Classifier Fusion Based on Cautious Discounting of Beliefs

Zhunga Liu and Quan Pan
School of Automation
Northwestern Polytechnical University
Xi'an, China
Email:liuzhunga@nwpu.edu.cn

Jean Dezert
ONERA - The French Aerospace Lab
F-91761 Palaiseau, France.
Email: jean.dezert@onera.fr

***Abstract*—Classifier fusion is a classical approach to improve the classification accuracy. The multiple classifiers to combine have in general different classification qualities (i.e. performances), and the proper evaluation of the classifier quality plays an important role for achieving the best global performance. We propose a new method for classifier fusion based on refined reliability evaluation (CF-RRE). For each object, the reliability of its classification result with a given classifier is characterized by a matrix $R_{c\times c}$ (c being the number of classes in the data set), which is estimated based on the classifier performance in the neighborhoods (i.e. K-nearest neighbors) of the object using the training data. The reliability matrix is used to make a cautious discounting of the classification result. More specifically, the probability (or belief) of the object associated with each class is cautiously redistributed according to the reliability matrix under the belief functions framework. The discounted classification results of each classifier can be combined by Dempster's rule for making the final class decision. Our simulation results illustrate the potential of this new method using real data sets, and they show that CF-RRE can improve substantially the classification accuracy.**

Keywords: classifier fusion, belief functions, reliability, discounting, classification.

I. Introduction

The classification accuracy can be efficiently improved by proper fusion of multiple classifiers, which usually provide complementary classification knowledge for the query pattern from different points of view. This complementarity can be achieved by extracting different features, by employing different classifiers, as well as by randomly selecting different training data sets [1]. The fusion approach is expected to reduce the error rate and enhance the robustness of classification compared with any individual classifier.

Many fusion methods have been developed for making a class decision from the individual classifiers [2]. The selection of appropriate fusion strategy mainly depends on the formats of classifier output. If the output of the classifier consists only of a label value (i.e. a hard-decision classifier), the simple majority voting method is often recommended. If the classifier can generate soft membership measures, like probability value, fuzzy memberships or belief functions, the linear combination way (average, sum, etc) [3], Bayesian combination [4], Bayesian Model Averaging (BMA) [5], fuzzy rules [6], evidential reasoning technique [1], [7] can be used. The soft classification result generally offers more useful information than a single hard label, and the fusion of the soft outputs of different classifiers can improve significantly the classification performance [8].

Belief functions (BF) [9]–[11] known as Dempster-Shafer theory (DST), provides an interesting framework to represent and combine uncertain information [12], [13]. Belief functions allow the object to be associated with not only the singleton classes but also any sets of classes according to a basic belief assignment (BBA), and we adopt it here for the ensemble of multiple classifiers. The belief functions have been already successfully applied in the information fusion [14], [24], data classification [15]–[17] and clustering [18], decision-making support [19], and so on. Particularly, several methods [1], [7], [20] have been introduced for the fusion of multiple classifiers based on belief functions. In [20], several combination strategies (i.e. majority voting, Bayesian formalism and DS model) were introduced, and the conditional probability of the object belonging to different classes was derived based on the confusion matrix. For DS model, each evidence is represented by dichotomous mass functions including three focal elements (e.g. A, $\bar{A}$ and ignorance element Ω), and it was defined according to the overall performance of classifiers. In [7], a class-indifferent method was proposed for multi-classifier fusion using DS rule, and the classifier decisions were modeled by triplet and quartet evidential structures. In [1], an optimal combination scheme was presented based on a parameterized family of t-norms for an ensemble of multiple classifiers that provides the partly dependent information, and the parameter can be optimized to achieve the minimum error criterion.

In the fusion process, the classifiers have in general different reliability factors which play an important role to improve the overall classification accuracy. The reliability factor is usually determined based on the overall classification performance (e.g. accuracy) in the training set, and many methods [21] have been proposed to compute it. In [22], a contextual discounting method has been introduced taking into account the refined reliability knowledge, where the reliability of information source were dependent of the knowledge of the ground truth (true value of the variable known in simulations), which is unfortunately rarely known in real applications.

In many applications, the reliabilities of classification results obtained by one classifier are related with the objects to classify. Different elements (i.e. the different probabilities assigned to each class) in the soft classification result of one object may also have different reliabilities, because the difference between the output value of classifier and the expected value (truth) usually is not the same for the different elements. As example, let's consider an object $\mathbf{y}$ with true class $c(\mathbf{y}) = \omega_1$ to classify over the frame of discernment $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Let us assume that the classifier provides the following probability assignments $p(\omega_1) = 0.4$, $p(\omega_2) = 0.5$, and $p(\omega_3) = 0.1$. If the classifier would have been 100% reliable, it should have provided $p_{true}(\omega_1) = 1$, $p_{true}(\omega_2) = p_{true}(\omega_3) = 0$ as correct output for the classification of $\mathbf{y}$. In order to improve the classification result, it seems very natural to develop a method for revising the classifier output thanks to a refined reliability evaluation (RRE). By doing this, one expects to improve substantially the accuracy of the classification result produced by each classifier.

In this paper we also want to improve the classification performance as far as possible and that is why we propose a new method for classifier fusion with refined reliability evaluation (CF-RRE). The refined reliability knowledge of each classifier is represented by a $R_{c\times c}$ matrix[1], where each element of this reliability matrix represents the likelihood of the object belonging to class $\omega_i, i = 1, \ldots, c$ when it is classified to class $\omega_j, j = 1, \ldots, c$ by the given classifier. This reliability matrix is estimated using the training data close to the object. The soft classification result provided by a classifier can be modified (revised) according to the reliability matrix by a new cautious discounting rule under the belief functions framework. The multiple discounted classification results from different classifiers will be combined altogether using DS rule for the final classification of the object.

