International Journal of Innovation in Computer Science and Information Technology
Vol.1, NO.1 , P:15 - 26
Received: 23 Nov 2018
Accepted: 19 Feb 2019

A new classification model based on Evidence theory

Hamidreza Tahmasbi
Department of Computer Engineering, Kashmar Branch, Islamic Azad University, Kashmar, Iran
htahma@gmail.com

Abstract
Studies have revealed that a combination of classifiers is often more accurate than an individual classifier. A multiple classifier system can take advantage of the strengths of the individual classifiers, avoid their weaknesses, and improve classification accuracy. This system can be considered as an efficient mechanism to achieve the highest possible accuracy in medical classification problem. In this paper, we propose a new method for combination of multiple classifiers using Dempster-Shafer theory of evidence combination for mining medical data. We combine the beliefs of three classifiers: Multi-Layer Perception Neural Network, K-Nearest Neighbor and Naïve Bayesian. Our experiments over the Breast Cancer Wisconsin dataset shows improvement compared to the classification results produced by the individual classifiers and other classifiers which use the combination methods.

Keywords: Medical Data Mining, Multiple Classification, Dempster-Shafer Theory.

Introduction

Many researchers have realized that there exist limitations on using a single classification technique [1, 2]. The combination of multiple classifiers has been intensively studied with the aim of overcoming the limitations of individual classifiers [1-4]. The performance of a multiple classifier system relies on both the complementary participating classifiers and the combination method. Hence, the research efforts in this field have focused on either the generation of complementary classifiers or the combination of a given set of classifiers [3].

A multiple classifier has a combination algorithm that fuses the information coming from the individual classifiers and takes a decision on the new combined information. Some commonly used algorithms for combining classifiers include weighted majority voting [5], Borda count Bayesian [6], Behaviour Knowledge Space (BKS) [7, 8] and Dempster-Shafer's theory [1, 9, 10].

Since the outputs of individual classifiers are inputs to the combination module, it is therefore important to analyze what kinds of output information classifiers can support. The output information that most of the classifiers support can be divided into three levels: abstract level, rank level, and measurement level [11]. The abstract level classifiers output only the class label, and the rank level classifiers output the rank for each class. The measurement level classifiers assign each class a measurement value to indicate the possibility that the input instance pertains to the class. Neural networks are representative examples of measurement level classifiers. The measurement level classifier is able to provide richer information than the abstract and the rank level classifiers.The problems that are encountered in combining classifiers consist of two major

aspects [12]. The first one is closely dependent on specific applications, including how many classifiers to select for a specific application; what types of classifiers to use and what types of feature representations to choose for each classifier. The second aspect is related to issues that are general and common to applications, including the best way to combine classifier outputs in terms of the best combination of classifiers so that precise classification decisions can be achieved [11]. In this work, we focus our interest on how classifier outputs can be modeled as pieces of evidence and then they can be combined by using the Dempster-Shafer theory of evidence for breast cancer diagnostic. The Dempster-Shafer theory of evidence is a powerful method for combining measures of evidence from different classifiers [13].

We propose a new Data Mining Multiple Classification approach that combines the output information of three individual classifiers, that are Multi-Layer Perceptron Neural Network, k Nearest Neighbors and Naive Bayes, based on the Dempster-Shafer theory of evidences, for improving the classification accuracy in the Breast Cancer Wisconsin (BCW) dataset. Classification results produced by the proposed Multiple Classification System shows improvement compared to the classification results produced by the individual classifiers and another tested common classifiers combination methods.

The rest of this paper is organized as follows: Section 2 reviews the Dempster-Shafer theory of evidence. Section 3 discusses the existing methods for computing evidence. The proposed combination technique is presented in section 4. Experimental results and evaluations on BCW dataset are stated in Section 5. This section compares the proposed system with other conventional methods, and an older implementation of the Dempster-Shafer theory. Finally in Section 6, the conclusion and future research are given.

I. The Dempster-Shafer theory of evidence

The Dempster-Shafer (DS) theory of evidence [1, 9, 10] developed by Arthur Dempster, then by Glenn Shafer. It is a powerful tool for representing uncertain knowledge [14], and a combination of two distinct believes is well defined and various methods can be implemented to represent the evidence provided by the individual classifiers. So, it appears as a good tool for multiple classifiers.

