training data. In [19], a class-indifferent method is developed to combine classifiers based on DST, in which a classifier's output is represented by evidential structures in terms of triplets and quadruplets. In order to achieve good fusion performance, the method prioritizes the class decisions to be combined, and employs an 'ignorance' element to model unknown information.

Information sources corresponding to multiple classifiers may have different reliabilities, which can be represented using different weighting factors in the fusion procedure. Reliability evaluation then plays an important role in weighted fusion methods [30]. In the multi-classifier fusion methods [23], [31], the weight of each classifier is often determined according to its training accuracy [24]. The higher the accuracy, the bigger the weight. Some often used weighted combination methods are introduced in [24], including simple, re-scaled, best-worst and quadratic best-worst weighted rule, and weighted majority rule. Weights can also be automatically learned using the training data [25], [26], and the optimal weight choice corresponds to the best fusion result under a certain criterion, e.g. minimal distance between fusion result and ground truth. When prior, training knowledge is not available, the degree of conflict among different sources of evidence can also be used to estimate the weight of each piece of evidence [27]. If one source exhibits large conflict with the others, this source will be considered not very reliable, and weighted less. Pieces

Zhun-ga Liu is with School of Automation, Northwestern Polytechnical University, Xi'an, China. Yu Liu is with Research Institute of Information Fusion, Naval Aeronautical and Astronautical University, Yantai, China. Jean Dezert is with ONERA - The French Aerospace Lab, F-91761 Palaiseau, France. Fabio Cuzzolin is with the School of Engineering, Computing and Mathematics, Oxford Brookes University, Oxford, UK.

Zhun-ga Liu is corresponding author, Email: liuzhunga@nwpu.edu.cn

of evidence consistent with the others will be assigned larger weights [28]. Conflict can be measured based on evidential distance or other similar metrics [29], [32], [33].

Once the weight of each classifier output (i.e., each basic belief assignment) is determined, weighted combination methods can be applied. In the traditional weighted averaging rule, the mean of the different pieces of evidence with different weights is calculated [34]. In Shafer's classical evidence discounting operation [3], part of the belief on each class is assigned to the whole universe of possible classes ('discounted') in proportion to the weight of the corresponding evidence, and the 'ignorance' element (the whole universe or 'frame', in Shafer's approach) plays a neutral role in the fusion process. Such traditional discounting strategies are mainly used to tune the influence of each source on the fusion result, but they cannot improve the accuracy of each source, because the belief/probability of each class is proportionally decreased or increased in the discounting procedure. As a response, contextual discounting [35] has been developed, as an extension of classical discounting, to take into account more refined reliability knowledge. A discount rate vector is used to represent the degrees of belief in sensor reliability in different conditions. Such a vector can be learnt by error minimization in the labeled data space. In a previous work [26], some of us have also developed a weighted evidence combination method for multiple classifiers. In [26], an iterative optimization strategy is presented to seek the proper weighting factors (including classifier weight, pattern weight, etc) by minimizing the distances between combination results and ground truth over the whole training set. For decision making support, a cautious rule is introduced. Any pattern hard to classify in terms of fusion results is assigned to a set of classes, as partial imprecision open to be refined by other techniques is considered preferable to definite errors in some applications. In [26], reliability is estimated using the entire training data, and classification results produced by the same classifier on different patterns have the same reliability.

For a given classifier, however, classification results for different patterns may also have different reliabilities. For example, patterns in the center of the region spanned by a class are usually more accurately classified than those lying in areas where several classes overlap. In [36], a contextual reliability evaluation method is developed for classifier fusion based on the concepts of 'inner' and 'relative' reliability. The inner reliability reflecting the quality of each classifier is evaluated based on the neighborhoods of each pattern, and it can be used to revise the classifier output by discounting some beliefs in favour of partial ignorance. Relative reliability is calculated based on an incompatibility measure among classifiers. Classical evidence discounting is applied using relative reliability to reduce conflict. Dempster's rule is finally employed to combine the discounted classifier outputs for final classification. Unfortunately, pattern attribute information may be unavailable in some applications, in which the fusion core only has at disposal decision-level knowledge coming from different sources. Moreover, the neighborhoods used to evaluate inner reliability may or may not represent well the test pattern to classify. Therefore, inner reliability so defined cannot be completely trusted for discounting purposes.

A. Contributions

In order to achieve the best possible fusion performance, a new Credal Beliefs Redistribution (CBR) method is proposed for combining classifiers. In CBR, the quality of each classifier output, represented by a BBA $\mathbf{m}$, is evaluated based on its training neighbors. If these neighbors are very close to $\mathbf{m}$, this indicates that they are likely to provide important prior knowledge for BBA correction, and a large proportion of belief is redistributed. Otherwise, the proportion of belief entering the redistribution process is small. Belief is redistributed based on an imprecision matrix Φ, which is estimated based on the local structure of the neighborhood. Entry $\phi_{i,j}$ in Φ represents the prior probability of the pattern to belong to class ω_i when classified as ω_j by the classifier at hand, thus encoding the degree to which the classifier is undecided/confused between classes ω_i and ω_j.

As part of the proposed Credal Beliefs Redistribution approach, both an 'optimistic' and a 'pessimistic' redistribution strategy are presented. As a function of the imprecision matrix, belief can be directly transferred among different singleton classes following the optimistic strategy. CBR also allows belief to be transferred from a single class (e.g., ω_i) to a meta-class containing it (e.g., the disjunction of two classes $\omega_i \cup \omega_j$) in a more pessimistic/cautious way. A balance parameter allows CBR to provide a reasonable trade-off between preserving the specificity of the revised BBA and mitigating the risk of mistaken BBA revisions whenever the imprecision matrix is not completely reliable.

B. Paper outline

This paper is organized as follows. The basics of evidence theory are briefly introduced in next Section. The Credal Belief Redistribution method is detailed in Section III. In section IV, suitable experiments are presented to test CBR's performance, in comparison with the state of the art. Conclusions are provided in section V.

II. Basics of evidence theory

Evidence theory [3] is a general theory for reasoning under uncertainty, based on the notion of *belief function*. Belief functions assume a *frame of discernment* $\Omega = \{\omega_1, \ldots, \omega_c\}$, which consists of a finite discrete set of mutually exclusive and exhaustive hypotheses – in our case, these correspond to the possible classes.

A *Basic Belief Assignment* (BBA) is defined over the power set 2^Ω of Ω, which collects all the subsets of Ω. For example, if $\Omega = \{\omega_1, \omega_2, \omega_3\}$, then $2^\Omega = \{\emptyset, \omega_1, \omega_2, \omega_3, \omega_1 \cup \omega_2, \omega_1 \cup \omega_3, \omega_2 \cup \omega_3, \Omega\}$. A BBA is then a function $m(.)$ from 2^Ω to $[0, 1]$, which satisfies the following constraints:

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

A is called a *focal element* of $m(.)$ if $m(A) > 0$. When the cardinality $|A|$ of A is greater than 2 ($|A| \geq 2$), A is

called *meta-element* here. A *Bayesian* BBA is such that all its focal elements are singletons elements of the frame of discernment. In this case, the Bayesian BBA is homogeneous to a (subjective) probability measure. If the focal elements of a BBA just consist of the singleton elements and of the 'ignorance' element Ω, the BBA is called *simple*.

In a classification context $m(A)$, the 'mass' of A, can be interpreted as the degree of classification of the object at hand associated with A. If $|A| \geq 2$ (e.g. $A = \omega_i \cup \omega_j$), $m(A)$ also reflects a classifier's degree of confusion among elements of A (e.g. ω_i, ω_j). The quantity $m(\Omega)$ measures total ignorance, i.e., the reluctance to commit the object to any particular class.

Please note that multi-label learning, the situation in which the object to classify can simultaneously belong to multiple classes (i.e., a set of classes) is not considered in this work. In this paper it is assumed that the object truly belongs to only one class, and that a set of classes (a meta-element, or meta-class) reflects imprecision in the classification under the framework of belief functions. Any meta-element A implies that the object is likely to belong to one of the classes in A, but that the true class of the object cannot be resolved.

The *belief function* $Bel(.)$ and *plausibility function* $Pl(.)$ associated with a BBA $m(.)$ are defined as

$$Bel(A) = \sum_{B \subseteq A} m(B); \quad Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \quad \forall A \in 2^{\Omega}, \tag{2}$$

and correspond to lower and upper bounds to the probability of the events A, respectively.

Multiple sources of evidence each represented by a BBA can be combined by *Dempster's rule* (here called DS rule for short). The DS combination of $\mathbf{m}_1, \ldots, \mathbf{m}_n$ is defined by

$$m_{DS}(A) = [\mathbf{m}_1 \oplus \ldots \oplus \mathbf{m}_n](A) = \frac{\sum_{\bigcap_{i=1}^{n} B_i = A} \prod_{i=1}^{n} m_i(B_i)}{1 - \sum_{\bigcap_{i=1}^{n} C_i = \emptyset} \prod_{i=1}^{n} m_i(C_i)} \tag{3}$$

for $A \neq \emptyset, B_i, C_i \in 2^{\Omega}$, whereas $m_{DS}(\emptyset) = 0$ for $A = \emptyset$.

DS rule is associative. Hence, multiple BBAs can be combined one by one, and the order of the combination has no influence on the final result. DS rule has been widely applied, but it should be approached with caution when the sources of evidence are highly in conflict, and in some special case even when conflict is low [37].

When the sources of evidence have different reliabilities, they can be discounted using their reliability value to reduce their weight in the fusion process. Given a BBA $\mathbf{m}$ with reliability (discounting factor) equal to α, classical discounting [3], due to Shafer, yields:

$$\begin{cases} {}^{\alpha}m(A) = \alpha \cdot m(A) \quad A \subset \Omega, A \neq \Omega, \\ {}^{\alpha}m(\Omega) = 1 - \alpha + \alpha \cdot m(\Omega). \end{cases} \tag{4}$$

It can be seen from Eq. (4) that the discounted beliefs are all committed to the ignorance element Ω in proportion to the discounting factor. The smaller the reliability value α, the higher the degree of ignorance. If $\alpha = 1$, this BBA is considered as completely reliable, and is not altered by discounting. If $\alpha = 0$, the mass of all its focal elements is reassigned to Ω, and the discounted BBA is the 'vacuous' belief function such that ${}^{\alpha}m(\Omega) = 1$, which plays a neutral role in DS combination.

Such discounting operation is mainly used to tune the influence of each source of evidence during fusion, but it cannot improve its accuracy. Consider, for example, a source of evidence committing maximum belief to a specific class ω_i, whereas the true class for the pattern at hand is $\omega_j, j \neq i$. After traditional discounting (4), the maximum mass among all singleton classes remains on ω_i (rather than on ω_j), no matter what the discounting factor is.

III. Credal belief redistribution method for pattern classification

Decision-level information fusion is widely used for target (pattern) classification [2]. In some applications, the fusion center just receives multiple independent information sources expressing the probabilities/beliefs of the object belonging to each class, and the original observation collecting the pattern's attributes is not available. Such information fusion system can be seen as combining multiple classification results for pattern classification.

Evidence theory as an important decision-level fusion method has been applied to classifier fusion for dealing with uncertain information [18], [19], with each classifier corresponding to one source of evidence. Classifier output can typically be represented by a probability (Bayesian BBA) or a simple BBA, including one additional 'ignorance' class. Therefore, in this work we mainly consider the combination of classification results in form of Bayesian or simple BBAs.