This paper is organized as follows. After a brief introduction of the belief functions in section II, we present the method for refined reliability evaluation in details in the section III, with classifier fusion approach. Simulations results are presented in the section IV to evaluate the performance of this new method for different data sets. Section V concludes this work.

II. Basics of belief function theory

The belief functions (BF) have been introduced by Shafer in his Mathematical Theory of Evidence, also known as Dempster-Shafer Theory (DST) [9], [11]. In DST, we work with a discrete frame of discernment as $\Omega = \{\omega_i, i = 1, 2, \ldots, c\}$ consisting of c exclusive and exhaustive hypotheses (classes) $\omega_i, i = 1, \ldots, c$. A basic belief assignment (BBA), also called a mass of beliefs, can be defined over the power-set of Ω denoted by 2^Ω, which is the set of all the subsets of Ω. For example, if the frame of discernment is $\Omega = \{\omega_1, \omega_2, \omega_3\}$, then its power-set is $2^\Omega = \{\emptyset, \omega_1, \omega_2, \omega_3, \omega_1 \cup \omega_2, \omega_1 \cup \omega_3, \omega_2 \cup \omega_3, \Omega\}$. A BBA is mathematically defined as a mapping $m(.)$ from 2^Ω to $[0, 1]$, which satisfies $m(\emptyset) = 0$ and

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

With a BBA $m(.)$, one can allow one object to belong to different elements (singletons, as well as their disjunctions) in 2^Ω with different masses of belief. All the elements $A \in 2^\Omega$ such that $m(A) > 0$ are called the focal elements of the BBA $m(.)$. $m(A)$ represents the support degree of the object associated with class ω_i. In pattern classification problem, if A is a set of classes (e.g. $A = \omega_i \cup \omega_j$), $m(A)$ can be used to characterize the imprecision (partial ignorance) degree among the class ω_i and ω_j in classification of the object. $m(\Omega)$ denotes the total ignorance degree, and it usually plays a particular neutral role in the fusion process, because $m(\Omega) = 1$ characterizes the vacuous belief source of evidence.

The lower and upper bounds of imprecise probability associated with a BBA respectively correspond to the belief function $Bel(.)$ and the plausibility function $Pl(.)$ defined $\forall A \subseteq \Omega$ by (see [9])

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

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

In a multi-classifier system, the output of each classifier can be considered as an evidence represented by a BBA. The well-known Dempster's rule (often called DS rule) is still widely applied for combining multiple BBA's mainly because of its commutative and associative properties, which makes it relatively easy to implement, and also because it offers a compromise between the specificity and complexity for the combination of BBA's. The DS combination of two distinct sources of evidence characterized by the BBA's $m_1(.)$ and $m_2(.)$ over 2^Ω is denoted $\mathbf{m} = \mathbf{m}_1 \oplus \mathbf{m}_2$, and it is mathematically defined (assuming the denominator is not equal to zero) by $m(\emptyset) = 0$, and $\forall A \neq \emptyset \in 2^\Omega$ by

$$m(A) = \frac{1}{1 - K_{12}} \sum_{B,C\in 2^\Omega | B\cap C = A} m_1(B) m_2(C) \tag{4}$$

where $K_{12} \triangleq \sum_{B,C\in 2^\Omega | B\cap C=\emptyset} m_1(B) m_2(C)$ is the total conjunctive conflicting mass.

In DS formula (4), the total conflicting mass K_{12} is redistributed back to all the focal elements due to choice of the normalization. This choice of normalization (conflicting mass redistribution) can however generate unreasonable results, specially in the high conflicting cases [14], but also in some special low conflicting cases [23] as well. So a number of alternative combination rules have been developed to overcome the limitations of DS rule, like Proportional Conflict Redistribution (PCR) rules [14]. These modified rules are unfortunately less attractive from the implementation standpoint because even if they provide better fusion results, they are much more complicate and not associative.

In the combination of multiple sources of evidence cor-

[1] c being the number of classes in the framework of discernment of the problem under concern.

responding to different classifiers, each source may have different reliabilitiy factors. The classical Shafer's discounting method was introduced in [9] to deal with the unreliable source of evidence, and it discounts the partial mass of belief in a BBA to the total ignorance according to the reliability factor. In Shafer's discounting method [9], the reliability of one source of evidence is described by a single number in $[0, 1]$, and the mass values of different focal elements are discounted with the same number. A contextual discounting operation considered as a general extension of the classical discounting has been developed in [22]. It allows to take into account the refined reliability knowledge, which is represented by a vector of discounting rates characterizing the reliability of source associated with different hypotheses (contexts). The contextual discounting operation is suitable for handling the cases where the reliability of source of information mainly depends on the truth of the object to be classified. However, such prior reliability knowledge is usually very difficult to obtain in the real applications, that is why a new method is proposed in the next section of this paper.