Let Θ be a finite set of possible hypotheses. This set is referred to as the frame of discernment, and its powerset denoted by 2^{Θ}. Following are the basic concepts of the theory:

Basic Belief Assignment (BBA). A basic belief assignment m is a function that assigns a value in [0, 1] to every subset A of Θ and satisfies the following:

$$m(\phi) = 0 \text{ and } \sum_{A \subseteq \Theta} m(A) = 1 \quad (1)$$

where ϕ is the empty set. m(A) represents the belief that the searched element is in A.

Belief Function. The belief function, Bel, associated with the BBA m is a function that assigns a value in [0, 1] to every nonempty subset B of Θ. It is called "degree of belief in B" and is defined by

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

Plausibility Function. The quantity $Pl(B) = 1 - Bel(\bar{B})$ called the plausibility of B defines to what extent one fails to doubt in B. i.e., what extent one finds B plausible. It is straight forward to show that:

$$Pl(B) = \sum_{A \cap B \neq \phi} m(A) \quad (3)$$

Combination rule. Consider two BBAs m_1 and m_2 for belief functions Bel_1 and Bel_2 respectively that coming from two different sources. Then m_1 and m_2 can be combined to obtain the belief mass committed to $A \subseteq \Theta$ according to the following combination or orthogonal sum formula,

$$m(A) = (m_1 \oplus m_2)(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - N}, A \neq 0 \quad (4)$$

where N is the conflict factor, $N \in [0,1]$:

$$N = \sum_{B \cap C = \phi} m_1(B) m_2(C) \quad (5)$$

When N = 0, m_1 and m_2 are the same, while

when N = 1 there is total conflict and the combination of m1 and m_2 results in an empty BBA m. The higher N is, the more the combined BBAs are in contradiction, thus N measures the conflict between both evidences represented in BBAs m_1 and m_2.

Combining several belief functions. The combination rule can be easily extended to several belief functions by repeating the rule for new belief functions. Thus the pairwise orthogonal sum of n BBAs, $m_1, m_2, ..., m_n$, can be formed as

$$((ml_1 \oplus m_2) \oplus m_3) ... \oplus m_n = \oplus_{i=1}^{n} m_i \quad (6)$$

II. Existing methods

Let T be the number of classifiers $e_1, \ldots, e_T$ and let C ={$C_1, \ldots, C_M$} be a set of classes with M Classes. For any instance X, each classifier produces an output vector $e_i(X)$, i=1, …, T. The output vector $e_i(X)$ is a vector where dimensions are classes of C and values are the degree of belonging to X to this class. Classifying X means assigning it into one class in C, i.e., deciding among a set of |C| hypotheses: X belongs to C_q , q=1,…,M according to $e_i(X)$. In DS terms, C is referred to as a frame of discernment, and the classifying process is regarded as one which decides the true value the proposition of that instance X belongs to C_q according to the knowledge e(X). e(X) can be regarded as a piece of evidence that represents the degrees of our support or belief for the proposition. Instead of 100% certainty, it only expresses some part of our belief committed to $\{C_q\} \in 2^M$ and the rest of our belief remains unknown or indiscernible which cannot be directly derived from e(X) and the negation of the proposition. In the DS formalism, such a situation is regarded as rejection, and belief functions provide us with an effective way to express it. This is one of the attractive features of DS-based methods. The following of this section review DS-based methods for representing evidence and defining belief functions based on e(X).

The reference [15] proposed a simple method of creating basic probability assignments (BBAs) by using recognition, substitution, and rejection rates ($\varepsilon_r^i, \varepsilon_s^i, 1-\varepsilon_r^i-\varepsilon_s^i$). For a new instance X, a piece of evidence $e_i(X)$ is represented by the following belief function:

$$m_q^i(\{C_q\}) = \varepsilon_r^i(e_i(X)) \quad i = 1,...,T, \exists q \in \{1,...,M\} \quad (7)$$

$$m_q^i(\{\overline{C}_q\}) = \varepsilon_s^i(e_i(X)) \quad i = 1,...,T, \exists q \in \{1,...,M\} \quad (8)$$

$$m_q^i(C) = 1 - m_q^i(\{C_q\}) - m_q^i(\{\overline{C}_q\}) \quad (9)$$

Where $\{\overline{C}_q\} = C - \{C_q\}$.