Classifiers (information sources) are usually diverse in nature. For instance, they might reach their classification results based on different attributes. Such diversity can provide complementary knowledge which is beneficial for fusion purposes. However, significant diversity is likely to cause high conflict among classifiers, leading different classifiers to make different decisions. In real applications, the classifiers to be combined generally have different 'qualities' (reliabilities).

In order to improve classification accuracy, we propose to estimate an imprecision matrix to characterize in a rather detailed, sophisticated way the reliability of a specific classification result, for a specific test pattern. The imprecision matrix is mined from the local structure of the training data around the test object, and then used to revise each classifier's output via Credal Belief Redistribution (CBR) before combination. This new method allows us to transfer belief from one class to either other singleton classes or to meta-classes. In such a way, the chance of poor classification results is mitigated by properly modeling partial imprecision, with the chance of it being further refined/clarified by combination with other classifiers carrying complementary knowledge.

A. Notation and setting

Let us consider multiple classifiers $\mathcal{C}_1, \ldots, \mathcal{C}_n$ learnt from different attribute data. The outputs of these classifiers need

to be combined for pattern classification over the frame of discernment $\Omega = \{\omega_1, \ldots, \omega_c\}$ containing all possible classes.

Assume a training set of s labelled patterns is available. For each classifier $\mathcal{C}_l$, $l = 1, ..., n$, the result of the classification of pattern $\mathbf{x}_i$, $i = 1, ..., s$, is denoted by $\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,c})$, where $p_{i,j} \triangleq p_i(\omega_j)$ represents the probability of $\mathbf{x}_i$ being committed to class ω_j. The true classification results on these s patterns are expressed by the vectors $\mathbf{t}_1, \ldots, \mathbf{t}_s$, such that $t_i(\omega_j) = 1$ and $t_i(\omega_g) = 0, g \neq j$ whenever the true class label of $\mathbf{x}_i$ is ω_j. For each classifier the outputs $\mathbf{p}_i, i = 1, \ldots, s$ generated in correspondence to these s patterns are considered as training data, whereas the patterns themselves $\mathbf{x}_i, i = 1, \ldots, s$ (the vectors of attribute values) are not used in the fusion.

Given a test pattern $\mathbf{y}$, the probabilistic output of each classifier $\mathcal{C}_l$, $l = 1, \ldots, n$ can be represented by a (Bayesian) BBA $\mathbf{m}_l, l = 1, \ldots, n$. The final classification of $\mathbf{y}$ will be determined by the combination of these BBAs. Since, as we said, each classifier is, in general, characterized by a different reliability level on the given classification task, and classifier accuracy may also vary across patterns, it is crucial to carefully evaluate the quality of the classification result for each target.

B. Estimation of the imprecision matrix

Let us see how to assess the quality of the classification result $\mathbf{m}_l$ produced by classifier $\mathcal{C}_l$, $l = 1, \ldots, n$. This is done by evaluating the nearest neighbors of $\mathbf{m}_l$ in the space of the training data (the outputs generated by the training patterns).

First, the K nearest neighbors (K-NN) of $\mathbf{m}_l$ are located, and denoted by $\mathbf{p}_1, \ldots, \mathbf{p}_K$, the corresponding ground truth vectors being $\mathbf{t}_1, \ldots, \mathbf{t}_K$. If the K selected neighbors are quite close to $\mathbf{m}_l$, they may provide important prior knowledge for quality evaluation, according to which a large proportion of belief originally committed by $\mathbf{m}_l$ will be redistributed. If, however, $\mathbf{m}_l$'s neighbors are not so close to it, the amount of belief to be redistributed will be small. Namely, the fraction of belief entering the redistribution procedure depends on the distances between $\mathbf{m}_l$ and its neighbors so defined. The larger the distance, the smaller the proportion and vice-versa. This fraction α_l, called *discounting factor*, is computed as:

$$\alpha_l = e^{-\gamma_l \delta_l}, \tag{5}$$

with

$$\delta_l = \frac{\sum_{k=1}^{K} d_{lk}}{K\bar{d}}; \quad \bar{d} = \frac{1}{Ks} \sum_{i=1}^{s} \sum_{k=1}^{K} d_{ik},$$

where $d_{lk} = \|\mathbf{m}_l - \mathbf{p}_k\|$ is the Euclidean distance between $\mathbf{m}_l$ and its neighbor $\mathbf{p}_k$, d_{ik} is the distance between the training data $\mathbf{p}_i$ and its k_{th} nearest neighbor.

The quantity δ_l can be computed step by step as follows. Given $\mathbf{m}_l$ and the s training datapoints $\mathbf{p}_i, i = 1, \ldots, s$, we can find the K nearest neighbors of $\mathbf{m}_l$ and those of $\mathbf{p}_i, i = 1, \ldots, s$ in the training data space, respectively. The average distance between $\mathbf{m}_l$ and its K neighbors is calculated as $\bar{d}_l = \sum_{k=1}^{K} d_{lk}/K$. The average distance between $\mathbf{p}_i$ and its K-NN is similarly calculated as $\bar{d}_i = \sum_{k=1}^{K} d_{ik}/K$. The mean value of the average distances from each training data to its K nearest neighbors is finally given by $\bar{d} = \frac{1}{s} \sum_{i=1}^{s} \bar{d}_i$. Summarising, δ_l is thus the ratio between the mean distance of $\mathbf{m}_l$ to its K nearest neighbors and the mean distance between training data. This normalized distance measure is employed here to reduce the influence of the value of K and of data dispersion on the quality evaluation process.

Given the discounting factor (5), the discounted beliefs to be redistributed are given by:

$$\mathbf{m}_{l1} = \alpha_l \mathbf{m}_l. \tag{6}$$

The remaining fraction of belief is committed to each class as in the original BBA, namely:

$$\mathbf{m}_{l2} = (1 - \alpha_l)\mathbf{m}_l. \tag{7}$$

In classical discounting, the discounted beliefs are all committed to the total ignorance element Ω as shown in Eq. (4). Such operation is mainly used to tune the influence of each piece of evidence in the combination, but it has no effect on its accuracy. In opposition, useful prior knowledge can be mined from a BBA's neighbors to revise the evidence it represents via belief redistribution. More precisely, if the majority of the K neighbors of BBA $\mathbf{m}_l$ with ground truth label ω_j are strongly committed to ω_i by a classifier, this indicates that the prior probability that the pattern truly belongs to ω_j, while it is actually classified as ω_i, is high. Thus, the weight of the hypothesis that the target is classified as ω_i while it truly belongs to class ω_j can be calculated by:

$$w_{j,i} = \sum_{t_k(\omega_j)=1} e^{-\lambda_l \tilde{d}_k} p_k(\omega_i), \tag{8}$$

where $\tilde{d}_k = \frac{\|\mathbf{m}_l - \mathbf{p}_k\|}{\min_k \|\mathbf{m}_l - \mathbf{p}_k\|}$ and λ_l is a parameter distinct for each classifier l. The quantity $\tilde{d}_k$ is a measure of relative distance, with as reference the minimum distance of $\mathbf{m}_l$ to its nearest neighbors. In (8), $e^{-\lambda_l d_k}$ is a distance penalizing factor which tunes the importance of each neighbor in the quality evaluation process. Neighbors far from the test pattern $\mathbf{y}$, quite reasonably, play a small role in the evaluation. This allows quality evaluation to be robust to the choice of K, since neighbors far from the test target (and therefore with a rather small distance penalizing factor attached) will have a very moderate influence. The K neighbors are found from the training data set, and the true class labels of the K neighbors (e.g. $t_k(.)$) are given as the prior training knowledge.

Finally, the prior probability of the target belonging to $\omega_j, j = 1, \ldots, c$ when it is classified as $\omega_i, i = 1, \ldots, c$ can be easily obtained via normalization:

$$\phi_{i,j} = \frac{w_{j,i}}{\sum_g w_{g,i}}, \tag{9}$$

so that one has $\sum_{j=1}^{c} \phi_{i,j} = 1$.

The prior probability $\phi_{i,j}$ reflects the degree of imprecision of the classification related to the pair of classes ω_i and ω_j,

so that we can term the matrix $\Phi_l = [\phi_{i,j}]_{c\times c}$ the *imprecision matrix* attached to the classification result $\mathbf{m}_l$. As explained, the imprecision matrix is estimated by applying K-NN to the classification result, and different classification outcomes correspond to different imprecision matrices, all computed in the same manner.

C. Belief redistribution with imprecision matrix

The imprecision matrix provides prior classification knowledge in a rather sophisticated form, which can be taken into account for correcting the associated classification result via proper belief redistribution. Assuming that the imprecision matrix Φ_l is completely reliable when it comes to the considered target, the mass values redistributed to $\omega_j, j = 1, \ldots, c$ can be computed as:

$$\tilde{m}_{l1}(\omega_j) = \sum_{i=1}^{c} \phi_{i,j} m_{l1}(\omega_i). \tag{10}$$

This can be rewritten in matrix form as:

$$\tilde{\mathbf{m}}_{l1} = \Phi_l^T \mathbf{m}_{l1}. \tag{11}$$

Eq. (10) clearly shows how masses are redistributed among the different singleton classes thanks to the imprecision matrix. In particular, one imprecision matrix always exists that can map the original BBA to the actual ground truth. This is shown by Lemma 1 below.

Consider, however, that Φ_l is estimated based on the neighbors of the classification results $\mathbf{m}_l$. Since such neighbors may well differ from $\mathbf{m}_l$ quite substantially, the resulting imprecision matrix may be unreliable for the purpose of revising $\mathbf{m}_l$. An 'unreliable' imprecision matrix is one that may potentially change the BBA to one that completely conflicts with the ground truth, severely harming the classification process. This point is also a corollary of Lemma 1.

Lemma 1. Let us consider a Bayesian BBA $\mathbf{m}$ defined over the frame of discernment $\Omega = \{\omega_1, \ldots, \omega_c\}$. Let $\tilde{\mathbf{m}}$ denote the modified version of $\mathbf{m}$ after application of an imprecision matrix Φ. It must hold that for all $j = 1, ..., c$ $\exists \Phi$ satisfying $\tilde{\mathbf{m}} = \Phi^T \mathbf{m}, \tilde{m}(\omega_j) = 1, \tilde{m}(\omega_i) = 0, j \neq i$.

Proof. According to Eq. (11) one has $\tilde{m}(\omega_j) = \sum_{i=1}^{c} \phi_{i,j} m(\omega_i)$. When $\phi_{i,j} = 1$, $i = 1, \ldots, c$, one gets:

$\tilde{m}(\omega_j) = \sum_{i=1}^{c} m(\omega_i) = 1$ and $\tilde{m}(\omega_i) = 0, j \neq i$.

Thus, there exists one imprecision matrix Φ with $\phi_{i,j} = 1, i = 1, \ldots, c$ that can make all mass focus on class ω_j.