III. Refined reliability evaluation (RRE) and Classifier Fusion

By convention, the objet to classify is denoted $\mathbf{y}$, its real class is denoted $c(\mathbf{y})$, and its estimated class declared by a classifier C_n is denoted $\hat{c}_n(\mathbf{y})$. In this work, we focus on the combination of multiple classifiers trained on different attribute sets. The class $c(\mathbf{y})$ of the object $\mathbf{y}$ to classify is assumed to belong to the frame of discernment $\Omega = \{\omega_1, \ldots, \omega_c\}$. We consider N classifiers, $C_1, \ldots, C_N$ trained on N different attribute spaces $\mathbb{S}_1, \ldots, \mathbb{S}_N$. Each classifier C_n provides as output a probabilistic mass function (pmf) denoted $\boldsymbol{\mu}_n \triangleq [\mu_n(1), \ldots, \mu_n(c)]$ based on the attribute knowledge of object in $\mathbb{S}_n$, where $\mu_n(i) \triangleq P(\hat{c}(\mathbf{y}) = \omega_i|\mathbb{S}_n)$, $i = 1, \ldots, c$. The value $\mu_n(i)$ represents the probability of the object belonging to the class ω_i estimated by the classifier. The classification performance can be improved in taking into account the quality of the classifier, which can be captured by the refined reliability evaluation of the output (pmf) $\boldsymbol{\mu}_n$ of each classifier. Then the output $\boldsymbol{\mu}_n$ will be modified accordingly before entering the classifier fusion process to make the final class decision.

A. Refined reliability evaluation

In this section, we will propose a very refined reliability evaluation method. In a c-class problem, the classification result of an object $\mathbf{y}$ by classifier C_n in the attribute space $\mathbb{S}_n$ is given as $\boldsymbol{\mu}_n$. The reliability of $\boldsymbol{\mu}_n$ is denoted by a matrix $\mathbf{R}_{c\times c}$[2] called reliability matrix, and this matrix expresses the conditional probability of the object $\mathbf{y}$ potentially belonging to class $\omega_i, i = 1, \ldots, c$ when it is classified to class $\omega_j, j = 1, \ldots, c$ by classifier C_n, i.e. $r_{ji} \triangleq P(c(\mathbf{y}) = \omega_i|\hat{c}(\mathbf{y}) = \omega_j)$.

Obviously, if this reliability matrix $\mathbf{R}$ can be well estimated, the accuracy of the classification result $\boldsymbol{\mu}_n$ could be efficiently improved taking into account this important knowledge. Now we will show how to estimate this reliability matrix $\mathbf{R}$.

Because the knowledge about the true class of the object is unavailable in the classification task, we will attempt to estimate the reliability matrix using the training knowledge. In the training data space, the patterns in the nearby neighborhood of the object $\mathbf{y}$ generally have the close attribute values with the object. Thus, the given classifier is expected to produce the similar performance on the object and on its close neighbors. Meanwhile, the ground truth of the class of the training patterns is always known. So the training data lying in the neighborhood of the object will be employed here for the reliability evaluation.

The K nearest neighbors (training patterns) of $\mathbf{y}$ are found at first in the attribute space $\mathbb{S}_n$. The selected neighbors denoted $\mathbf{x}_k$, $k = 1\ldots, K$ will be classified by the given base classifier C_n[3], and the classification result $\hat{c}_n(\mathbf{x}_k)$ provided by C_n is represented by the vector $\mathbf{P}_k = [P_k(1), \ldots, P_k(c)]$, where $P_k(i) \triangleq P(\hat{c}(\mathbf{x}_k) = \omega_i)$ is the estimated probability of $\mathbf{x}_k$ belonging to the class ω_i, for $i = 1, \ldots, c$.

If a neighbor $\mathbf{x}_k$ with the real class label ω_i (i.e. $c(\mathbf{x}_k) = \omega_i$) is classified by the base classifier into class ω_j (i.e. $\hat{c}(\mathbf{x}_k) = \omega_j$) with the corresponding probability $P_k(j)$, it indicates that the conditional probability of $\mathbf{x}_k$ classified to ω_j is $P_k(j)$ knowing $\mathbf{x}_k$ truly lies in ω_i as $P_k(j) \triangleq P(\hat{c}(\mathbf{x}_k) = \omega_j|c(\mathbf{x}_k) = \omega_i)$. Because $\mathbf{x}_k$ is a close neighbor of the object $\mathbf{y}$, the given classifier C_n likely produces the similar performance on $\mathbf{x}_k$ and $\mathbf{y}$. We can estimate the conditional probability of the object $\mathbf{y}$ classified to ω_j if its real class label is ω_i, i.e $P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_i)$, according to $P(\hat{c}(\mathbf{x}_k) = \omega_j|c(\mathbf{x}_k) = \omega_i)$.

Moreover, there may be multiple patterns with the real class label ω_i in the K selected neighbors, and all of them will be employed to estimate $P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_i)$. Meanwhile, the distance[4] between the object $\mathbf{y}$ and the neighbor $\mathbf{x}_k$ must be additionally taken into account in the calculation of $P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_i)$. If $\mathbf{y}$ is far from $\mathbf{x}_k$, then $\mathbf{x}_k$ is considered with a small influence on the estimation. Thus, the bigger distance, the smaller weight of the neighbor. The weighted sums of the conditional probabilities of the neighbors $\mathbf{x}_k$ belonging to class ω_i but classified to ω_j (denoted by β_{ij}) is computed by

$$\begin{aligned}\beta_{ij} &= \sum_{\mathbf{x}_k} P(\hat{c}(\mathbf{x}) = \omega_j|c(\mathbf{x}_k) = \omega_i) \cdot \delta_k \\ &= \sum_{\mathbf{x}_k|c(\mathbf{x}_k)=\omega_i} P_k(j) \cdot \delta_k \end{aligned} \tag{5}$$

with

$$\delta_k = e^{-\gamma \cdot d_k} \tag{6}$$

[2] For notation convenience, the classifier index n is omitted in the sequel.