With T pieces of evidence existing, represented by T belief functions, the degrees of support for classes can be calculated through combining these belief functions by formula (4). A final class decision for a given instance is made on selecting the class with the largest degree of support. The drawback of this method is the way evidence is measured. This method ignores the fact that normally a classifier does not have the same performance on different classes, this might consequently degrade the combined performance of classifiers[13] .

Rogova [16], proposed a model for combining the results of neural network classifiers using the DS theory. He used several proximity measures between a reference vector and a classifier's output vector. The proximity measure that gives the highest classification accuracy was later transformed into evidences. The reference vector used was the mean vector, μ_q^i of the output set of each classifier e_i and each class label q. A number of proximity measures, d_q^i for μ_q^i and y_i were considered. y_i is the output of classifier e_i. For each classifier e_i, the proximity measure of each class C_q is transformed into the following BBAs:

$$m_q^i(\{C_q\}) = d_q^i \quad , m_q^i(C) = 1 - d_q^i \quad (10)$$

$$m_{j(j \neq q)}^i(\{\overline{C}_q\}) = 1 - \prod_{r=1, r \neq q} (1 - d_r^i) \quad (11)$$

$$m_{j(j \neq q)}^i(C) = 1 - m_j^i(\{\overline{C}_q\}) = \prod_{r=1, r \neq q} (1 - d_r^i) \quad (12)$$

Finally, Dempster's combination rule was used to combine evidences for all classifiers to obtain a measure of confidence for each class label. The major drawback of Rogova's method is the way the reference vectors are calculated, where the mean of output vectors may not be the best choice [13]. Al-Ani and Derichi [13] proposed a similar method to apply the DS theory of evidence to combining neural network outputs. This method also treated the distance between a reference vector and a classifier output as a piece of evidence. But the difference from the previous work is in the way it obtains reference vectors. It first initializes reference vectors for each class, and then iteratively uses training instances to optimize reference vectors through minimizing the mean square errors between combined classifier outputs and the target outputs, ensuring the optimized reference vectors can be achieved.

Finally, the distance between the optimized reference vectors and classifier outputs is defined as a piece of evidence and is represented by a simple support function. Let μ_q^i be an optimized reference vector. For any instance X, each classifier produces an output vector $e_i(X)$, i=1, …, T, a simple support function is defined below:

$$m_q^i(\{C_q\}) = \frac{d_q^i}{\sum_{j=1}^{M} d_j^i + g^i} \quad (13)$$

$$m_q^i(C_q) = \frac{g^i}{\sum_{j=1}^{M} d_j^i + g^i} \quad (14)$$

Where $d_q^i = \exp(-||\mu_q^i - e_i(X)||^2)$ and g^i is a coefficient to be tuned.

In this method the way of obtaining reference vectors through minimizing the overall mean square error makes the process of combining classifiers trainable, which may lead to better performance than Rogova's Method, but with the additional cost for training as well as additional training data [10].

Unlike the methods above which were designed for combining classifiers in ensemble learning, Denoeux [14] proposed an evidence theoretic k-nearest neighbours (kNN) method for classification problems based on the DS theory. this method focuses on a single classifier in classifying new instances, by accounting for distances from their neighbors to determine class labels.

Let D be a training data set, for instance, $d \in D$ and let Φ be a set of the k-nearest neighbors of d according to some distance measures (e.g. Euclidian distance).

Classifying d means assigning it to one of the classes $C_q \in C$ based on the weights of representative classes of its neighbors. Thus the distance between d and neighbor $d_i \in \Phi$ considered as a piece of evidence to support a proposition about the class membership of d. The evidence is represented by a simple support function as follows [14]:

$$m^i(\{C_q\}) = \varphi(d, d_i), \quad d_i \in \Phi \quad (15)$$

$$m^i(C) = 1 - \varphi(d, d_i) \quad (16)$$

$$m^i(A) = 0, \quad \forall A \in 2^C \setminus \{\{C_q\}, C\} \quad (17)$$

where φ was suggested to be $\exp(-\gamma(\mu^i)^2)$ and $\mu^i = ||d - d_i||$.

Denoeux's method shows the advantage of permitting a clear distinction between the presence of conflicting information, as happens when an instance is close to several neighbors from different classes and incomplete information, when an instance is far away from any instances in its neighborhood. It proves to be very competitive with the standard kNN methods. Denoeux adapted a similar idea to a neural network classifier (ANN) [17].