D. Cautious redistribution

To avoid the issue of unreliable imprecision matrices, we illustrate an alternative cautious redistribution approach. There, the probability $\phi_{i,j}$ is considered to represent the classification's degree of imprecision concerning ω_i and ω_j for the target at hand. In response, the masses $m_{l1}(\omega_i)$ are transferred to the meta-class $\omega_i \cup \omega_j, j = 1, \ldots, c$ (i.e., the disjunction of the considered pair of classes) rather than to the singleton class ω_j, proportionally to $\phi_{i,j}, j = 1, \ldots, c$. Any individual meta-class, e.g. $\omega_i \cup \omega_j$, receives mass transferred from both the involved singleton classes ω_i and ω_j, namely:

$$\begin{cases} \phi_{j,i} m_{l1}(\omega_j) \rightarrow \tilde{m}_{l1}(\omega_j \cup \omega_i); \\ \phi_{i,j} m_{l1}(\omega_i) \rightarrow \tilde{m}_{l1}(\omega_i \cup \omega_j), \end{cases} \tag{12}$$

so that the total mass committed to the meta-class $\omega_i \cup \omega_j$ is:

$$\tilde{m}_{l1}(\omega_i \cup \omega_j) = \phi_{i,j} m_{l1}(\omega_i) + \phi_{j,i} m_{l1}(\omega_j). \tag{13}$$

In Eq. (13), $\tilde{m}_{l1}(\omega_i \cup \omega_j)$ represents the degree of indistinguishability/imprecision associated with the pair of classes ω_i and ω_j when classifying the test pattern at hand.

This is a very cautious belief redistribution method. As an effect of this revision the plausibility value $Pl(.)$ (cfr. Eq. (2)) of each class increases, whereas its belief value $Bel(.)$ decreases. When interpreting belief and plausibility functions as lower and upper bounds to the value of an unknown, 'true' probability measure, it follows that the uncertainty interval $[Bel(.), Pl(.)]$ associated with the classification result become wider. This is proved in Lemma 2.

Lemma 2. Let us consider one simple BBA $\mathbf{m}$ defined over the frame of discernment $\Omega = \{\omega_1, \ldots, \omega_c\}$. Assume $\mathbf{m}$ is revised as in Eq. (13) using an imprecision matrix Φ, and denote by $\tilde{\mathbf{m}}$ the revised BBA. It holds that $\tilde{Pl}(\omega_i) \geq Pl(\omega_i)$ and $\tilde{Bel}(\omega_i) \leq Bel(\omega_i), i = 1, \ldots, c$.

Proof. According to Eq. (2), since $\mathbf{m}$ is simple (it has only singletons and Ω as focal elements), we have that $Pl(\omega_i) = m(\omega_i) + m(\Omega)$. As for the revised BBA $\tilde{\mathbf{m}}$:

$$\begin{aligned} \tilde{Pl}(\omega_i) &= \sum_{\omega_i \in A} \tilde{m}(A) \\ &= \tilde{m}(\omega_i) + \sum_{i \neq j} \tilde{m}(\omega_i \cup \omega_j) + m(\Omega) \\ &= \phi_{i,i} m(\omega_i) + \sum_{i \neq j} [\phi_{i,j} m(\omega_i) + \phi_{j,i} m(\omega_j)] + m(\Omega) \\ &= \sum_{j=1}^{c} \phi_{i,j} m(\omega_i) + \sum_{i \neq j} [\phi_{j,i} m(\omega_j)] + m(\Omega). \end{aligned}$$

Since $\phi_{i,j} \in [0, 1]$ $\forall i = 1, \ldots, c, j = 1, \ldots, c$ and $\sum_{j=1}^{c} \phi_{i,j} = 1$, it follows that:

$$\begin{aligned} \tilde{Pl}(\omega_i) &= m(\omega_i) + \sum_{i \neq j} [\phi_{j,i} m(\omega_j)] + m(\Omega) \\ &\geq m(\omega_i) + m(\Omega) = Pl(\omega_i). \end{aligned}$$

Thus, one always has $\tilde{Pl}(\omega_i) \geq Pl(\omega_i)$.
Since $\mathbf{m}$ is simple, $Bel(\omega_i) = m(\omega_i)$. As for the revised BBA $\tilde{\mathbf{m}}$:

$$\tilde{Bel}(\omega_i) = \tilde{m}(\omega_i) = \phi_{i,i} m(\omega_i) \leq m(\omega_i),$$

and we have as desired.

Lemma 2 shows that the plausibility value of any (singleton) class does not decrease after cautious redistribution, no matter what the imprecision matrix is. When the imprecision matrix estimated based on the pattern's training neighbors is not very reliable (in the sense explained above), the proposed cautious redistribution is able to suppress the harmful influence of the

imprecision matrix on belief redistribution, and to increase its robustness. Nevertheless, this cautious behavior is inherently pessimistic, since it leads to higher imprecision which is undesirable for efficient classification.

E. Credal redistribution

A sensible conclusion is that, when revising a BBA via redistribution, it is not a good idea to exclusively adopt a single strategy, either the optimistic one described in Eq. (10), or the pessimistic one of Eq. (17), whereas it is best to strike a balance between the two approaches. To this extent, a *credal redistribution* method is introduced below, which allows belief to be transferred not only to singleton classes but also to meta-classes.

Namely, an additional parameter β_l is introduced to balance the assignment of belief to singleton classes in the optimistic strategy, and to meta-classes in the pessimistic one. In this way, the risk of classification error is reduced by properly modeling the partial imprecision of the classification process. As we mentioned, the imprecise information on the meta-classes can be reduced by combination with other classifiers. Any misclassification, however, is typically difficult to overcome via fusion. Credal redistribution, then, provides a good trade-off between classification accuracy and specificity.

Specifically, the masses committed to the singleton classes $\omega_i, i = 1, \ldots, c$ and meta-classes $\omega_i \cup \omega_j, i = 1, \ldots, c; j = 1, \ldots, c$ under credal redistribution can be determined, as a function of the fraction m_{l1} (6) of the original BBA, as:

$$\tilde{m}_{l1}(\omega_i \cup \omega_j) = \beta_l \big[\phi_{i,j} m_{l1}(\omega_i) + \phi_{j,i} m_{l1}(\omega_j)\big]; \quad (14)$$

$$\tilde{m}_{l1}(\omega_j) = \phi_{j,j} m_{l1}(\omega_j) + (1 - \beta_l) \sum_{i=1, i\neq j}^{c} \phi_{i,j} m_{l1}(\omega_i). \quad (15)$$

In the revised BBA, the total belief on the singleton class ω_j is then calculated by compounding Eq. (6) and (15) as:

$$\begin{aligned} \hat{m}_l(\omega_j) &= m_{l2}(\omega_j) + \tilde{m}_{l1}(\omega_j) \\ &= (1 - \alpha_l) m_l(\omega_j) + \phi_{j,j} m_{l1}(\omega_j) \\ &\quad + (1 - \beta_l) \sum_{i=1, i\neq j}^{c} \phi_{i,j} m_{l1}(\omega_i). \end{aligned} \quad (16)$$

Since no mass is committed to meta-classes in the original BBA, the masses of the meta-classes in the revised BBA can be directly obtained as:

$$\begin{aligned} \hat{m}_l(\omega_i \cup \omega_j) &= \tilde{m}_{l1}(\omega_i \cup \omega_j) \\ &= \beta_l \big[\phi_{i,j} m_{l1}(\omega_i) + \phi_{j,i} m_{l1}(\omega_j)\big]. \end{aligned} \quad (17)$$

The mass of the 'ignorance' class $m_l(\Omega)$ (if applicable) remains the same after the redistribution procedure.
Each classifier output $\mathbf{m}_1, \ldots, \mathbf{m}_n$ can be revised one by one in the way described above. Then, the revised BBAs $\hat{\mathbf{m}}_i, i = 1, \ldots, n$ are combined by DS rule (3) to yield:

$$\mathbf{m}^f = \hat{\mathbf{m}}_1 \oplus \hat{\mathbf{m}}_2 \ldots \oplus \hat{\mathbf{m}}_n, \quad (18)$$

where $\oplus$ denotes DS combination.

F. Influence of parameters

In the proposed credal redistribution method, three factors play a very important role: the discounting factor α, the imprecision matrix Φ and the balance factor β. A BBA can be suitably corrected by tuning these parameters. Here we briefly discuss their influence.

Assume that the test pattern to classify has true class label ω_i, and that the original classifier output $\mathbf{m}$ is corrected using an imprecision matrix Φ determined as a function of its training neighbors. If the output of the classifier on these neighbors is quite different from $\mathbf{m}$, the discounting factor α will be very small, indicating that a small fraction of the original mass is redistributed, with most of the original mass preserved for each class. In this situation, the tuning of Φ and β has little influence on the revision of $\mathbf{m}$. If the classifier's outputs for the selected neighbors are very consistent with $\mathbf{m}$, α will be big and a large fraction of the original belief will be redistributed.

According to credal redistribution, if the estimated imprecision matrix Φ is beneficial for correcting the classifier's output (i.e., the entries of Φ in the i_{th} column, $\phi_{j,i}, j = 1, ..., c$ are large), the largest share of mass can be assigned to the true class ω_i by Φ. In such a situation the balance parameter β should be small, so that most belief is transferred from the other singleton classes to the singleton class ω_i. In this way, a more specific and accurate BBA can be obtained. Finally, if the imprecision matrix Φ is not very reliable (i.e., the elements $\phi_{\cdot,j}, j \neq i$ are large where i is the true class of the target), using it may harm the revision of $\mathbf{m}$. In this case a large β value is required in order to redistribute belief to meta-classes of the form $\omega_i \cup \omega_j$. This reduces the risk of a misclassification error at the price of partial imprecision. Criteria for choosing proper values for these parameters are discussed in the remainder of the paper.

Let $P^{(l)} = [\mathbf{p}_1^{(l)}, ..., \mathbf{p}_s^{(l)}]$ be a matrix whose columns are the output of classifier $\mathcal{C}_l$ on the available training patterns $\mathbf{x}_1, ..., \mathbf{x}_s$.

Table I
CREDAL BELIEF REDISTRIBUTION

Input:	Given classifiers: $\mathcal{C}_l, l = 1, \ldots, n$
	Classifier output on training data: $P^{(1)}, \ldots, P^{(n)}$
	Classifier output on test data-point $\mathbf{y}$: $\mathbf{m}_l, l = 1, \ldots, n$
Par:	K: number of nearest neighbors
	$\lambda_l > 0$: distance penalizing weight
	$\gamma_l > 0$: discounting weight
	$\beta_l \in [0, 1]$: balance number
for	l=1 to n
	Select the K-nearest neighbors of $\mathbf{m}_l$ from $P^{(l)}$;
	Calculate the discounting factor α_l by Eq. (5);
	Determine the redistributed masses by Eqs. (6), (7);
	Estimate the imprecision matrix Φ_l by Eq. (9);
	Correct the classifier output $\mathbf{m}_l$ as in Eqs. (16), (17).
end	
	Combine corrected results for classification by Eq. (18).

For convenience of implementation, our credal belief redis-

tribution method for the combination of multiple classifiers is outlined in Table I, which assumes that all parameters are shared across the various classifiers.

G. Illustrative example

We think it is useful to provide a numerical example to illustrate more clearly how to implement the proposed credal belief redistribution method.