[3] The base classifier can be selected according to the actual application, like Artificial neural network, Bayesian classifier, etc. The classifier can work with probabilistic framework or belief functions framework. In this work, we just consider the belief-based classifier with the output represented by a simple BBA, which includes the singleton focal elements and only one ignorant element. The evidential neural network [15] classifier producing the simple BBA's as output is employed as base classifier in our sequel simulations.

[4] The Euclidean distance is used here.

$$d_k \triangleq \frac{d(\mathbf{y},\mathbf{x}_k)}{\min\limits_{k\in[1,K]} d(\mathbf{y},\mathbf{x}_k)} \tag{7}$$

where δ_k denotes the distance weights, and γ is a tuning parameter used to control the influence of distance, and d_k is the relative distance of the object to the neighbor $\mathbf{x}_k$ with respect to the minimum distance to the nearest neighbors.

β_{ij} can be interpreted as the weighting factor of the hypothesis that the object is really from class ω_i but classified to ω_j. The conditional probability $P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_i)$ should be proportional to β_{ij} as $P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_i) \propto \beta_{ij}$, and it is defined by $P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_i) = \rho\beta_{ij}$ ($\rho \in (0,1]$ being a positive proportional coefficient). Then the reliability matrix $\mathbf{R}$ expressed by the probability $r_{ji} \triangleq P(c(\mathbf{y}) = \omega_i|\hat{c}(\mathbf{y}) = \omega_j)$ can be easily derived according to Bayes rule, one gets

$$\begin{aligned} r_{ji} =& P(c(\mathbf{y}) = \omega_i|\hat{c}(\mathbf{y}) = \omega_j) \\ =& \frac{P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_i)P(c(\mathbf{y}) = \omega_i)}{\sum\limits_{l=1}^{c} P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_l)P(c(\mathbf{y}) = \omega_l)} \end{aligned} \tag{8}$$

Without extra knowledge, the priori probability $P(c(\mathbf{y}) = \omega_l), l = 1,\ldots,c$ is usually assumed uniformly distributed. Therefore, the probability $P(c(\mathbf{y}) = \omega_i|\hat{c}(\mathbf{y}) = \omega_j)$ can be obtained by

$$\begin{aligned} r_{ji} =& \frac{P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_i)}{\sum\limits_{l=1}^{c} P(\hat{c}(\mathbf{y}) = \omega_j|c(\mathbf{y}) = \omega_l)} \\ =& \frac{\rho\beta_{ij}}{\rho\sum\limits_{l=1}^{c}\beta_{lj}} \\ =& \frac{\beta_{ij}}{\sum\limits_{l=1}^{c}\beta_{lj}} \end{aligned} \tag{9}$$

Then the reliability matrix $\mathbf{R}$ is determined, and we will modify the classification result $\boldsymbol{\mu}_n$ to make it closer to the potential truth using $\mathbf{R}$. We recall that the matrix is estimated according to a limited number of neighborhoods of the object to classify. Generally, there are more or less differences between the object and these neighbors. Therefore we must not be completely confident about the estimation of this matrix for revising the classifier result of the object $\mathbf{y}$, and the confidence degree about this matrix seems quite difficult to obtain. That is why we propose a very cautious discounting method to transfer the classification knowledge to the associated partial ignorance (e.g. $\omega_i \cup \omega_j$) rather than to the specific class (e.g. ω_i). By doing this, one can efficiently reduce the risk of misclassification error at the price of partial imprecision, and such imprecision can be specified through the combination with other classifiers. More specifically, the contribution of belief from the classifier output $\mu_n(j)$ and the probability $P(c(\mathbf{y}) = \omega_i|\hat{c}(\mathbf{y}) = \omega_j)$ is transferred by

$$m_{n1}(\omega_i \cup \omega_j) = P(c(\mathbf{y}) = \omega_i|\hat{c}(\mathbf{y}) = \omega_j)\cdot\mu_n(j) \tag{10}$$

$\omega_i \cup \omega_j$ represents the imprecision between ω_i and ω_j, and it plays a neutral role in the classification between ω_i and ω_j.

Another contribution of belief on $\omega_j \cup \omega_i$, $j \neq i$ is also obtained from $\mu_n(i)$ by considering

$$m_{n2}(\omega_j \cup \omega_i) = P(c(\mathbf{y}) = \omega_j|\hat{c}(\mathbf{y}) = \omega_i)\cdot\mu_n(i) \tag{11}$$

So that the discounted BBA derived from $\boldsymbol{\mu}_n$ is given for $i = 1,\ldots,c$ and $j = 1,\ldots,c$

$$\begin{aligned} m_n(\omega_i \cup \omega_j) &= m_{n1}(\omega_i \cup \omega_j) + m_{n2}(\omega_j \cup \omega_i) \\ &= P(c(\mathbf{y}) = \omega_i|\hat{c}(\mathbf{y}) = \omega_j)\cdot\mu_n(j) \\ &+ P(c(\mathbf{y}) = \omega_j|\hat{c}(\mathbf{y}) = \omega_i)\cdot\mu_n(i), \quad \text{if} \quad i \neq j \end{aligned} \tag{12}$$

$$m_n(\omega_i) = P(c(\mathbf{y}) = \omega_i|\hat{c}(\mathbf{y}) = \omega_i)\cdot\mu_n(i), \text{ if } \quad j = i. \tag{13}$$

One can see that some partial imprecision has arisen due to the cautious discounting operation, but these imprecise information will be clarified by the combination with other classifiers in the sequel.