Fadi Elmasri [18] used the kNN and ANN methods that proposed by Denoeux, and proposed a new Classification approach that combine the output information of three individual classifiers, that are Neural network, k Nearest Neighbors and Naive Bayes, based on the Dempster-Shafer theory of evidences, and showed that his idea improve the classification accuracy.

III. The Proposed multiple classification system

In this section, we propose a new Data Mining Multiple Classification approach that combines the output information of three

individual classifiers i.e. Multi-Layer Perceptron Neural network (MLP), k Nearest Neighbors (kNN) and Naive Bayes (NB), based on the Dempster-Shafer theory of evidences, for improving the classification accuracy. Figuse (1) shows the structure of the proposed Multiple Classification system.

Let C = {C_1, …, C_M} be a set of classes where M is the number of classes.

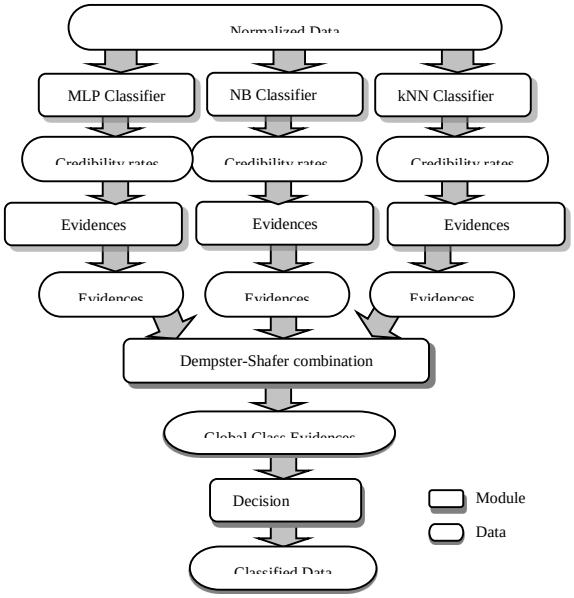

Figure (1) The structure of the proposed Multiple Classification system

A. Proposed Multiple Classification Data Flow Steps

A.

Following is the description of the data flow through the proposed Multiple Classification system:

- In the first step, the normalized data forms the input of each individual classifier disjointedly.
- Each Classifier independently processes the normalized data and produces its classification results in the form of measurement values for each data sample. We consider this values as a credibility rate.
- The credibility rates produced by the different classifiers are simultaneously fed as an input to the evidences extraction module.
- The extracted evidences from the evidences extraction module are fed as an input to the evidences combination module. The evidences combination module performs an evidence combination process using the Dempster-Shafer combination rule to produce the global class evidences for each data sample.
- The global class evidences cross the threshold in the decision-making module to either assign the data sample to one class or reject it.
- The final output results are either obtained as an abstract class label C_j (j=1, …, M) or as a rejection indication M+1.

B. Structure of The Proposed Multiple Classification System

The proposed multiple classification system consists of several, different purpose, connected modules. The modules are connected to emulate the data flow scheme mentioned in the previous section. The proposed multiple classification system consists of the following modules:

a. kNN Module

For Extraction the evidences from outputs of the classifiers, these classifiers must produce their output in the form of measurement level. In the kNN Classifier, each credibility rate could be equal to $\frac{n_q}{k}$. n_q is the number of objects belonging to class C_q among the k nearest neighbors. However, such a method does not take into account the significant information that are distances $d_q = d(X, X_q)$, q = 1, ..., k associated to the k nearest neighbors $X_1, X_2, \ldots, X_k$. We use the k-nearest neighbours method that proposed by Denoeux that produce output in the form of measurement level. This output can be in form credibility (belief) or plausibility.

If m is the BBA function, the credibility produced for each class C_q is:

$$Bel(C_q) = m(C_q) \qquad q = 1,..., M+1$$

(18)

and the plausibility produced for each class C_q is:

$$Pl(\{C_q\}) = m(\{C_q\}) + m(C) \qquad q = 1,\ldots,M$$

(19)

The $Bel(C_q)$ function is considered when the option of having rejected data samples is applicable. $Pl(C_q)$ function is considered when there is no rejection option.

In this module, we use the $Pl(C_q)$ function and we decide about rejection option in the decision module. We denote the $Pl(C_q)$ with Cr(q) that is credibility rate and normalized in [0,1]. Thus the output of this module is:

$$C_{KNN} = \{Pl(C_1),\ldots,Pl(C_M)\} = \{Cr(1),\ldots,Cr(M)\}$$

(20)

b. MLP Module

In the Multi-Layer Perceptron Neural Network (MLP) module the network used is a Multi-Layer Perceptron (MLP). The outputs of Network are $O_1, O_2, ..., O_M$ that are normalized so that their sum is 1. these values considered as credibility rates.