Let us consider a single source of evidence $\mathbf{m}$ defined over a frame of discernment $\Omega = \{\omega_1, \omega_2, \omega_3\}$, with $m(\omega_1) = 0.5$, $m(\omega_2) = 0.4$, $m(\omega_3) = 0.1$. Assume that CBR yields discounting factor $\alpha = 0.8$, imprecision matrix $\Phi = [\phi]_{3\times 3} = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.1 & 0.9 & 0 \\ 0.2 & 0.6 & 0.2 \end{bmatrix}$ and balance number $\beta = 0.5$.
The discounted masses entering the redistribution process are calculated by Eq. (6), $\mathbf{m}_1 = \alpha\mathbf{m}$, that is:
$m_1(\omega_1) = 0.4, m_1(\omega_2) = 0.32, m_1(\omega_3) = 0.08$.
The remaining mass is retained by each class according to $\mathbf{m}_2 = (1-\alpha)\mathbf{m}$, so that one gets:
$m_2(\omega_1) = 0.1, m_2(\omega_2) = 0.08, m_2(\omega_3) = 0.02$.
The masses of singleton classes revised from $\mathbf{m}_1$ via the imprecision matrix Φ and the balance number β are obtained via Eq.(15):

$$\begin{aligned}
\tilde{m}_1(\omega_1) &= \phi_{1,1}m_1(\omega_1) + \beta[\phi_{2,1}m_1(\omega_2) + \phi_{3,1}m_1(\omega_3)] \\
&= 0.144; \\
\tilde{m}_1(\omega_2) &= \phi_{2,2}m_1(\omega_2) + \beta[\phi_{1,2}m_1(\omega_1) + \phi_{3,2}m_1(\omega_3)] \\
&= 0.412; \\
\tilde{m}_1(\omega_3) &= \phi_{3,3}m_1(\omega_3) + \beta[\phi_{1,3}m_1(\omega_1) + \phi_{2,3}m_1(\omega_2)] \\
&= 0.056.
\end{aligned}$$

In the revised BBA, the total mass of each singleton class is calculated by adding $\mathbf{m}_2$ as in Equation (16):

$$\begin{aligned}
\hat{m}(\omega_1) &= \tilde{m}_1(\omega_1) + m_2(\omega_1) = 0.244; \\
\hat{m}(\omega_2) &= \tilde{m}_1(\omega_2) + m_2(\omega_2) = 0.492; \\
\hat{m}(\omega_3) &= \tilde{m}_1(\omega_3) + m_2(\omega_3) = 0.076.
\end{aligned}$$

The masses cautiously discounted to meta-elements from $\mathbf{m}_1$ are determined using Eq. (17), as a function of the imprecision matrix and the balance number:

$$\begin{aligned}
\hat{m}(\omega_1 \cup \omega_2) &= (1-\beta)[\phi_{1,2}m_1(\omega_1) + \phi_{2,1}m_1(\omega_2)] = 0.116; \\
\hat{m}(\omega_1 \cup \omega_3) &= (1-\beta)[\phi_{1,3}m_1(\omega_1) + \phi_{3,1}m_1(\omega_3)] = 0.048; \\
\hat{m}(\omega_2 \cup \omega_3) &= (1-\beta)[\phi_{2,3}m_1(\omega_2) + \phi_{3,2}m_1(\omega_3)] = 0.024.
\end{aligned}$$

One can easily verify that normalization is satisfied.

In the original BBA $\mathbf{m}$, most belief is committed to ω_1. The available imprecision matrix, however, indicates that patterns belonging to ω_2 are likely to be misclassified as ω_1 and ω_3 since $\phi_{1,2}$ and $\phi_{3,2}$ are large values. Thus, in the revised BBA $\hat{\mathbf{m}}$, large masses are transferred from ω_1 and ω_3 to ω_2 thanks to the imprecision matrix, which also causes some imprecise information to focus on the meta-classes. As a result, ω_2 receives the largest belief/mass in $\hat{\mathbf{m}}$, and the degree of 'confusion' $\hat{m}(\omega_1 \cup \omega_2)$ between ω_1 and ω_2 is also high. The risk of the revision affecting accuracy can be reduced by modeling such imprecision, leaving open the possibility that this may be subsequently clarified by combining $\hat{m}$ with evidence provided by other sources (i.e., other classifiers).

H. Some discussion

1) CBR versus boosting: Although CBR, as an ensemble classification method may show some superficial similarities with the popular boosting and random forest methods [38], [39], the underpinning principles in the two cases are quite different.

In both boosting and random forests, the (weighted) averaging or voting rule is usually employed to combine multiple (weak) sub-classifiers, and the generation of a diverse collection of sub-classifiers plays a crucial role in improving classification accuracy. In this work, however, the sources of information (corresponding to the classifiers) are assumed to be given, and the generation of sub-classifiers is out of the scope of this paper.

What we propose, instead, is an efficient combination method for dealing with the multiple classification results produced by different sources of information, implemented via an evidence combination method based on credal belief redistribution. CBR is used to revise the given classification results by properly redistributing the masses of the classes and their disjoint unions under the framework of evidence theory, after which the well-known DS rule is employed to combine the revised classification output. Other differences between CBR and boosting approaches are pointed out in Section IV-G.

2) Complexity analysis: In the proposed CBR method, for each classification outcome its K nearest neighbors within the training data are sought to estimate the related imprecision matrix, which is in turn used to correct the classification result via proper belief redistribution. This is a rather refined belief redistribution method, as each classification result corresponds to a distinct imprecision matrix. The complexity of K-NN is $\mathcal{O}(s)$, s being the number of training data. If m is the number of test patterns to be classified based on the combination of n sources of information (i.e., classifiers), the K-NN algorithm needs to be run $m \times n$ times to calculate $m \times n$ imprecision matrices. After that, a linear revision of the classification result is conducted, and DS rule is implemented for conjunctive combination of the multiple sources of information. The computing time of both these steps is not related to s. The complexity of the proposed method is therefore $\mathcal{O}(s \times m \times n)$, which is rather time consuming.

For other weighted fusion methods, such as the weighted DS (WDS) combination rule or the weighted averaging fusion (WAF) rule [24], each information source is given only one weight to reflect its influence/importance in the fusion procedure, which is shared by all the classification results produced by this source, without the need to calculate a new weight for each classification result before combination. As a result, their computation burden is lower than that of CBR. However, as we will empirically see in Section IV, this is a price worth paying for a significantly improved classification accuracy.

3) Guidelines on parameter tuning: As explained our credal belief redistribution method involves three tuning parameters: λ, γ and β. In particular, the proportion $\alpha \in [0, 1]$ of

mass entering the redistribution procedure (Eq. (5)) is driven by the parameter γ, which is used to tune the influence of the distance ratio δ. As the latter is a ratio between average distances, rather than absolute ones, γ is relatively robust to the changes in distance values. Some heuristics have been tested for the choice of γ, and good results can, on average, be obtained when γ lies in the interval of [0.5, 1.5]. This is briefly explained here. In applications, the ratio $\delta_l = \sum_{k=1}^{K} d_{lk}/(K\bar{d})$ of the average distance between $\mathbf{m}_l$ (the evidence to correct) and its K nearest neighbors and the average distance between training patterns usually lies in the interval $[1/3, 3]$. If $\delta_l = 1/3$ and $\gamma \in [0.5, 1.5]$, the proportion of redistributed mass lies in the interval $\alpha_l \in [60.65\%, 84.65\%]$, by Eq. (5). If $\delta_l = 3$, this indicates that the neighbors are quite far from $\mathbf{m}_l$, and one has $\alpha_l \in [1.11\%, 22.31\%]$. In general, such intervals in the proportion coefficient look reasonable for $\gamma \in [0.5, 1.5]$. $\gamma = 1$ usually can be recommended as default value for simplicity.

Parameter $\beta \in [0, 1]$ is used in Eq. (14) to balance the belief redistribution processes concerning singleton classes and meta-classes, respectively. The larger the value of β, the more mass is transferred from the singleton classes to the associated meta-classes, and the higher the imprecision of the classification. Whereas, if β is too small, the lion's share of belief will be directly transferred between singleton classes, which may lead to misclassification when the imprecision matrix determined based on the neighbors of the target pattern is not very reliable.

Default values of $\lambda = 1$, $\gamma = 1$ and $\beta = 0.5$ usually yield good classification results according to our many tests with different datasets, and are recommended in most cases. Moreover, *leave-one-out cross validation* can be applied to the training data to retrieve the optimal values of these three parameters for a specific dataset. Such optimal parameter values should minimize the distances between the combination results and the ground truth, namely:

$$\{\gamma, \beta, \lambda\} = \arg\min_{\gamma,\beta,\lambda} \sum_{i=1}^{s} ||\mathbf{m}_i^f - \mathbf{t}_i||, \tag{19}$$

where s is the number of training datapoints, $\mathbf{m}_i^f$ is the result of combining multiple sources of evidence with respect to the i-th training datapoint, and $\mathbf{t}_i$ is the corresponding truth vector. A grid search strategy can be used to find the optimal values of $\gamma \in [0.5, 1.5]$ and $\beta \in [0, 1]$ within the respective suggested intervals, using a small step length (e.g. 0.1). Once γ and β are set, the remaining parameter $\lambda > 0$ can be optimized by minimizing the above objective function (19). The interior-point algorithm [40], [41] can be used to solve such optimization problem. In Matlab™, the interior-point algorithm is implemented by the function *fmincon* to find the optimal value of λ.

IV. Experimental validation

To validate the proposed Credal Belief Redistribution (CBR) method for pattern classification via classifier combination, in this Section we apply CBR to a number of benchmark datasets and compare its performance with that of several other related combination methods, such as DS rule, Weighted DS rule (WDS), Averaging Fusion rule (AF), Weighted Averaging Fusion rule (WAF) [24] and OWDS [26], the latter previously proposed by some of us.

A. Base classifiers

In our tests, Support Vector Machine (SVM) [42], Naive Bayesian classifier (BAY) [43], and Evidential Neural Network (ENN) [8] are employed as base classifiers to generate pieces of evidence. The SVM classifier adopts the one-versus-rest decomposition strategy and the standard linear kernel, and is trained using the L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) algorithm. The regularization coefficient is set to 0.1. The output of the SVM is converted to a probability measure (a process often called 'calibration') to preserve any useful classification information as much as possible.

In a c-class problem, the output of an SVM for a test pattern $\mathbf{y}$ is given by a vector $\mathbf{f} = (f_1, f_2, \ldots, f_c)$, where f_i represents the value of the hyperplane function associated with the SVM trained on class ω_i versus the other classes. The resulting probability distribution can be denoted by $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_c)$, where $\mu_i = \frac{f_i - \min_j f_j}{\sum_{g=1}^{c}(f_g - \min_j f_j)}$. The larger the hyperplane function value the higher the probability, computed in a way similar to max-min normalization. Of course, different kernels or calibration strategies may be adopted, depending on the actual application. In our naive Bayesian classifier implementation, a Gaussian distribution is assumed to hold for each attribute. In ENN, the maximum number of iterations is set to 500 and the stopping threshold is 10^{-4}. In all these base classifiers, the optimal parameter values (e.g. for the regularization coefficient in SVM, the maximum number of iterations and the stopping threshold in ENN) can be determined, as usual, by cross-validation on the training data[1]. Two popular ensemble learning methods, Adaboost [38] and Random Forest (RF) [39], are also included for sake of comparison. Both the number of learning cycles in Adaboost and the number of trees in RF are set to 100.