If the probabilities of the K neighbors committed to ω_i are all zeros. In this case, the probability of the object $\mathbf{y}$ belonging to ω_i, i.e. $\mu_n(i)$, will be discounted to total ignorance by taking

$$m_n(\Omega) = 1 - \sum_{A\subset\Omega} m_n(A) \tag{14}$$

$m_n(\Omega)$ captures the total ignorant information about the classification done by the classifier C_n, and it plays a neutral role in the combination with the (modified) output of other classifiers. In fact, $m_n(\Omega)$ will always be redistributed to other more specific focal elements in the classifier fusion process based on the conjunctive rule of combination.

To show how formulas (12), (13) and (14) work for making a cautious discounting, let us consider the following example.

Example 1: One assumes that the result obtained by the classifier C_n for one object is the following probability mass function (pmf): $\mu_n(1) = 0.5$, $\mu_n(2) = 0.3$, and $\mu_n(3) = 0.2$. Suppose that this object has three close neighbors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, and two of them $\mathbf{x}_1, \mathbf{x}_2$ are truly labeled by ω_1, and the third one $\mathbf{x}_3$ is labeled by ω_2. The three neighbors are respectively classified using classifier C_n, and the classification results of the $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are respectively given by the following pmf:

$$\begin{aligned} \boldsymbol{P}_1 : &P_1(1) \triangleq P(\hat{c}(\mathbf{x}_1) = \omega_1|c(\mathbf{x}_1) = \omega_1) = 0.9, \\ &P_1(2) \triangleq P(\hat{c}(\mathbf{x}_1) = \omega_2|c(\mathbf{x}_1) = \omega_1) = 0.1 \\ \boldsymbol{P}_2 : &P_2(1) \triangleq P(\hat{c}(\mathbf{x}_2) = \omega_1|c(\mathbf{x}_2) = \omega_1) = 0.6, \\ &P_2(2) \triangleq P(\hat{c}(\mathbf{x}_2) = \omega_2|c(\mathbf{x}_2) = \omega_1) = 0.4 \\ \boldsymbol{P}_3 : &P_3(1) \triangleq P(\hat{c}(\mathbf{x}_3) = \omega_1|c(\mathbf{x}_3) = \omega_2) = 0.2, \\ &P_3(2) \triangleq P(\hat{c}(\mathbf{x}_3) = \omega_2|c(\mathbf{x}_3) = \omega_2) = 0.8 \end{aligned}$$

In the reliability evaluation, the distance weights of $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ can be easily calculated using eq. (6), and let us assume that we have $\delta_1 = 0.6$, $\delta_2 = 0.3$, and $\delta_3 = 0.5$.

So we can estimate the weighted sums of the conditional probability of the object classified to ω_j if the real class is ω_i

$i = 1, 2, 3$ using eq.(5) as

$$\begin{aligned}\beta_{11} =& 0.6 \times 0.9 + 0.3 \times 0.6 = 0.72\\ \beta_{12} =& 0.6 \times 0.1 + 0.3 \times 0.4 = 0.18\\ \beta_{21} =& 0.5 \times 0.2 = 0.1\\ \beta_{22} =& 0.5 \times 0.8 = 0.4\end{aligned}$$

Then conditional probability of the object classified to ω_i but truly coming from ω_g can be derived by eq.(9).

$$\begin{aligned}r_{11} &= \frac{\beta_{11}}{\beta_{11}+\beta_{21}} = 0.88\\ r_{12} &= \frac{\beta_{21}}{\beta_{21}+\beta_{11}} = 0.12\\ r_{21} &= \frac{\beta_{12}}{\beta_{12}+\beta_{22}} = 0.31\\ r_{22} &= \frac{\beta_{22}}{\beta_{22}+\beta_{12}} = 0.69\end{aligned}$$

Then the classification result μ_n of the object obtained by the classifier C_n will be cautiously discounted based on the reliability matrix $\mathbf{R}$ using eq.(12), (13) and eq. (14). One finally gets:

$$\begin{aligned}m_n(\omega_1) =& r_{11} \cdot \mu_n(1) = 0.88 \times 0.5 = 0.44;\\ m_n(\omega_2) =& r_{22} \cdot \mu_n(2) = 0.69 \times 0.3 = 0.21;\\ m_n(\omega_1 \cup \omega_2) =& r_{12} \cdot \mu_n(1) + r_{21} \cdot \mu_n(2)\\ =& 0.12 \times 0.5 + 0.31 \times 0.3 = 0.15;\\ m_n(\Omega) =& 1 - 0.44 - 0.21 - 0.15 = 0.20\end{aligned}$$

In fact, the probability value $\mu_n(3)$ is transferred to the mass of ignorance $m_n(\Omega)$, since no neighbors are from class ω_3 and the probability of the selected three neighbors committed to class ω_3 is zero. After the cautious discounting operation, one observes also that some masses of beliefs are transferred to the partial ignorant element (i.e. $\omega_1 \cup \omega_2$), and such imprecise information can be specified by the combination with other classifiers. By doing this, one can reduce the classification error rate using the complementarity of the classifiers.

B. Classifier fusion process and decision-making

The popular DS rule defined by the formula (4) requiring relatively small computation burden is often used to combine the uncertain and imprecise information, and it will be employed here to combine the discounted classification results from different classifiers. Since DS rule is associative, the BBA's can be combined sequentially in any sequence order. In the final fusion results, some beliefs may remain in the (partial) imprecise focal element (imprecise classes) due to the discounting procedure. So the plausibility functions $Pl(.)$ taking into account all the beliefs of the associated classes is used here for decision making support, and the object is considered belonging to the class receiving the biggest plausibility value, e.g. ω_g satisfying $\omega_g = \arg\max_j Pl(\omega_j)$.

