Adaptively evidential weighted classifier combination

Liguo Fei, Bingyi Kang, Van-Nam Huynh and Yong Deng

Abstract—Classifier combination plays an important role in classification. Due to the efficiency to handle and fuse uncertain information, Dempster-Shafer evidence theory is widely used in multi-classifiers fusion. In this paper, a method of adaptively evidential weighted classifier combination is presented. In our proposed method, the output of each classifier is modelled by basic probability assignment (BPA). Then, the weights are determined adaptively for individual classifier according to the uncertainty degree of the corresponding BPA. The uncertainty degree is measured by a belief entropy, named as Deng entropy. Discounting-and-combination scheme in D-S theory is used to calculate the weighted BPAs and combine them for the final BPA for classification. The effectiveness of the proposed weighted combination method is illustrated by numerical experimental results.

Index Terms—Classifier combination, Dempster-Shafer evidence theory, Deng entropy, Classification, Weight.

I. INTRODUCTION

Classification is a method of integrated learning [1] which belongs to machine learning techniques [2], [3], [4], [5], [6], [7], [8] as one branch. It attracts much attention of researchers along with the perfection of the theoretical basis. And its application is widely published in different fields, such as text classification and retrieval [9], image recognition and speech recognition [10], [11]. There exist a large number of well-known classifiers: support vector machine (SVM) [6], [12], radial basis function (RBF) [13], naive Bayes (NB) [14], decision tree learner (REPTree), multilayer perceptron (MP), 1 nearest neighbor (1NN, or IB1), and RBFnetwork (RBFN). NB and SVM are in the top ten data-mining algorithms [15]. However, it is noteworthy that the ability to collect and deal with information for a single classifier is limited [16]. Moreover, this limitation has a serious impact to the accuracy of the classification results [17]. On the other hand, it’s apparent that there exist a lot of patterns that cannot be classified using different learning algorithms or techniques in the classification systems. And these sets of patterns will not overlap necessarily [18]. It means that different classifiers can provide different information from different aspects, which can complement each other for better classification results [19]. In other words, the combination of different classifiers is more beneficial to take advantages of their own strengths to improve the quality of the classification.

Taking notice of the significance and the potential applications of classifiers combination, more and more researches and exploration are done to build an ensemble classifier [20] which could perfect the performance of the individual classifier. Fattah et al. [21] presented the comprehensive investigation of different proposed new term weighting schemes for sentiment classification, and exploit the class space density based on the class distribution in the whole documents set as well as in the class documents set. Dlez-Pastor et al. [22] proposed a new approach to build ensembles of classifiers for two-class imbalanced data sets which can lead to larger AUC compared to other ensembles of classifiers. Ahmadvand et al. [23] applied the combination of multiple classifiers to medical image processing to supervise the segmentation of MRI brain images. Moosavian et al. [24] put forward a new method to recognize the spark plug fault based on sensor fusion and classifier combination using Dempster-Shafer evidence theory. Due to the effectiveness to handle uncertainty, D-S theory is paid more and more attention in multi-classifiers fusion. Yager et al. [25] proposed the ordered weighted averaging (OWA) to aggregate the information in the uncertainty profile for obtaining representative values in decision-making. Quost et al. [26] presented optimized t-norm in the Dempster-Shafer framework based combination rules to combine non independent classifiers. Marek et al. [27] built ensemble classifiers using belief functions and OWA operators for classification.

Recently, Huynh et al. [28] presented an evidential reasoning based framework for weighted combination of classifiers for word sense disambiguation (WSD). Within this framework, the probability distributions (PD) are obtained from multi-classifiers. Then, the authors presented a method to weight the PDs for discounting their own uncertainty measured by Shannon entropy. Next, the BPAs are determined from each classifier’s PD by the discounting operation. Finally, all obtained BPAs are combined using Dempster’s rule to obtain the final results as the ensemble classifier for classification. What is certain is that the evidential reasoning based framework conducts itself well for WSD than others con-generic method by their experimental results. However,
this algorithm still has its limitations to handle more general case. The method of Huynh et al. [28] obtains PDs firstly. However, the output of each classifier may be BPA due to the high uncertain environment. In other words, the output \( \psi_i(x) \) can be BPAs directly instead of a posterior probability distribution on \( \varphi \) in many practical application. In these situations, the method of Huynh et al. [28] will be no more applicable. To address this issue, we proposed a new evidential reasoning based framework based on D-S theory. And the process of the two methods are comparing in Figure 1.