B. Benchmark datasets

Fifteen datasets from the UCI repository (available at http://archive.ics.uci.edu/ml) are used to evaluate the performance of the various combination methods analyzed. The basic features of these datasets, including number of classes, number of attributes (#Attr.) and number of instances (#Inst.), are all shown in Table II.

In our experimental setting, for each benchmark dataset the attribute set is divided into n folds, and a different classifier is

[1]Default parameter values, such as setting the regularization coefficient to 0.1, the maximum number of iterations to 500 and the stopping threshold to 10^{-4}, are often used in applications. We found that good results on average can be obtained using such values, according to our many tests on different data sets: hence, we adopted those settings for sake of simplicity. In fact, it turns out that the classification results are not very sensitive to the tuning of these parameters, and small changes of these settings do not have a significant effect on the performance of these classifiers.

Table III
CLASSIFICATION ACCURACY OF DIFFERENT COMBINATION METHODS WITH SVM AS BASE CLASSIFIER (IN %).

Data	n	$[\eta^l,\ \eta^u]$	AF	WAF	DS	WDS	OWDS	CBR
PB	5	[89.88±0.08, 91.50±0.26]	89.84±0.13	89.90±0.08	89.84±0.27	89.84±0.29	91.81±0.36	**93.52**±0.52
PB	3	[90.44±0.28, 93.77±0.13]	91.72±0.78	91.65±0.87	91.74±0.75	91.81±0.76	93.90±0.72	**95.15**±0.19
Te	10	[56.44±0.34, 78.18±0.17]	92.09±0.18	69.20±0.26	92.31±0.16	92.64±0.29	92.64±0.38	**98.02**±0.21
Te	4	[76.47±0.37, 87.51±0.41]	96.42±0.20	95.98±0.15	96.65±0.14	96.97±0.15	97.08±0.13	**99.22**±0.11
Sat	12	[60.30±2.01, 78.90±0.86]	79.89±0.28	71.03±1.76	78.03±0.53	79.67±0.44	80.53±0.35	**84.63**±0.18
Sat	3	[71.48±0.24, 75.21±0.23]	76.38±0.25	76.24±0.23	75.98±0.13	76.32±0.18	76.96±0.27	**80.31**±0.11
Ta	2	[36.82±1.95, 48.09±1.29]	47.69±3.98	43.98±4.48	47.16±4.46	47.16±4.14	49.23±4.15	**54.81**±3.82
Veh	3	[49.76±1.95, 56.38±1.65]	60.28±1.51	62.88±1.52	59.22±0.48	61.70±1.52	63.35±1.57	**64.50**±1.48
Veh	6	[37.94±2.51, 50.71±1.52]	58.04±1.61	56.74±2.73	60.17±2.65	58.16±2.58	60.29±2.76	**63.92**±1.27
Rwq	5	[43.34±1.19, 54.60±0.43]	54.03±0.34	55.97±0.27	55.04±0.55	54.72±0.46	56.58±0.39	**63.78**±0.83
Rwq	2	[49.53±1.25, 56.04±0.06]	57.54±0.16	55.72±0.38	57.54±0.25	58.04±0.27	59.85±0.31	**62.74**±0.88
New	2	[84.65±1.58, 89.30±1.20]	88.84±1.34	90.23±1.22	88.84±1.50	90.70±1.42	**93.06**±1.51	92.94±1.93
ORHD	4	[66.73±0.42, 84.57±0.35]	92.26±0.18	91.25±0.28	92.37±0.23	92.24±0.25	94.15±0.25	**95.31**±0.15
ORHD	8	[39.32±0.47, 64.88±0.22]	77.69±0.22	77.40±0.27	79.23±0.18	81.98±0.26	83.10±0.29	**85.08**±0.21
Vow	3	[16.67±1.08, 34.75±1.24]	34.55±1.87	37.17±2.09	35.35±1.63	42.63±1.68	45.83±1.77	**87.97**±0.92
Vow	2	[33.84±0.17, 37.68±1.40]	43.64±0.59	40.00±0.06	44.65±1.19	44.34±0.27	48.38±0.86	**89.34**±1.17
Pen	4	[53.83±0.82, 62.69±0.19]	78.72±0.09	77.71±0.09	79.28±0.07	78.59±0.21	81.21±0.19	**94.37**±0.33
Pen	8	[26.16±0.43, 44.81±0.80]	72.86±0.63	65.84±0.55	74.60±0.09	74.21±0.14	78.35±0.32	**85.26**±0.37
Hay	2	[43.13±3.54, 43.75±3.82]	43.75±4.16	43.75±3.57	46.25±0.73	45.00±2.70	46.89±3.15	**52.10**±3.98
Kno	2	[83.83±1.46, 92.54±0.47]	88.31±0.97	88.31±1.00	88.31±1.05	88.81±0.87	90.81±0.91	**96.81**±0.45
ML	6	[35.00±1.95, 48.61±2.37]	61.11±1.47	58.89±2.08	61.94±1.00	61.39±1.53	63.17±2.27	**86.64**±2.92
ML	15	[25.28±0.94, 44.72±2.21]	59.44±2.24	33.06±2.19	60.83±2.73	59.72±2.50	61.95±2.65	**83.64**±2.89
Seg	2	[41.90±1.05, 80.43±1.13]	74.94±2.84	72.55±2.68	74.46±2.16	83.94±1.86	84.84±2.15	**86.51**±1.34
Seg	5	[31.69±0.85, 80.13±0.90]	79.18±1.09	88.48±1.19	77.84±1.55	86.71±1.26	90.58±1.33	**91.40**±0.33
WQ	2	[46.82±0.19, 48.65±0.56]	48.63±0.62	48.53±0.50	48.67±0.29	48.61±0.38	51.59±0.52	**61.98**±0.76

Table II
BASIC CHARACTERISTICS OF THE DATASETS EMPLOYED HERE.

Dataset	#Classes	#Attr.	#Inst.
Page-blocks (PB)	5	10	5472
Texture (Te)	11	40	5500
Satimage (Sat)	7	36	6435
Tae (Ta)	3	5	151
Vehicle (Veh)	4	18	946
Red wine quality (Rwq)	6	11	1599
Newthyroid (New)	3	4	215
Opt. Rec. Hand. Digits(ORHD)	10	64	5620
Vowel-context (Vow)	11	10	990
Pen-Based Recognition (Pen)	10	16	10992
Hayes-Roth (Hay)	3	5	160
Knowledge (Kno)	4	5	403
Movement-libras (ML)	15	90	360
Segment (Seg)	7	19	2310
Wine quality (WQ)	7	11	4898

trained on a distinct fold of the attribute set[2]. This generates n sources of information associated with the learnt classifiers. In total we consider 25 cases (each corresponding to a dataset with a specific value of n), and test the performance of our proposed CBR approach over these cases. For each case, the base classifiers generate n pieces of classification results for each query pattern to classify, each represented by a (Bayesian/simple) BBA. These are combined as described in Section III to obtain the final class decision. The number of base classifiers n ranges from 2 to 15 for different datasets[3].

In WDS and WAF, the weighting factors are determined according to the normalized training accuracy of each classifier, as done in [24]. This is expressed by Equation (20):

$$w_l = \frac{\eta_l}{\sum_{i=1}^{n} \eta_i}, \tag{20}$$

where η_l denotes the individual accuracy of classifier $\mathcal{C}_l$ on the whole data set, given by $\eta_l \triangleq \frac{n_c}{T}$, n_c being the number of patterns correctly classified and T the total number of patterns.

C. Performance evaluation

k-fold cross validation [44] is often used for classification performance evaluation, but k remains a free parameter [44]. Here we use the simplest 2-fold cross validation method[4], since training and test sets are of equal size, and each sample can be respectively used for training and testing on each fold. For each fold, the program is randomly run 10 times, and

[2]Naive Bayes is not applicable to the Pen and ORHD datasets, for the within-class variance of several attributes is not positive. Consequently, 13 datasets are used in conjunction with the Naive Bayesian classifier. Adaboost and random forest are implemented using all the original attributes.

[3]In this paper the sources of information are considered already given, and our proposed combination method focusses on the best way to combine any number n of such sources. We are not concerned with selecting the optimal number of sources to combine, as this is a given in our approach. In out tests we create n sources for each dataset by randomly splitting the set of attributes (feature components) into n folds, and learning a classifier for each fold. The classification performance is reported in Tables III, IV and V for two possibly such splits to show that CBR consistently outperforms its competitors no matter the number of sources considered.

[4]The commonly used 10-fold cross validation approach can be also applied here. In our tests, however, it produces performances close to that of 2-fold cross validation.

we report CBR's classification performance for K (number of training neighbors considered) in the range $\{5, 6, \ldots, 15\}$. The mean classification accuracy together with the related standard deviation for the various competing combination methods, using respectively SVM, ENN and Bayesian base classifiers, are reported in Tables III, IV and V. The maximal and minimal accuracy of the individual classifiers are also presented as $\eta^u \triangleq \max_j \eta_j$, $\eta^l \triangleq \min_j \eta_j$. The highest classification accuracy for each data set across the various methods is labeled in bold.

A *sign test* is used to check whether significant differences exist between our proposed CBR and other methods. In the sign test, a Z value is calculated as:

$$Z \triangleq \frac{(r - 0.5) - 0.5N}{0.5N}, \tag{21}$$

where $r > 0.5N$ is the number of cases in which CBR delivers higher accuracy than the compared method, and N is the total number of cases to test. The results are shown in Table VI.

Table VI
PERFORMANCE COMPARISON BETWEEN CBR AND OTHER METHODS VIA THE SIGN TEST.

	AF	WAF	DS	WDS	OWDS
Z_{SVM}	4.8	4.8	4.8	4.8	4.4
Z_{ENN}	4.4	4.8	4.8	4.4	4.0
Z_{BAY}	4.36	4.36	4.36	3.93	4.36

D. Discussion

From Tables III-V, one can appreciate how combining classifiers generally produces higher accuracy than the individual models, in most cases. This indicates that the different classifiers may provide important complementary knowledge, a very helpful feature for improving classification performance.

One can also note that the proposed CBR method typically yields the highest accuracy, when compared to other combination methods (i.e., DS, WDS, AF, WAF). In traditional weighted fusion methods, the weight is mainly used to control the influence of each classifier in the fusion procedure. In the weighted averaging fusion (WAF) rule, for instance, the probability of each class in the classifier's output is reduced in proportion to the given weight in the fusion process. In the weighted DS (WDS) rule, the belief on each class is roughly discounted (reassigned) to the 'total ignorance' class.

As already pointed out, such weighted fusion strategy cannot improve the accuracy of each individual piece of classification result at all. In our previous OWDS method, the weighting factors are optimized using the whole training data, and different targets share a common weight for a given classifier.

In actual reality, classification accuracy varies across different test patterns for any given classifier. In our proposed CBR method, an efficient and sophisticated belief redistribution strategy is plugged in to correct the output of a classifier for each test pattern. The involved weighting parameters (discounting factor, imprecision matrix and balance number) are estimated as a function of the training neighbors of the classifier output for each specific test pattern, inspired by the notion that the local structure of such neighborhoods can well reveal useful information on the classifier's performance in the local region around the target. Moreover, in CBR belief is allowed to be discounted not only to other singleton classes, but also to the associated the meta-classes according to the imprecision matrix. By doing this, one can properly redistribute masses among different classes to improve accuracy. When the imprecision matrix is not reliable enough, CBR also provides a robust belief redistribution strategy based on meta-classes, aimed at reducing misclassification risks by modeling partial imprecision/confusion involving pairs of classes. These are the reasons why CBR globally outperforms other fusion methods.