C. Guideline for parameters tuning

In this new CF-RRE method, the parameter γ involved in eq.(5) should be tuned in the real applications. γ is used to penalize the influence of the neighbors in the determination of the reliability according to the distance between the object and its neighbor. The bigger γ value, the smaller influence of the neighbor (through its distance to the object) for the reliability evaluation. According to many heuristics tested with various real data sets, we find that γ must belong to $[5, 20]$ in practice, and we recommend to take $\gamma = 10$ as default value. In applications, the tuning parameters γ can be optimized by cross validation in the training data space, and the optimized value corresponding to the highest accuracy can be chosen.

IV. Experiment applications

The classification performance of this new CF-RRE method is evaluated by comparisons with several other fusion methods including weighted majority voting (WMV), weighted averaging fusion (WAF) and weighted DS (WDS) combination rule. Here we will test the formulas (15)-(16) that are commonly used in practice to calculate the weighting factors based on the classification accuracy η (see [21]).

$$w_n = \frac{\eta_n}{\sum\limits_l \eta_l} \tag{15}$$

$$w_n = \frac{\eta_n - \eta_W}{\eta_B - \eta_W}, \tag{16}$$

where $\eta_B \triangleq \max\limits_n \eta_n$, $\eta_W \triangleq \min\limits_n \eta_n$, $\eta_n \triangleq \frac{N_c}{T}$, and where N_c is the number of patterns correctly classified, and T is the number of patterns to classify. We consider here the local accuracy η_n, which is calculated according to $T = K$ nearest neighbors of objects in training data space. η_n denotes the individual local accuracy of the classifier C_n.

The three used fusion methods including WMV, WAF and WDS are briefly explained here for comparisons in this work.

- In WMV rule, the fusion result is calculated by $\mathbf{l} = \sum\limits_{n=1}^{N} w_n \mathbf{l}_n$, and $\mathbf{l}_n$ is the hard classification result of classifier C_n.
- In WAF method, the fusion is defined as $\mathbf{p} = \sum\limits_{n=1}^{N} w_n \mathbf{p}_n$, and $\mathbf{p}_n$ is the output of classifier C_n.
- In WDS method, the classifiers are combined by $\mathbf{m} = {}^{\alpha_1}\mathbf{m}_1 \oplus \ldots \oplus {}^{\alpha_N}\mathbf{m}_N$ with $\alpha_n = \frac{w_n}{\max\limits_i w_i}$. The BBA ${}^{\alpha_n}\mathbf{m}_n$ denotes the BBA $\mathbf{m}_n$ discounted using Shafer's discounting rule [9] with the reliability factor α_n. The weights w_n $(n = 1, \ldots, N)$ are normalized to make the sum of fusion result equal to one.

The base classifier can be selected according to the actual applications. In this work, the Evidential neural network (ENN) [15] classifier is employed as the base classifiers[5],

[5] Any other classical classifiers can be also be used here as base classifier, and the selection of proper base classifier mainly depends on the actual application, which is out of scope of this paper.

since it usually produces good performance. The base classifier(s) will be respectively trained using different subsets of attributes, and the multiple classification results obtained by different classifiers will be fused for classifying the objects. In this work, the predicted class of the object (i.e. the final decision made) corresponds to the class that has received (after the classifier fusion) the maximum of plausibility.

Five real data sets from UCI repository [26] have been used in this work to evaluate the performance of this new CF-RRE method, and to compare it with respect to other three fusion methods. The basic knowledge of the used data sets are shown by Table I. The patterns in these data sets contain multiple attributes. For each data set, the whole set of attributes will be randomly divided into N distinct subsets[6], and each subset of attributes will be respectively used to train the base classifier. For example, Texture data set has 40 attributes that can be divided into 4 distinct sub-sets, and each subset contains 10 attributes. The base classifier ENN will be respectively learnt based on each subset of attributes.

The k-fold cross validation is often used for the classification performance evaluation, but k remains a free parameter. We use the simplest 2-fold cross validation here, since the training and test sets are large, and each sample can be respectively used for training and testing on each fold. In the K nearest neighbors selection, we have tested the classification performance with the K value ranging from 5 to 20 for the local weighted fusion methods, and for our proposed CF-RRE method. The two derivations of weights according to (15)-(16) have been tested and the best results are reported. In the CF-RRE method, the parameters $\gamma \in [5, 20]$ can be optimized using the training data, and optimized value corresponding to the highest accuracy is adopted. The average classification results (mean accuracy value) with $K \in [5, 20]$ for different methods are reported in Table II, and the accuracy curves with the different K values in different methods are shown by Figs. 1 and 2.

In Table II, the N value is the number of classifiers, and each classifier corresponds to a subset of attributes. AC_l and AC_u represent respectively the lower and upper bounds of the classification accuracy of these individual classifiers that are combined, and the accuracy is calculated by $AC = \frac{N_c}{T}$ where N_c is the number of correctly classified object, and T is number of total test patterns.

The analysis of the results of Table II shows that all the used fusion approaches generally improve the classification accuracy with respect to the individual classifier. This demonstrates the advantage and interest of combining classifiers. Meanwhile, one can see that this new CF-REE method produces much higher accuracy rate than other methods thanks to the use of the refined reliability evaluation strategy. In the other weighted fusion methods, the weighting factors mainly depend on the overall performance of classifier, and the refined classification knowledge (e.g. the variety of the misclassifications) is ignored. In the new method, the different misclassification cases of neighborhoods play different roles in the cautious discounting of classification results of object.