Comparing with the method of Huynh et al. [28], our proposed method deals with the BPAs from classifiers directly. It is recognized that BPA itself exists uncertainty degree, and the higher uncertainty degree of the BPA, the less information provided by the output of a classifier and then the lower weight it should be assigned. The weighting process is obvious different in the two methods. A new method named Deng entropy [29] is utilized to measure the uncertainty degree of BPAs, and the weight with regard to each classifier are defined adaptively based on the input pattern under classification. Finally, we combine multi-classifiers with Dempster’s rule based on the weighed BPAs. In conclusion, there are two major improvements in the proposed method comparing with the method of Huynh et al. [28]. The first one is that we use BPAs instead of PDs to represent more uncertain information. The second one is that the Deng entropy is made use of to determine the weights of multi-classifiers. It should be pointed out the proposed method can be seen as the generalization of the method of Huynh et al. [28]. If the output of classifiers are PDs, the proposed method degenerated as the method in [28]. From this aspect, the proposed method is more efficient to handle uncertain information. In addition, one of the advantages of the proposed method keeps obtaining the BPAs dynamic with the changes of the output of classifiers. Then the weights and the weighted BPAs also change adaptively. This fully embodies the characteristics of our method adaptive and this recognizes the adaptive quality of our proposed method profoundly.

The organization of the rest of this paper is as follows. Section 2 starts with a brief presentation of the D-S theory and its basic rules and some necessary related concepts. The proposed method for the D-S theory based framework for weighted combination of classifiers is presented in Section 3. Section 4 presents and analyzes the experimental results. Conclusion is presented in Section 5.

A. Dempster-Shafer evidence theory

Dempster-Shafer evidence theory (D-S theory) is proposed by Dempster and developed later by Shafer [30], [31]. This theory extends the elementary event space in probability theory to its power set named as frame of discernment and constructs the basic probability assignment(BPA) on it. In addition, there is a combination rule presented by Dempster to fuse different BPAs. In particular, D-S theory can definitely degenerate to the probability theory if the belief is only assigned to single elements. Therefore, the D-S theory is the generalization of probability theory with the purpose of handling uncertainty and is widely used to uncertainty modeling [32], [33], [34], decision making [35], [36], [37], [38], [39], [40], information fusion [41], [42] and uncertain information processing [43], [44], [45]. The basic definitions about D-S theory is shown as follows:

1) Frame of discernment: D-S theory supposes the definition of a set of elementary hypotheses called the frame of discernment, defined as:

\[
\theta = \{H_1, H_2, \ldots, H_N\}
\]

That is, \( \theta \) is a set of mutually exclusive and collectively exhaustive events. Let us denote \( 2^\theta \) the power set of \( \theta \).

2) Mass functions: When the frame of discernment is determined, a mass function \( m \) is defined as follows.

\[
m : 2^\theta \rightarrow [0, 1]
\]

which satisfies the following conditions:

\[
m(\emptyset) = 0
\]

\[
\sum_{A \in 2^\theta} m(A) = 1
\]

In D-S theory, a mass function is also called a basic probability assignment (BPA).

3) Evidence discounting: The discounting operation is used when an evidence provides a BPA, but the evidence is believed by probability \( \alpha \). In this circumstance, The BPA \( m^\alpha \) is redefined based on the probability of reliability \( \alpha \) as follows

\[
m^\alpha(A) = \alpha \times m(A), \quad A \subset \theta
\]

\[
m^\alpha(\emptyset) = (1 - \alpha) + \alpha \times m(\emptyset)
\]

where \( A \) is the focal element, and \( m \) is the mass function.

4) Dempster’s rule of combination: In a real system, there may be many evidence originating from different sensors, so we can get different BPAs. Dempster [31] proposed orthogonal sum to combine these BPAs. Suppose \( m_1 \) and \( m_2 \) are two mass functions. The Dempster’s rule of combination denoted by \( m = m_1 \oplus m_2 \) is defined as follows:

\[
m(A) = \frac{\sum_{B \cap C = A} m_1(B)m_2(C)}{1 - K}
\]

with

\[
K = \sum_{B \cap C = \emptyset} m_1(B)m_2(C)
\]

Note that the Dempster’s rule of combination is only applicable to such two BPAs which satisfy the condition \( K < 1 \).
Fig. 1. The comparison between the proposed method and the method in [28]
B. Deng entropy

Deng entropy [29] is presented to measure the uncertainty degree of basic probability assignment as a generalized Shannon entropy in D-S evidence theory. Deng entropy can be described as follows

$$E_d = - \sum_i m(F_i) \log \frac{m(F_i)}{|F_i| - 1}$$

(9)

where $F_i$ is a proposition in mass function $m$, and $|F_i|$ is the cardinality of $F_i$. Deng entropy is similar with Shannon entropy in form. The difference is that the belief for each proposition $F_i$ is divided by a term $(2^{F_i} - 1)$ which represents the potential number of states in $F_i$ (the empty set is not included). So Deng entropy is the generalization of Shannon entropy, which is used to measure the uncertainty degree of BPA [29].