The probability related to each class is then:

$$P(C_q) = O_i$$

(21)

We denote the $P(C_q)$ by $Cr(_q)$ that is credibility rate. The output of this module is then:

$$C_{MLP} = \{O_1,\ldots,O_M\} = \{Cr(1),\ldots,Cr(M)\}$$

(22)

c. NB Module

We show each output $P(C_q \mid X)$ of the Naive Bayes classifier (NB) in the form Cr(q) credibility rate. Thus the output of the NB module is:

$$C_{NB} = \{P(C_1 \mid X),\ldots,P(C_M \mid X)\} = \{Cr(1),\ldots,Cr(M)\}$$

(23)

d. Evidences Extraction Module

For extraction the evidences or BBAs, we use the Evidences Extraction Module. In this module for a classifier e(X), the classes in which consecutive variations between credibility rates Cr(q), q=1,…, M, are smaller than threshold $\lambda_t \in [0,1]$ form a proposition. In other words, this module groups the classes with close credibility rates into the same proposition, according to the credibility rate imprecision threshold λ_t. To determine λ_t conveniently for a given simple classifier, a series of tests with different training sets should be performed to calculate the average standard deviation of the credibility rates.

The propositions and then the BBAs can be made with the following algorithm:

Algorithm 1

1. Order all classes i by decreasing order of Cr(i) in a vector V, where V = [V(1), V(2), ..., V(M)] and where V(1) corresponds to the class with the highest credibility rate.
2. Let two indexes a and j initialized to a = 1 and j = 1.
3. Create a new empty proposition A_j.
4. Put V(a) in A_j.
5. Increment index a of 1.
6. If a = M + 1, go to step 8.
7. If Cr(V(a - 1)) - Cr(V(a)) < λ_t , return to step 4, if not, increment j by 1 and return to step 3.
8. Determine the m(A_j) that is associated BBA for each proposition A_j by the average credibility rate μ_{A_j} of all classes included in the proposition :

$$m(A_j) = \frac{\mu_{A_j}}{S} \tag{24}$$

with

$$\mu_{A_j} = \frac{\sum_{i \in A_j} Cr(i)}{|A_j|} \tag{25}$$

where S is a normalization constant:

$$S = \sum_{j \in A} \mu_{A_j} \tag{26}$$

9. End

E .Evidence Combination Module

This module is designed to combine the class evidences produced by different evidences extraction modules, using Dempster-Shafer combination rule. Thus the combination of the BBAs for the kNN classifier, the BBAs for the neural network and the BBAs for the Bayes classifier is made following combination rule:

$$m = m_{kNN} \oplus m_{MLP} \oplus m_{NB} \tag{27}$$

A. Decision-making Module

In this module the global evidences obtained from evidences combination module are examined to make the final decision about the tested data sample class label. A final class decision for a given data sample X is made on the following conditions:

1. If $N < \lambda_{rej}$, the class C_J (J=1,…M) is chosen to represent the class label of the data sample X, if

$$P(J) = \max\{P(1), P(2), \ldots, P(M)\} \tag{28}$$

where P(J) is the pignistic probability function with

$$P(J) = \sum_{J \in A} \frac{m(A)}{|A|} \tag{29}$$

2. If at least two classes C_J and C_I share maximum pignistic probability, that is

$$P(I) = P(J) = \max\{P(1), \ldots, P(M)\} \tag{30}$$

Then, the data sample X is rejected.

In such a situation the proposed multiple classification system returns the answer M + 1.

3. Else if $N \geq \lambda_{rej}$, then the data sample X is rejected and returned the answer M + 1. In other words. the decision is rejected if the conflict factor N exceeds a pre-established threshold λ_{rej}. N is a result of the combination.

If λ_{rej} is 1, all decisions are accepted. As λ_{rej} lowers, the rejection rate increase. If λ_{rej} is 0, all decisions are rejected.

IV. Experimental evaluation

A. Experimental settings

The submitting author is responsible for obtaining the agreement of all coauthors and any consent required from sponsors before submitting a paper. It is the obligation of the authors to cite relevant prior work. In our experiments, we used medical data to test and validate the accuracy of the proposed

multiple classification system. Breast Cancer Wisconsin (BCW) dataset is the data set used to test the accuracy of our proposed system downloaded from the UCI machine learning repository [19].