In the Figs. 1 and 2, the x-axis represents the number of K value, whereas y-axis corresponds to the accuracy. According to the Figs. 1 and 2, we also observe that the classification performance of the new method is not very sensitive to the K value contrary to the other methods. This is because the influence of the distance from the object to its neighbors is additionally taken into account. The farther distance from a neighbor to the object will yield the smaller weight (influence) of this neighbor in the reliability evaluation. So the neighbors which are quite far from the object will have very little influence on the classification of the object. The experiment results show that the new method is robust with respect to the K value. Thus, the K value can be easily selected in the real applications for this new method.

For each data set, we have considered two cases with different number of classifiers. We find that the bigger number of classifiers does not necessarily lead to higher accuracy. So the proper selection of classifiers for the fusion procedure according to the reliability evaluation may be an interesting topic to investigate in the future.

Table I
BASIC INFORMATION OF THE USED DATA SETS.

Data	Classes	Attributes	Instances
Texture (Te)	11	40	5500
Vehicle (Ve)	4	18	946
Movement-libras (ML)	15	90	360
Sonar (So)	2	60	208
Segment (Se)	7	19	2310

Table II
CLASSIFICATION RESULTS OF DIFFERENT METHODS WITH ENN CLASSIFIER (IN %).

Data	N	$[AC_l,\ AC_u]$	WMV	WAF	WDS	NEW
Te	4	[59.18, 65.35]	81.62	80.87	83.13	95.38
Te	8	[53.00, 68.36]	84.59	82.59	85.28	94.44
Ve	2	[38.18, 49.53]	52.25	51.57	51.56	65.17
Ve	6	[38.53, 49.29]	55.16	52.40	55.78	64.93
ML	9	[26.67, 44.17]	55.38	54.36	61.55	76.68
ML	15	[24.44, 38.06]	56.56	49.03	62.10	71.65
So	6	[53.37, 73.08]	74.37	73.29	77.31	81.97
So	20	[53.37, 74.04]	72.57	71.36	75.81	78.85
Se	7	[32.73, 67.10]	76.60	80.28	80.61	90.70
Se	2	[63.72, 69.87]	82.34	81.63	82.36	91.79

V. CONCLUSION

A new method for classifier fusion with refined reliability evaluation (CF-RRE) has been proposed based on belief function theory. The reliability represented by a matrix

[6] There is no overlapping attributes in different subsets.

$R_{c\times c}$ (c being the number of classes in the data set) is estimated based on the local classifier performance in the neighborhoods of the object. Each element of the reliability matrix characterizes the conditional probability of the object potentially belonging to class $\omega_i, i = 1, \ldots, c$ when it is classified to $\omega_j, j = 1, \ldots, c$ by the given classifier. Then the classification result is cautiously discounted according to the elements of reliability matrix, and the partial probability (or belief) of each class is prudently redistributed to the associated imprecise classes (i.e. the disjunction of several classes) under belief functions framework. This cautious discounting operation is able to reduce the error risk by modeling the imprecision, which can be specified by combining with other (more or less) complementary classifiers. The popular Dempster's rule (also called DS rule) is employed to globally fuse the discounted classification results provided by different classifiers. The uncertainty and imprecision of the individual classifiers can be efficiently decreased through the fusion procedure. The effectiveness of the new method has been validated by experiments using various real data sets with respect to several other related methods, and the new method is able to produce much higher accuracy than others. Some more base classifiers (e.g. Support Vector Machine, Bayesian classifier) and more real data sets will be used to further test the potential of the proposed method in our future work.

Acknowledgement

This work has been partially supported by National Natural Science Foundation of China (Nos.61135001, 61403310).

References

[1] 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), pp.353–374, 2011.

[2] Z.W. Yu, L. Li, J.M. Liu, G.Q. Han, *Hybrid Adaptive Classifier Ensemble*, IEEE Trans. on Cybernetics, Vol. 45(2), pp.177–190, 2015.

[3] L.I. Kuncheva, *A Theoretical Study on Six Classifier Fusion Strategies*, IEEE Trans. on Pattern Anal. Mach. Intell., Vol.24(2), 2002.

[4] L. Kuncheva, J. Bezdek, R. Duin, *Decision templates for multiple classifier fusion: an experimental comparison*, Pattern Recognition, Vol. 34(2), pp. 299–314, 2001.

[5] J.A. Hoeting, C.T. Volinsky, *Bayesian Model Averaging: A Tutorial*, Statistical Science, Vol.14(4):382–417, 2010.

[6] N.J. Pizzi, W. Pedrycz, *Aggregating multiple classification results using fuzzy integration and stochastic feature selection*, International Journal of Approximate Reasoning, Vol. 51(8), pp. 883–894, 2010.

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

[8] D. Ruta, B. Gabrys, *An Overview of Classifier Fusion Methods*, Computing and Information Systems, Vol.7, pp. 1–10, 2000.

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

[10] J.b. Yang, D.l. Xu, *Evidential reasoning rule for evidence combination*, Artif. Intell., Vol. 205, pp.1–29, 2013.

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

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

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

[14] F. Smarandache, J. Dezert, *Information Fusion Based on New Proportional Conflict Redistribution Rules*, in Proc. of Fusion 2005 Int. Conf. on Information Fusion, USA, 2005.

[15] T. Denœux, *A neural network classifier based on Dempster-Shafer theory*, IEEE Trans. on Systems, Man and Cybernetics A, Vol. 30, No. 2, pp. 131–150, 2000.

[16] Z.-g. Liu, Q. Pan, J. Dezert, G. Mercier, *Credal classification rule for uncertain data based on belief functions*, Pattern Recognition, Vol. 47, No. 7, pp. 2532–2541, 2014.