Specially, Deng entropy can definitely degenerate to Shannon entropy as follows

$$E_d = - \sum_i m(\theta_i) \log \frac{m(\theta_i)}{2^{\theta_i} - 1} = - \sum_i m(\theta_i) \log m(\theta_i)$$

(10)

Numerical examples are given to illustrate the computational process of Deng entropy.

Example I.1. Suppose a frame for discernment $X = \{S, C, V\}$, for a mass function $m(C) = 0.5554, m(C, V) = 0.4420, m(S, C, V) = 0.0026$.

$$E_d = -0.5554 \times \log \frac{0.5554}{2 - 1} - 0.4420 \times \log \frac{0.4420}{2 - 1} - 0.0026 \times \log \frac{0.0026}{2 - 1} = 1.7220$$

Example I.2. And another mass function $m(S) = m(C) = m(V) = 1/19, m(S, C) = m(S, V) = m(C, V) = 3/19, m(S, C, V) = 7/19$.

$$E_d = - \frac{1}{19} \times \log \frac{1/19}{2 - 1} - \frac{1}{19} \times \log \frac{1/19}{2 - 1} - \frac{1}{19} \times \log \frac{1/19}{2 - 1} - \frac{3}{19} \times \log \frac{1/19}{2 - 1} - \frac{3}{19} \times \log \frac{3/19}{2 - 1} - \frac{3}{19} \times \log \frac{3/19}{2 - 1} - \frac{7}{19} \times \log \frac{7/19}{2 - 1} = 4.2479$$

II. THE PROPOSED METHOD OF WEIGHTED COMBINATION OF CLASSIFIERS

Let us suppose that there are $M$ classes in the decision system representing as $\varphi = \{c_1,...,c_M\}$. Also suppose that there are $R$ classifiers $\psi_i (i = 1,...,R)$ can be used for combination. For each input pattern $x$, let us denote by

$$\psi_i(x) = [m_{i1}(x),...,m_{iM}(x)]$$

the right-hand side of this equality is a mass function obtained from $ith$ classifier. We determine the BPA of each classifier from the selected training set using the normal distribution method which is mentioned above.

Each BPA $\psi_i(x)$ is now considered as the belief degree distribution derived from information source provided by classifier $\psi_i$ for classifying $x$. However, the evidence has a certain extent uncertainty by itself resulting in a decline in the degree of trust. Therefore, it is necessary to quantify somehow the quality of information offering form $\psi_i$ regarding the classification of $x$ and to consider the uncertainty degree when combining classifiers. Obviously, the greater the uncertainty degree, the lower the accuracy of classification and the larger confusion to us to make classification. Based on these findings we define weights with respect to classifiers according to Deng entropy as follows

$$w_i(BPA) = 1 - \frac{E_d(BPA_i)}{\max_i \{E_d(BPA_i)\}}$$

(11)

where $E_d$ is the Deng entropy expression of the BPA, i.e. The weights are different from one classifier to another depending on how much belief degree the BPA has provided from each classifier.

Based on the mass function and its corresponding weight $w_i(BPA)$, we can obtain the discounted mass function before combining them, expressed as follows

$$m_i^w(A) = w_i(BPA) \times m_i^w(A)$$

(12)

$$m_i^w(\theta) = (1 - w_i(BPA)) + w_i(BPA) \times m_i(\theta)$$

(13)

where $\theta$ is the universal set of mass function.

As of now the weighted BPAs have been determined for individual classifiers. Next, we devote to combine all the evidence BPAs originating from each classifier $\psi_i$ on the classification of input $x$, based on the combining rule of D-S theory, to determine an overall mass function for making the final classification decision. The final mass function can be calculated for the expression as follows

$$m_i(BPA) = \bigotimes_{i=1}^R (m_i^w(BPA))$$

(14)

where $\otimes$ is a combination operator.

Until now, we have determined BPAs of individual classifier as well as their weights, respectively. Moreover, we also obtain the weighted BPAs by making use of the combination rule of D-S theory. Next, we describe the core algorithm of this paper as follows.