Quantitatively, the Z values produced by the sign test for all competing fusion methods, and reported in Table VI, are all bigger than $Z_{0.05/2} = 1.96$. This confirms that CBR yields significantly better performance than other fusion methods.

E. Influence of K

In order to test the influence of K on the classification performance of CBR, curves plotting accuracy versus $K \in [5, 15]$, for different base classifiers on various datasets, are reported in Figure 1. In the legends, $\mathcal{C}$-n (e.g., SVM-5) means base classifier $\mathcal{C}$ (e.g., SVM) and number of classifiers being combined equal to n (e.g., 5). The x-axis represents K values, and y-axis represents the classification accuracy.

As it can be appreciated, variations in accuracy associated with different K values as generally very small. This is because training neighbors far from the pattern to classify end up playing a minor role in the belief redistribution procedure. This indicates that the performance of CBR is robust to the tuning of K, as desirable for real applications.

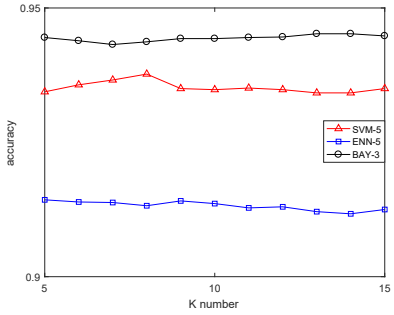
F. Computational cost

The execution time (in seconds, s) of the different methods with SVM as base classifier is shown in Table VII.

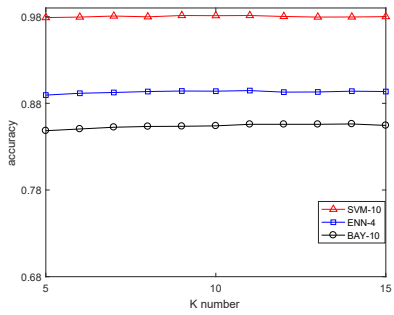
Table VII shows that the proposed CBR method is indeed more time consuming than other methods, since the K nearest neighbors of each classification result need to be found to estimate the imprecision matrix. A tradeoff thus exists between accuracy and computation when using this approach. Generally speaking, CBR is more suitable for applications in which high classification accuracy is required whereas efficient computation is not a strong requirement.

G. Comparison with boosting

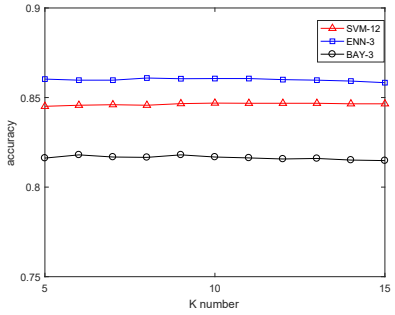
Two popular ensemble learning methods, i.e. Adaboost [38] and Random Forest (RF) [39], are also included here for sake of comparison, and their classification results are shown in Table VIII. For each dataset, two cases (two values of n) are considered in our evaluation of CBR. We selected the case with the highest accuracy here for comparison, and the corresponding accuracy is also reported in Table VIII for the convenience. A total of 15 datasets were considered in this experiment. We found that random forests produce



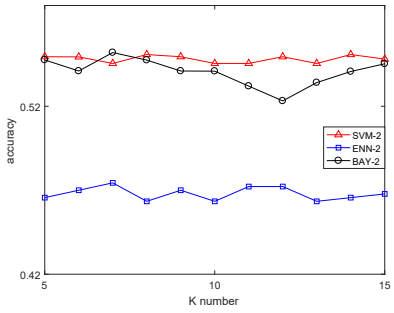
(a) Pb data.



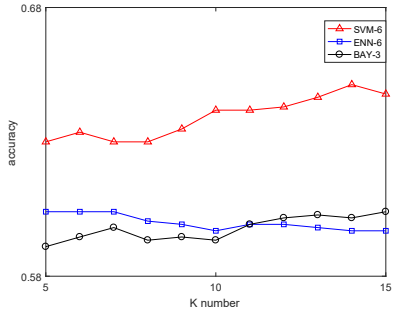
(b) Te data.



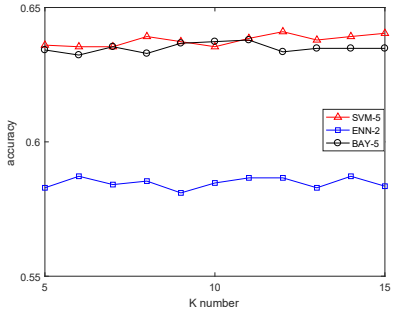
(c) Sat data.



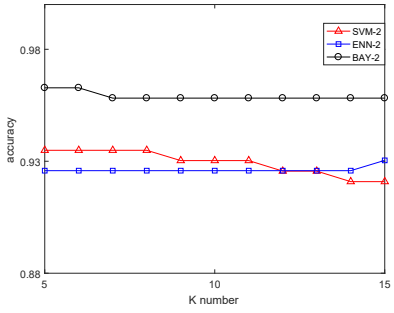
(d) Ta data.



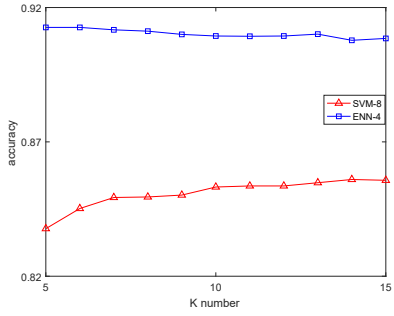
(e) Veh data.



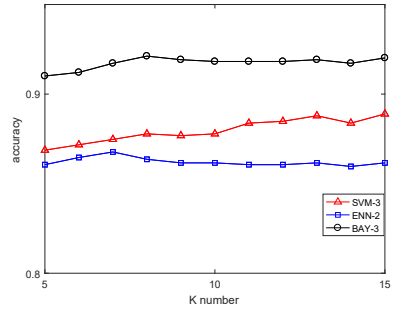
(f) Rwq data.



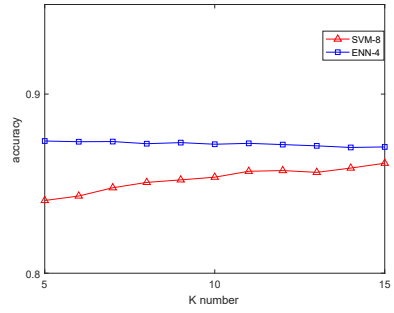
(g) New data.



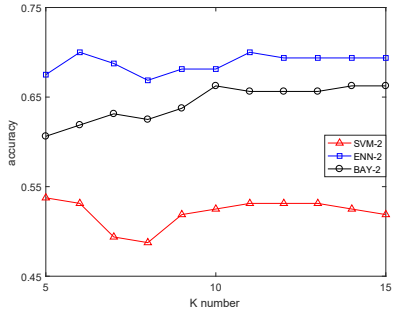
(h) Orhd data.



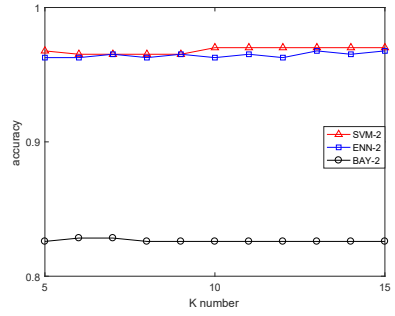
(i) Vow data.



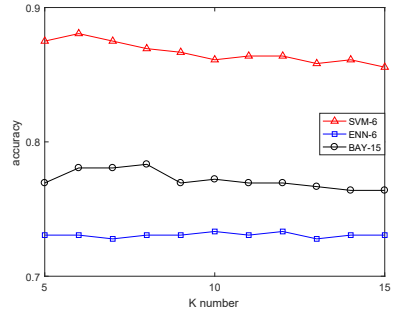
(j) Pen data.



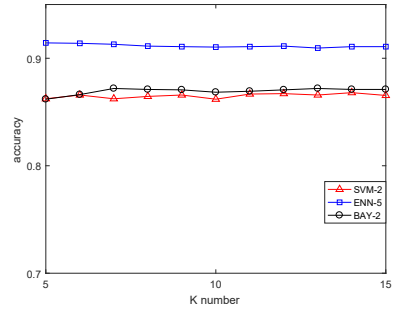
(k) Hay data.



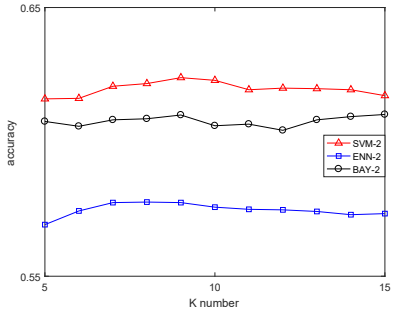
(l) Kno data.



(m) ML data.



(n) Seg data.



(o) WQ data.

Figure 1. Plots of CBR's classification accuracy as a function of $K \in [5, 15]$, over different datasets and for different base classifiers.

Table IV
CLASSIFICATION ACCURACY OF DIFFERENT COMBINATION METHODS WITH ENN BASE CLASSIFIER (IN %).