[17] Z.-g. Liu, Q. Pan, G. Mercier, J. Dezert, *A new incomplete pattern classification method based on evidential reasoning*, IEEE Trans. on Cybernetics, Vol. 45, No. 4, pp. 635–646, 2015.

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

[19] 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, pp.133–141, 2011.

[20] L. Xu, A. Krzyzak, C.Y. Suen, *Several methods for combining multiple classifiers and their applications in handwritten character recognition*, IEEE Trans. on System Man & Cybernetics, Vol. 2(3), pp. 418–435, 1992.

[21] 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.

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

[23] 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, No. 3, pp. 223–252, March 2014.

[24] J. Dezert, A. Tchamova, F. Smarandache, P. Konstantinova, *Target Type Tracking with PCR5 and Dempster's rules: A Comparative Analysis*, in Proc. of Fusion 2006 Int. Conf. on Information Fusion, Firenze, Italy, 2006.

[25] X.-R. Li, *Probability, Random Signals, and Statistics*, CRC Press, 1998.

[26] M. Lichman, UCI Machine Learning Repository http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science, 2013.

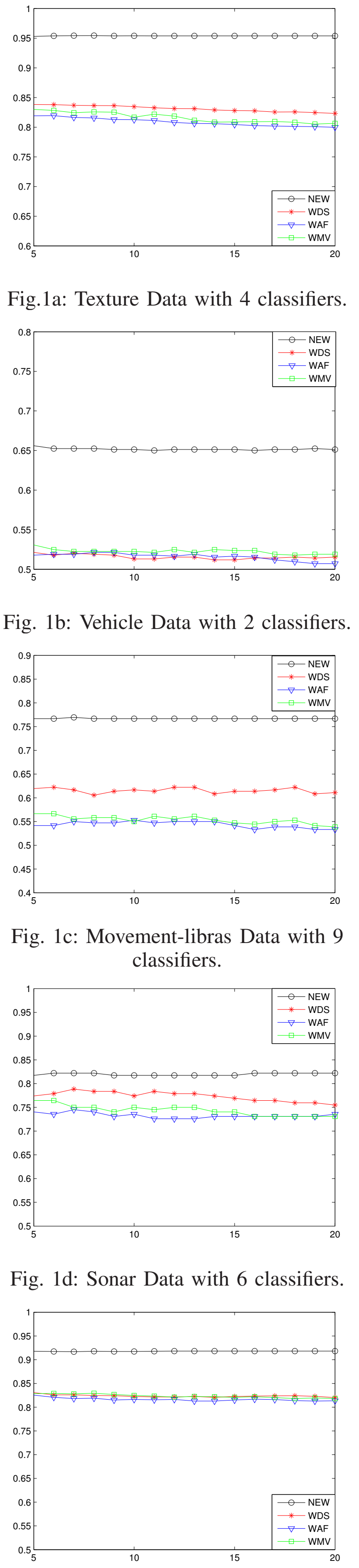

Fig.1a: Texture Data with 4 classifiers.

Fig. 1b: Vehicle Data with 2 classifiers.

Fig. 1c: Movement-libras Data with 9 classifiers.

Fig. 1d: Sonar Data with 6 classifiers.

Fig. 1e: Segment Data with 2 classifiers.

Figure 1. Classification results by fusion of few classifiers.

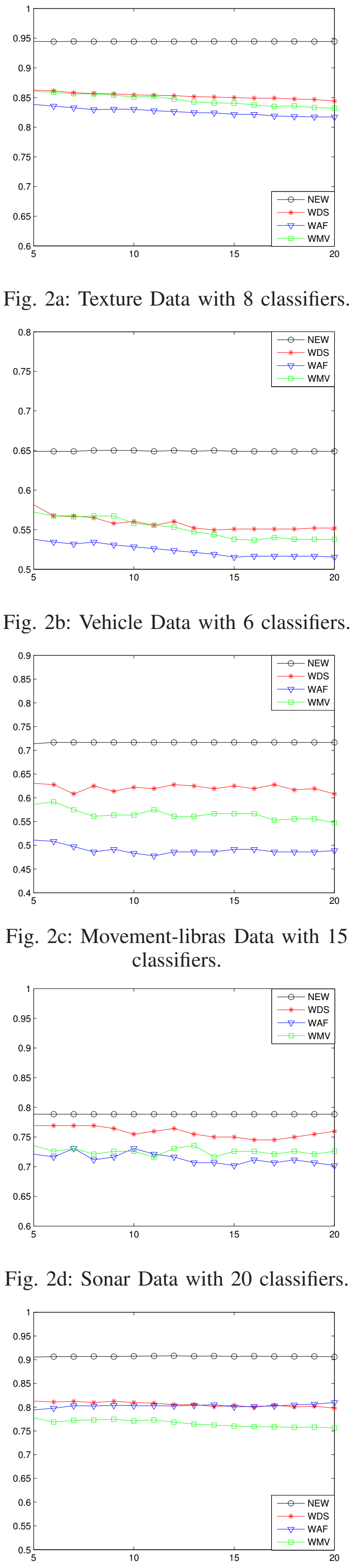

Fig. 2a: Texture Data with 8 classifiers.

Fig. 2b: Vehicle Data with 6 classifiers.

Fig. 2c: Movement-libras Data with 15 classifiers.

Fig. 2d: Sonar Data with 20 classifiers.

Fig. 2e: Segment Data with 7 classifiers.

Figure 2. Classification results by fusion of more classifiers.