The proposed multiple classification system tries to improve the breast cancer diagnosis for new patient records. However, the BCW dataset forms a two-category classification problem regarding to the patient's diagnosis results (Benign - Malignant). The objective is to identify each record in the dataset as a benign or a malignant record.

We first test the three individual classifiers, that are Multi-Layer Perceptron Neural Network (MLP), k Nearest Neighbors (kNN) and Naive Bayes (NB) on this dataset and then the following combination methods were tested: the weighted linear combination (WLC), average (Av), median (Md), maximum (Mx), majority voting (MV), Elmasri's DS method (DS0) [18], and our proposed classifiers combination method. Throughout the experiments, the validation method is 10-fold cross-validation [20]. The individual classifiers, MLP and NB, are taken from the Waikato Environment for Knowledge Analysis (Weka) version 3.4 [21]. Parameters used for these two classifiers were set at the default settings in Weka.

B. BCW Dataset

BCW dataset has 699 records; each record consists of 11 features; record ID, diagnosis (2 for benign, 4 for malignant) and 9 integer-value features between {1, …, 10}. The classes' distribution for the BCW is 458 benign (65.5%) and 241 malignant (34.5%) records and there are 16 records that contain a single missing (i.e., unavailable) feature value, now denoted by "?". we first replace this missing values with mean of values for this feature in other records. The experiments are performed using 90% of the dataset as training set (629 records) and 10% of the dataset as testing set (70 records).

C. Experimental results

The proposed kNN classifier evaluated for different values of k in 10 runs. The mean of the accuracy of 10 runs for each value of k is displayed in Table 1. It was observed that k=9 gives the best accuracy in most cases. Thus the number of k nearest neighbors is set to 9.

Table 1. Mean of the accuracy of kNN with variation k in 10 runs

K	1	3	5	7	9	11	13
Accuracy mean	0	86.71	89.57	89.43	93.14	92.29	91.71

To determine λ_t and λ_{rej} in the proposed multiple classification system, values 0, 0.2, 0.4, 0.6, 0.8 and 1 selected for λ_t, then for each of this values, values 0.2, 0.4, 0.6 , 0.8 and 1 selected for λ_{rej}, and then for each selected values for λ_t and λ_{rej}, the proposed classifiers combination method tested for 10 Runs. The highest recognition rate mean obtained by 98.43% with $\lambda_t \neq 1$ and $\lambda_{rej} = 1$ $\lambda_t = 0.4$. Thus we selected and for $\lambda_{rej} = 1$ our proposed classifiers combination method.

The confusion matrices for the classification results obtained from the three used individual classifiers and the proposed classifiers combination method in the first running are displayed in Table 2. Rows in the confusion matrix represent the actual class labels of the records to be classified, and the column with the label U represents the number of the rejected records (Uncertainty) of each classifier. B and M denote Benign and Malignant respectively. The confusion matrix of the classification results obtained using the proposed classifiers combination method shows an improvement in the classification accuracy up to 95.71%.

The comparison between the classification accuracy obtained from 10 Runs for each one of the individual classifiers and for the proposed classifiers combination method are displayed in Table 3. This table shows that the proposed combination method produces classification accuracy that is higher than the classification accuracy produced by any individual classifier in Run 1, 2, 3, 4, 5, 6, 7, 9, and 10. Also, it produces at least the same classification accuracy of the best individual classifier in Run 8. Moreover, the proposed combination method mean classification accuracy of 98.43% is higher than the best classifier mean classification accuracy of 94.29%.

This experiment clearly shows that the classification accuracy is improved by 4.14% using the proposed combination method.

The confusion matrices for the classification results obtained from 7 different classifiers combination methods including the proposed classifiers combination method in the first running are displayed in Table 4. These matrices demonstrate the distribution of correct classified, misclassified and rejected records of the different tested combination methods. From the seven confusion matrices listed in Table 4 for the proposed classifiers combination method and the other 6 different combination methods, it is evident that the proposed classifiers combination method has the highest classification accuracy of 95.71%.

The classification accuracy and the mean classification accuracy produced by 10 Runs of the 7 different classifier combination methods including our proposed classifiers combination method are displayed in Table 5.