In the following section we will use Iris dataset [46] to conduct some experiments to illustrate our method and demonstrate its effectiveness as well as the dynamic and adaptive nature for classification applying to the combination of multi-classifiers.

III. EXPERIMENTS AND ANALYSIS

A large amount of methods for determining the BPA have been proposed by researchers with the more and more application in D-S theory. Zhu et al. [47] presented the method using fuzzy membership degrees to obtain the mass function. Within this method, fuzzy c-means (FCM) plays a key role to denote the gray levels as fuzzy
sets. Yager et al. [48] applied the D-S belief structure to the entire class of fuzzy measures, and studied the entropy from the point of fuzzy measure. Bloch et al. [49] associated cluster centers with distance to determine the BPA. Bloch et al. used an unsupervised way to obtain the BPA and considered the ambiguity between pixels in medical image processing making use of fuzzy membership functions. It is vagueness instead of randomness leading to the ambiguity. Le Hegarat-Mascle et al. [50] and Salzenstein et al. [51] used probability density functions (PDFs) to simulate the knowledge derived from all the information source. And then they put forward a subtractive scheme to transform these PDFs into belief degree. Wang et al. [52] got mass functions from common multivariate data spaces systematically. In recent years, Xu et al. [53] proposed a new method to determine basic probability assignment from training data based on normal distribution assumption. Within his method, normality test is performed for the training set firstly, it will be transformed to an equivalent normal space if training set doesn’t meet the normal distribution. And then to construct the models for different attributes. Next, the relationship between the test sample and the normal distribution models will be determined. Finally, the BPA can be calculated on the basis of the intersections of the selected attributes. Comparing with the above-mentioned measures for determining BPA, we consider that the method based on normal distribution is more effective and practical. So in this paper, we will use this method to obtain BPA for each classifier as the preparation for weighted combination of multi-classifiers.

The experiment is based on the Iris data. There are 150 samples of Iris data including 4 attributes for each sample named as Sepal Length (SL), Sepal Width (SW), Petal Length (PL) and Petal Width (PW), respectively. These samples are divided into three classes named as Setosa, Versicolour and Virginica, respectively. There are 50 samples for each of the three classes, and 30 samples are selected randomly as the training set, and the remaining 20 samples regarded as the test set. Each of the four attributes is considered as an information source as well as a classifier, and there are three training sets and three test sets correspondingly. In other words, each attribute is treated as an evidence from a classifier \( \phi_i \). The data can be obtained from the UCI repository of machine learning databases (http://archive.ics.uci.edu/ml/dataset/Iris).

Next step, we will determine the BPAs of each attribute of the Iris data, namely, the mass functions of individual classifiers using the above mentioned normal distribution assumption.

Now an example is given to show the process of the classifier combination for classification. Supposing that the training sets and test sets have been obtained from Iris data using normal distribution method. We then select an instance as test sample from the test set of Virginica. The four attribute values are shown as follows:

\[
SL = 6.3\text{cm}, \quad SW = 2.5\text{cm}, \quad PL = 5.0\text{cm}, \quad PW = 1.9\text{cm}
\]

The BPAs of these attributes (classifiers) are shown in Table 1, and the S, V and C represent class Setosa, Versicolour and Virginica, respectively.

Taking the attribute SL as an example to explain the calculation of the proposed algorithm.

It is obvious the Deng entropy of attribute SL is 1.7220 from Example 2.1. And it can be proved for three elements in frame of discernment, the Deng entropy gets maximal value when the BPA distributes as Example 2.2. So, the \( \max[E_d(BPA)] \) is 4.2479 in this example. Then the weight can be obtained by Eq. (12) as follows

\[
W_{SL}(BPA) = 1 - \frac{1.7220}{4.2479} = 0.5946
\]

Also, the weighted BPA can be calculated by Eqs. (13) and (14).

\[
m(\{C\}) = 0.5946 \times 0.5554 = 0.3303
\]
\[
m(\{C, V\}) = 0.5946 \times 0.4420 = 0.2628
\]
\[
m(\{S, C, V\}) = (1 - 0.5946) \times 0.5946 \times 0.0026 = 0.4069
\]

Next, the relationship between the test sample and the classifier combination for classification. Supposing that the training sets and test sets have been obtained from Iris data using normal distribution method. We then select an instance as test sample from the test set of Virginica. The four attribute values are shown as follows:
The weights and weighted BPAs of the other three attributes are shown in Table 2 and Table 3.