Data	n	$[\eta^l, \eta^u]$	AF	WAF	DS	WDS	OWDS	CBR
PB	5	[89.77±0.01, 90.97±0.05]	89.77±0.01	89.77±0.01	89.77±0.01	89.77±0.01	90.23±0.05	**91.32**±0.20
PB	3	[89.77±0.01, 91.38±0.06]	89.77±0.01	89.77±0.01	89.77±0.01	89.77±0.01	90.35±0.03	**91.32**±0.27
Te	10	[50.85±0.69, 68.55±1.70]	80.76±1.05	59.13±0.39	80.33±0.87	80.65±1.44	82.53±0.91	**89.89**±0.46
Te	4	[56.38±0.56, 67.47±0.26]	79.22±0.24	76.71±0.47	80.05±0.46	79.04±0.31	81.65±0.39	**89.31**±0.55
Sat	12	[69.23±0.51, 82.35±2.27]	84.34±0.31	73.70±0.45	84.07±0.18	84.16±0.31	85.03±0.25	**85.52**±0.23
Sat	3	[77.56±1.21, 78.48±1.03]	83.20±1.19	83.12±1.04	83.34±0.87	83.48±0.97	84.19±1.02	**85.99**±1.14
Ta	2	[31.78±3.52, 37.51±4.48]	38.83±2.01	37.94±2.11	39.27±2.65	39.05±1.97	42.61±2.86	**46.80**±3.65
Veh	3	[41.13±1.91, 42.43±1.01]	52.72±1.56	49.65±0.89	52.13±0.61	52.60±1.54	54.10±0.94	**56.67**±1.18
Veh	6	[37.23±0.36, 49.29±3.72]	55.79±1.08	55.32±0.21	55.44±0.89	55.32±0.89	56.82±0.91	**59.99**±1.91
Rwq	5	[40.71±0.43, 56.72±0.56]	56.47±1.11	57.72±1.09	56.66±1.02	57.35±0.94	59.25±0.72	**60.82**±0.16
Rwq	2	[47.72±0.90, 47.84±0.42]	49.84±1.09	49.72±0.93	49.59±1.46	49.72±1.17	51.63±1.29	**58.47**±0.36
New	2	[80.46±2.16, 92.11±3.19]	92.11±2.13	92.58±1.68	92.58±1.94	93.04±1.61	**93.15**±1.55	92.62±1.26
ORHD	4	[65.14±0.85, 72.10±1.72]	85.04±0.20	86.35±0.45	85.09±0.29	85.44±0.34	86.21±0.25	**91.03**±0.31
ORHD	8	[43.72±0.16, 66.76±0.23]	88.24±0.43	65.70±0.23	87.86±0.15	88.56±0.38	88.91±0.21	**89.47**±0.23
Vow	3	[34.75±2.52, 42.02±2.12]	65.25±2.73	58.59±3.09	67.78±2.65	67.27±3.20	69.35±2.29	**84.39**±3.32
Vow	2	[35.56±0.76, 55.96±2.24]	61.01±2.68	55.45±2.86	61.11±2.44	61.01±2.05	64.25±2.18	**86.22**±2.39
Pen	4	[54.45±1.56, 64.69±0.93]	80.19±0.82	79.97±1.14	81.04±0.60	80.29±0.39	83.33±0.46	**87.22**±0.80
Pen	5	[48.44±0.87, 62.44±1.76]	81.51±0.81	79.79±0.07	81.39±1.02	81.87±1.03	84.16±0.89	**85.44**±0.62
Hay	2	[54.37±0.44, 58.13±0.88]	**70.63**±0.44	66.88±0.88	66.25±1.76	68.13±0.44	69.05±0.52	68.81±0.75
Kno	2	[71.39±3.31, 77.36±2.82]	85.82±0.86	86.32±1.80	87.06±1.25	86.57±2.24	88.15±1.16	**96.45**±1.44
ML	6	[30.56±1.37, 45.83±2.55]	62.50±3.92	56.39±1.96	64.17±1.96	63.33±3.92	67.20±3.84	**73.06**±3.53
ML	15	[24.72±0.20, 40.00±1.18]	65.00±3.73	30.56±3.14	66.94±2.35	65.00±3.14	68.13±3.42	**70.71**±2.39
Seg	2	[47.10±0.93, 69.52±1.01]	79.78±2.18	78.83±2.08	82.55±2.22	79.00±2.35	83.26±2.85	**85.71**±1.27
Seg	5	[36.41±1.21, 71.00±2.47]	78.40±0.84	79.05±1.15	84.68±0.46	77.10±0.82	85.65±0.73	**91.15**±0.55
WQ	2	[44.75±0.31, 45.02±0.30]	44.81±0.37	44.92±0.05	44.86±0.31	44.79±0.40	45.68±0.45	**57.47**±0.65

Table V
CLASSIFICATION ACCURACY OF DIFFERENT COMBINATION METHODS WITH BAYESIAN BASE CLASSIFIER (IN %).

Data	n	$[\eta^l, \eta^u]$	AF	WAF	DS	WDS	OWDS	CBR
PB	5	[87.19±2.98, 91.94±0.20]	91.78±0.12	88.23±0.21	93.06±0.14	92.14±0.10	93.81±0.11	**94.10**±0.12
PB	3	[86.11±2.27, 92.34±0.36]	92.91±1.50	90.81±2.12	93.92±0.49	93.17±0.36	93.92±0.43	**94.44**±0.29
Te	10	[46.05±0.17, 66.53±0.21]	74.35±0.16	53.51±0.19	77.45±0.12	75.00±0.09	77.55±0.15	**85.36**±0.34
Te	4	[54.15±0.04, 66.80±0.21]	72.02±0.36	70.91±0.34	77.45±0.16	73.38±0.49	77.97±0.38	**83.08**±0.41
Sat	12	[61.71±0.28, 79.55±0.09]	77.28±0.05	77.27±0.06	76.85±0.07	77.11±0.12	77.68±0.09	**79.10**±0.11
Sat	3	[76.27±0.08, 79.53±0.09]	81.18±0.08	81.17±0.09	80.00±0.12	80.64±0.10	80.82 ±0.16	**81.64**±0.23
Ta	2	[39.96±2.01, 47.01±2.64]	48.56±2.13	49.24±1.36	49.01±2.89	48.96±2.91	50.18±2.65	**54.05**±3.28
Veh	3	[39.72±1.34, 51.89±1.92]	49.53±2.59	52.25±1.42	47.28±3.93	50.35±2.17	52.19±2.75	**59.77**±3.18
Veh	6	[36.76±1.50, 43.38±0.17]	47.28±0.67	49.53±2.34	45.51±0.08	48.23±0.33	57.95±0.46	**63.84**±1.09
Rwq	5	[39.15±0.68, 55.10±0.33]	56.10±0.55	56.22±0.52	56.22±0.73	56.60±0.80	57.65 ±0.86	**63.49**±1.05
Rwq	2	[42.90±2.15, 56.10±0.23]	55.85±1.72	53.66±1.52	56.85±1.18	57.16±0.88	59.08±1.01	**62.75**±1.04
New	2	[88.85±0.83, 91.65±0.60]	95.82±0.53	95.83±0.76	95.82±0.53	95.83±0.96	95.83±0.83	**95.90**±0.91
Vow	3	[36.97±1.14, 42.32±1.79]	67.88±1.00	61.62±2.24	71.62±0.29	68.08±2.57	75.35±1.89	**91.74**±1.48
Vow	2	[47.68±1.21, 58.99±0.86]	66.97±2.71	63.74±1.76	71.62±1.62	66.67±1.52	75.96±1.65	**90.56**±1.19
Hay	2	[53.13±2.03, 56.25±1.20]	61.87±4.19	58.75±3.70	59.38±3.72	**65.63**±3.92	64.23±3.83	64.32±3.40
Kno	2	[31.84±1.56, 79.10±1.35]	80.10±1.43	80.10±1.51	83.33±1.52	79.85±1.36	81.85±1.49	**82.63**±1.60
ML	6	[27.78±1.59, 40.00±1.67]	50.28±0.86	47.50±2.24	69.44±1.20	50.00±0.93	71.33±1.18	**73.48**±1.89
ML	15	[21.11±0.83, 33.06±1.87]	48.06±1.74	27.78±2.78	69.44±4.03	48.89±2.53	72.45 ±2.41	**77.17**±2.39
Seg	2	[49.91±0.41, 79.52±0.76]	67.53±0.40	59.57±0.26	80.39±0.57	80.17±0.70	83.52±0.55	**86.94**±0.87
Seg	5	[30.22±0.32, 75.89±0.74]	77.19±0.82	80.09±0.34	80.39±0.54	80.22±1.06	83.98±0.73	**89.28**±0.92
WQ	2	[41.40±0.29, 46.82±0.37]	47.73±0.19	46.04±0.62	48.00±0.06	48.04±0.22	50.85±0.42	**60.77**±0.72

performances comparable to that of CBR. While CBR yields higher accuracy than RF on some datasets, RF outperforms CBR on others. More precisely, in our tests CBR coupled with SVM is able to produce higher accuracy than Adaboost and RF on 9 datasets, whereas RF yields the highest accuracy on the remaining 6 benchmarks.

We need to stress, however, that CBR operates on a principle which is quite distinct from those upon which both Adaboost and random forests are built. Specifically, in this work the various pieces of information (i.e., the classifiers outputs) are passed to a fusion center, where the CBR approach is used to efficiently combine classification results. The original attribute values of the patterns are not at all involved in such a decision level fusion process. In contrast, both Adaboost and RF require access to the attribute values of the patterns in order to generate (weak) sub-classifiers, which are then

Table VIII
CLASSIFICATION ACCURACY OF RANDOM FOREST, ADABOOST AND CBR METHODS(IN %).

Data	Ada	RF	CBR
PB	93.44±0.20	**97.14**±0.23	95.15±0.19
Te	56.92±3.84	78.96±0.64	**99.22**±0.11
Sat	78.60±0.46	**90.68**±0.19	84.63±0.18
Ta	40.01±3.42	51.33±3.78	**54.81**±3.82
Veh	54.48±2.49	**73.78**±0.87	64.50±1.48
Rwq	56.35±0.36	63.66±1.09	**63.78**±0.83
New	94.23±1.15	**94.70**±0.76	92.94±1.93
ORHD	68.10±1.09	87.96±0.44	**95.31**±0.15
Vow	30.23±3.11	72.38±2.24	**89.34**±1.17
Pen	51.30±1.35	88.55±0.25	**94.37**±0.33
Hay	56.63±5.23	**79.63**±3.04	52.10±3.98
Kno	79.23±1.68	93.76±1.07	**96.81**±0.45
ML	18.92±1.67	45.11±2.79	**86.64**±2.92
Seg	69.84±1.84	91.04±0.47	**91.40**±0.33
WQ	47.36±1.29	**63.47**±0.56	61.98±0.76

Table VII
EXECUTION TIME(S) OF THE DIFFERENT COMBINATION METHODS WITH SVM AS BASE CLASSIFIER.

Data	n	AF	WAF	DS	WDS	OWDS	CBR
PB	5	0.2421	0.2429	0.3819	0.3860	0.3941	0.5996
PB	3	0.2418	0.2420	0.3784	0.3817	0.3910	0.5253
Te	10	0.0041	0.0046	0.0544	0.0612	0.0779	1.1307
Te	4	0.0038	0.0038	0.0285	0.0312	0.0360	1.1307
Sat	12	0.3205	0.3277	0.5720	0.5913	0.6054	0.8337
Sat	3	0.3175	0.3175	0.2337	0.2856	0.2915	0.8337
Ta	2	0.0011	0.0011	0.0020	0.0032	0.0041	0.0063
Veh	3	0.0007	0.0007	0.0023	0.0029	0.0037	0.0176
Veh	6	0.0011	0.0011	0.0036	0.0049	0.0053	0.0197
Rwq	5	0.0012	0.0012	0.0069	0.0074	0.0079	0.0753
Rwq	2	0.0011	0.0011	0.0037	0.0044	0.0048	0.0420
New	2	0.0002	0.0002	0.0006	0.0008	0.0008	0.0026
ORHD	4	0.0041	0.0042	0.0258	0.0294	0.0337	0.7269
ORHD	8	0.0041	0.0042	0.0462	0.0498	0.0522	0.9690
Vow	3	0.0008	0.0008	0.0038	0.0043	0.0056	0.0547
Vow	2	0.0008	0.0008	0.0027	0.0032	0.0032	0.0429
Pen	4	0.0084	0.0084	0.0678	0.0714	0.0764	1.3895
Pen	8	0.0085	0.0088	0.1013	0.1098	0.1115	1.9052
Hay	2	0.0002	0.0002	0.0005	0.0008	0.0008	0.0020
Kno	2	0.0003	0.0003	0.0010	0.0014	0.0015	0.0059
ML	6	0.0005	0.0005	0.0027	0.0036	0.0045	0.0563
ML	15	0.0005	0.0005	0.0067	0.0090	0.0107	0.1189
Seg	2	0.0014	0.0015	0.0076	0.0085	0.0092	0.0856
Seg	5	0.0017	0.0018	0.0094	0.0113	0.0129	0.1521
WQ	2	0.0027	0.0031	0.0138	0.0155	0.0159	0.2572

integrated to improve accuracy. CBR and these two ensemble learning methods, therefore, are really designed to deal with completely different fusion settings. Whenever the fusion center receives multiple classification results from different information sources but the original pattern attributes cannot be obtained, then CBR is a sensible option. If pattern attributes are available, RF is a reasonable alternative choice for the classification task.