Now we are committed to combine all the four BPAs from individual classifiers \(\psi_i\) on the classification of the test sample by Eq. (18). The result is show as follows

\[
\begin{align*}
    m^w(\{C\}) &= 0.0933, \\
    m^w(\{V\}) &= 0.8356, \\
    m^w(\{C, V\}) &= 0.0599, \\
    m^w(\{S, C, V\}) &= 0.0112
\end{align*}
\]

The process of our experiment is over for this test sample. The combination results illustrate that the belief degree for \(V\) (Virginica) is 0.8314 in the combined BPA, and the effectiveness can be demonstrate from this experiment. Other discounted BPAs of rest test samples, namely, the classification of input \(x\), can be obtained by this process. In order to demonstrate the results of our experiments more visually and effectively, another parts of experiment results are given based on Iris dataset using the proposed method. The results are shown in Figures 2-4.

In Figure 2-4, the \(x\)-coordinate represents 20 test samples of three test sets from Setosa, Versicolour and Virginica, respectively. And the \(y\)-coordinate means the probability values of the class which the test sample belong to in the BPAs of individual classifiers. We can find that the all probability values of Setosa test set are close to 1, and for Versicolour the most values exceed 0.9. There exist a few classifiers out of operation in making classification of Virginica test set, but most of the rest part perform a good job. Suppose that 0.5 is the demarcation point deciding whether the proposed method is effective. Thus, the recognition rate of all pieces of classifiers using our method approaches reaches to 95% approximately in this experiment.

We select a test sample which was worst suitable in classification for class Virginica to analyze the causes of this phenomenon. And we give the four classifiers’ BPAs,
In addition, we conduct 100 times random experiments with the purpose of further explaining the accuracy of the proposed classification method. We list the average classification accuracy rates for Class Setosa, Versicolour, Virginica and the average of the three classes conducting the random experiments 10, 20,...,100 times, respectively. The results are shown in Table 4.

For expressing the results more unambiguous and visualized, we give the results of the average classification accuracy rates varying the random experiments from 1 to 100 in Figure 5.

From Table 4 and Figure 5, we can find that the average classification accuracy rates of the three classes and their average are considerably high. In conclusion, the experiments manifest classification validity of our proposed weighted combination algorithm for multi-classifiers based on D-S theory.

IV. CONCLUSION

In this paper, the basic framework of D-S theory has been constructed for weighted combination of multi-classifiers for classification. A new method has been proposed to define adaptively weights of individual classifier based on Deng entropy which is used to measure the uncertainty degree of BPAs. Then we combine the weighted BPAs derived from individual classifier to obtain the final BPA for the classification decision. It should be pointed out that the proposed method can be seen as the generalization of the method of Huynh et al. [28]. If the output of classifiers are PDs, the proposed method is degenerated as the method of Huynh et al. [28]. From this point of view, the proposed method is more efficient to handle uncertain information. Moreover, our method can determine corresponding BPAs as the output of classifiers have changed, namely the proposed method has good adaptability.

In the experimental section, we determine BPAs of three test sets using normal distribution method based on Iris dataset. Then, the weights are calculated and weighted BPAs are determined making use of our proposed method for each classification. Finally, these weighted BPAs are combined by Dempster’s rule. The experimental results illustrate the effectiveness of our method for classification.

ACKNOWLEDGMENTS

The work is partially supported by National High Technology Research and Development Program of China (863 Program) (Grant No. 2013AA013801), National Natural Science Foundation of China (Grant Nos. 61174022,61573290), China State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (Grant No.BUAA-VR-14KF-02).

REFERENCES

TABLE VII
AVERAGE CLASSIFICATION ACCURACY RATES OF T TIMES

<table>
<thead>
<tr>
<th>Class</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setosa</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Versicolor</td>
<td>94.5</td>
<td>94.25</td>
<td>93.5</td>
<td>93.25</td>
<td>93.2</td>
<td>93.17</td>
<td>93.0</td>
<td>92.94</td>
<td>93.11</td>
<td>93.2</td>
</tr>
<tr>
<td>Virginica</td>
<td>90</td>
<td>90.25</td>
<td>91</td>
<td>90.63</td>
<td>90.3</td>
<td>90.33</td>
<td>90.71</td>
<td>90.75</td>
<td>90.56</td>
<td>90.15</td>
</tr>
<tr>
<td>Average of all the classes</td>
<td>94.83</td>
<td>94.83</td>
<td>94.83</td>
<td>94.63</td>
<td>94.5</td>
<td>94.5</td>
<td>94.57</td>
<td>94.56</td>
<td>94.56</td>
<td>94.45</td>
</tr>
</tbody>
</table>

Fig. 5. Average classification accuracy rates varying random experiments from 1 to 100