V. CONCLUSIONS

Evidence theory has been widely applied to classifier fusion, as each classifier output can be represented by a piece of evidence in the form of a Basic Belief Assignment. The various pieces of evidences may have different qualities of classification, and they are usually discounted using different weights before combination.

In order to achieve the best possible fusion performance, in this paper a new Credal Belief Redistribution (CBR) method was proposed to amend the evidence prior to combination. In CBR, the K-Nearest Neighbors of each piece of evidence are found at first to calculate the parameters of the discounting process. The amount of mass/belief to be redistributed mainly depends on the distance between the piece of evidence and its training neighbors. Larger distances allow less mass to enter the redistribution procedure. An imprecision matrix is estimated based on the K neareast neighbors to characterize the degree of imprecision of the classification, i.e., the probability of a target to belong to each possible class, given the classification result declared by the classifier at hand. Since the neighbors may or may not adequately represent the evidence, we cannot completely trust the related imprecision matrix for evidence correction. As a consequence, CBR allows us to reassign masses from single classes to both other singleton classes or to meta-classes (i.e., disjunctions of two classes). The transfer of belief between different singleton classes can efficiently improve the accuracy of evidence when the imprecision matrix is very reliable, but carries misclassification risks when this is not the case. Redistributing a fraction of the mass to the meta-classes can mitigate this risk by properly modeling partial imprecision. As a result, CBR is able to well balance degree of specificity and misclassification risk in the evidence correction process. The multiple pieces of corrected evidence are finally combined by DS rule for pattern classification.

Various real datasets from the UCI repository were employed to evaluate the performance of CBR. Our experimental results show that CBR can indeed improve classification accuracy over other related fusion methods. The classification accuracy of CBR is robust to the choice of K, the number of neighbors, which is appealing for real applications.

As a limitation, the computation burden of CBR is significant, but unavoidable in our strive to improve accuracy. So the application of CBR, as currently formulated, is constrained in cases in which fast computation is required, for instance

when dealing with truly big datasets. In our future work we will indeed work towards developing a faster, more efficient credal belief redistribution method, by adding elements of unsupervised learning. Namely, close classification results can be grouped together in order to have to calculate only one imprecision matrix per group, thus significantly reducing the computation burden. In parallel, we seek to extend CBR's range of applications to include complex, real-world tasks, such as multi-source information fusion (using e.g. SAR, traditional and infrared imagery) for target identification in uncertain environments.

ACKNOWLEDGEMENTS

This work has been partially supported by National Natural Science Foundation of China (Nos.61672431, 61790552, 61790554) and Shaanxi Science Fund for Distinguished Young Scholars (No.2018JC-006).

REFERENCES

[1] E. Blasch, P. Valin, A.L. Jousselme, D. Lambert, E. Bosse, *Top ten trends in High-Level Information Fusion*, 15th International Conference on Information Fusion, Singapore, 2012.

[2] P. Hui Foo, G. Wah Ng, *High-level Information Fusion: An Overview*, Journal of Advance in Information Fusion, Vol. 8(1):33-72, 2013.

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

[4] P. Smets, *The combination of evidence in the transferable belief model*, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 12(5):447–458,1990.

[5] J.B. Yang, D.l. Xu, *Evidential reasoning rule for evidence combination*, Artificial Intelligence, Vol. 205:1–29, 2013.

[6] F. Smarandache, J. Dezert (Editors), *Advances and applications of DSmT for information fusion*, American Research Press, Rehoboth, Vols. 1-4, 2004-2015. http://www.onera.fr/staff/jean-dezert?page=2

[7] A.-L. Jousselme, C. Liu, D. Grenier, E. Bossé, *Measuring ambiguity in the evidence theory*, IEEE Trans. Systems, Man & Cybernetics - Part A: systems, Vol.36(5):890–903, 2006.

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

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

[10] Z.g Liu, Q. Pan, J. Dezert, G. Mercier, *Hybrid classification system for uncertain Data*, IEEE Trans. Systems, Man, and Cybernetics: Systems, Vol.47(10):2783–2790, 2017.

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

[12] C. Lian, S. Ruan, T. Denœux, *Dissimilarity metric learning in the belief function framework*, IEEE Trans. Fuzzy Systems, Vol. 24(6): 1555–1564, 2016.

[13] L. Jiao, T. Denœux, Q. Pan, *A hybrid belief rule-based classification system based on uncertain training data and expert knowledge*, IEEE Trans. Systems, Man and Cybernetics: Systems, Vol. 46(12):1711–1723, 2016.

[14] Z.g. Su, T. Denœux, Y.-S. Hao, M. Zhao, *Evidential K-NN Classification with Enhanced Performance via Optimizing a Class of Parametric Conjunctive t-Rules*, Knowledge-Based Systems, Vol.142:7–16, 2018.

[15] T. Denoeux, S. Li and S. Sriboonchitta, *Evaluating and comparing soft partitions: an approach based on Dempster-Shafer theory*, IEEE Trans. Fuzzy Systems, Vol. 26(3):1231–1244, 2018.

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

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

[18] 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, Vol.52(3):353–374, 2011.

[19] Y.X. Bi, J.W. Guan, D. Bell, *The combination of multiple classifiers using an evidential reasoning approach*, Artificial Intelligence, Vol. 172: 1731–1751, 2008.

[20] Z.g. Liu, J. Dezert, Q. Pan, G. Mercier, *Combination of sources of evidence with different discounting factors based on a new dissimilarity measure*, Decision Support Systems, Vol.52(1):133–141, 2011.

[21] S. Huang, X. Su, Y. Hu, S. Mahadevan, Y. Deng, *A new decision-making method by incomplete preferences based on evidence distance*, Knowledge-Based Systems, Vol.56:264–272, 2014.

[22] Z. Feng, Z. Zhou, C. Hu, L. Chang, G. Hu, F. Zhao, *A new belief rule base model with attribute reliability*, IEEE Trans. Fuzzy Systems, 2019, DOI 10.1109/TFUZZ.2018.2878196.

[23] D. Ruta, B. Gabrys, *An overview of classifier fusion methods*, Computing and Information Systems, Vol.7:1–10, 2000.

[24] F. Moreno-Seco, J.M. Inesta, P.J. Ponce de Leon, L. Mico, *Comparison of classifier fusion methods for classification in pattern recognition tasks*, D.-Y. Yeung et al. (Eds.): Springer-Verlag Berlin Heidelberg, pp. 705–713, 2006.

[25] Z.W. Yu, L. Li, J.M. Liu, G.Q. Han, *Hybrid adaptive classifier ensemble*, IEEE Trans. Cybernetics, Vol. 45(2):177–190, 2015.

[26] Z.g. Liu, Q. Pan, J. Dezert, A. Martin, *Combination of classifiers with optimal weight based on evidential reasoning*, IEEE Trans. Fuzzy Systems, Vol. 26(3):1217–1230,2018.

[27] A. Martin, *About conflict in the theory of belief functions*, 2nd International conference on Belief Functions, Compiegne, France, May, 2012.

[28] A. Martin, A.L. Jousselme, C. Osswald, *Conflict measure for the discounting operation on belief functions*. 11st International Conference on Information Fusion, Cologne, Germany, July, 2008.

[29] A.-L. Jousselme, P. Maupin, *Distances in evidence theory: comprehensive survey and generalizations*, International Journal of Approximate Reasoning, Vol.53(2):118–145, 2012.

[30] S. Destercke, P. Buche, B. Charnomordic, *Evaluating data reliability: an evidential answer with application to a web-enabled data warehouse*, IEEE Trans. Knowledge and Data Engineering, Vol.25(1):92–105, 2013.

[31] L. Kuncheva, *Combining pattern classifiers: methods and algorithms*, Wiley, 2004.

[32] Y. Yang, D. Han, *A new distance-based total uncertainty measure in the theory of belief functions*, Knowledge-Based Systems, Vol. 94:114-123, 2016.

[33] W. Liu, *Analyzing the degree of conflict among belief functions*, Artificial Intelligence, Vol.170(11):909–924, 2006.

[34] L.I. Kuncheva, *A theoretical study on six classifier fusion strategies*, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.24(2):281–286, 2002.

[35] D. Mercier, B. Quost, T. Denœux, *Refined modeling of sensor reliability in the belief function framework using contextual discounting*, Information Fusion, Vol.9(2):246–258, 2008.

[36] Z.g. Liu, Q. Pan, J. Dezert , J.w. Han, Y. He, *Classifier fusion with contextual reliability evaluation*, IEEE Trans. Cybernetics, Vol.48(5):1605–1618, 2018 .

[37] J. Dezert, A. Tchamova, *On the validity of Dempster's fusion rule and its interpretation as a generalization of Bayesian fusion rule*, International Journal of Intelligent Systems, Vol. 29(3):223–252, 2014.

[38] Y. Freund, *A more robust boosting algorithm*, arXiv:0905.2138v1, 2009.

[39] L. Breiman, *Random Forests*, Machine Learning, Vol.45:5–32, 2001.

[40] George B. Dantzig, Mukund N. Thapa, *Linear Programming 2: Theory and Extensions*, Springer-Verlag, 2003.

[41] Florian A. Potra, Stephen J. Wright,*Interior-point methods*, Journal of Computational and Applied Mathematics, Vol.124 (1ï¿½ $\frac{1}{2}$C2): 281–302, 2000.

[42] N. Cristianini, J. Shawe-Taylor, *An introduction to support vector machines*, Cambridge University Press, Cambridge, 2000.

[43] S. Russell, P.Norvig, *Artificial intelligence: a modern approach*, Prentice Hall (second edition), 2003.

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

Zhun-ga Liu was born in China. He received the Bachelor, Master and Ph.D degree from Northwestern Polytechnical university (NPU), Xi'an, China in 2007, 2010 and 2013 respectively. He also studied in Telecom Bretagne, France for Ph.D during 2010 and 2013. He is full time professor in School of Automation, NPU since 2017. His research interest mainly focuses on belief functions, pattern recognition and information fusion.

Yu Liu was born in China. He received the Bachelor, Master and Ph.D degree from Naval Aviation University, Yantai, China in 2008, 2010 and 2014. He worked as post-doctor in Beihang University during 2016 and 2018. He has been an associate professor in Naval Aviation University since 2018. His research interest mainly focuses on information fusion and pattern recognition.

Jean Dezert was born in France. He received the Electrical Engineering degree in 1985, and his Ph.D degree from the University Paris XI, 1990. Since 1993, he has been senior research scientist in the Information Processing and Systems Dept. (DTIS) at the French Aerospace Lab. His main research interest focuses on information fusion, belief function theory and multi-criteria decision-making.

Fabio Cuzzolin received Ph.D. degree from the University of Padua in 2001. He was later with the Washington University in St. Louis, Politecnico di Milano and the University of California at Los Angeles, before been awarded a Marie Curie Fellowship at INRIA Rhone-Alpes, France. He joined Oxford Brookes University in 2008. He is a Professor of Artificial Intelligence since 2016, and he is also Director of the Visual Artificial Intelligence Lab. His research interests mainly focus on the uncertainty theory, belief functions and machine learning.