Table 5 illustrates the advantage of the classification accuracy results obtained from the proposed combination method over the classification accuracy results obtained from each one of the tested combination method. This advantage is confirmed based on the quality and consistency of the classification accuracy results of the proposed combination method in each Run's classification results and the mean classification accuracy results of the 10 Runs. In this experiment, the proposed combination method produces classification accuracy that is higher than the classification accuracy produced by any tested combination method in Runs 1, 2, 3, 4, 5, 6 and 10. Also, it produces at least the same classification accuracy of the best other 6 different combination methods in Run's 7 and 9. However, our proposed method does not produce the highest classification accuracy in the Runs 8 but is produced relatively acceptable and consistent classification results accuracy. Also, the mean classification accuracy of the proposed combination method of 98.43% is higher than the best combination method mean classification accuracy of 97% produced by Elmasri's DS method (DS0). This experiment clearly shows the classification accuracy superiority of the proposed combination method of 1.43% at least over any combination method tested in this experiment.

Table 2. Confusion matrices for the individual classifiers and proposed method in run No.1

kNN				**MLP**				**NB**				**Proposed Method**			
	B	M	U		B	M	U		B	M	U		B	M	U
B	33	2	0	B	33	2	0	B	33	2	0	B	35	0	0
M	5	29	1	M	4	31	0	M	6	29	0	M	2	32	1
Accuracy: 88.57 %				Accuracy: 91.43 %				Accuracy: 88.57 %				Accuracy: 95.71 %			

Table 3. Classification results of 10 Runs for the individual classifiers

Run No.	KNN	MLP	NB	Proposed Method
1	88.57%	91.4%	88.57%	**95.71%**
2	92.86%	88.57%	94.29%	**100%**
3	94.29%	97.14%	94.29%	**98.57%**
4	91.43%	92.86%	90%	**95.71%**
5	94.29%	97.14%	94.29%	**98.57%**
6	97.14%	95.71%	97.14%	**100%**
7	95.71%	92.86%	97.14%	**97.14%**
8	97.14%	**98.57%**	97.14%	**98.57%**
9	98.57%	98.57%	91.43%	**100%**
10	92.86%	88.57%	94.29%	**100%**
Mean	94.29%	94.14%	93.86%	**98.43%**

Table 4. Confusion matrices for 7 different classifiers combination methods in run No.1

MV				**MX**				**AV**				**MD**			
	B	M	U		B	M	U		B	M	U		B	M	U
B	33	2	0	B	33	2	0	B	33	2	0	B	33	2	0
M	4	31	0	M	4	31	0	M	4	31	0	M	4	31	0
Accuracy: 91.43 %				Accuracy: 91.43 %				Accuracy: 91.43%				Accuracy: 91.43 %			

WLC				**DS0**				**Proposed Method**			
	B	M	U		B	M	U		B	M	U
B	33	2	0	B	34	1	0	B	35	0	0
M	3	32	0	M	2	32	1	M	2	32	1
Accuracy: 92.86 %				Accuracy: 94.29 %				Accuracy: 95.71 %			

Table 5. Classification results for 10 Run of seven classifiers combination method

Run No.	MV	MX	AV	MD	WLC	DS0	Proposed Method
1	91.43%	91.43%	91.43%	91.43%	92.86%	94.29%	**95.71%**
2	95.71%	95.71%	95.71%	95.71%	95.71%	97.14%	**100%**
3	97.14%	97.14%	97.14%	97.14%	97.14%	97.14%	**98.57%**
4	94.29%	95.71%	94.29%	94.29%	94.29%	92.86%	**95.71%**
5	97.14%	95.71%	97.14%	97.14%	97.14%	97.14%	**98.57%**
6	98.57%	97.14%	98.57%	98.57%	98.57%	97.14%	**100%**
7	**97.14%**	**97.14%**	**97.14%**	**97.14%**	**97.14%**	**97.14%**	**97.14%**
8	98.57%	98.57%	98.57%	98.57%	98.57%	**100%**	**98.57%**
9	**100%**	98.57%	**100%**	**100%**	**100%**	**100%**	**100%**
10	97.14%	98.57%	97.14%	97.14%	97.14%	97.14%	**100%**
Mean	96.71%	96.57%	96.71%	96.71%	96.86%	97%	**98.43%**

I. Conclusion

The proposed multiple classification approach has been implemented using the Dempster-Shafer theory of evidence as a data fusion tool to combine the classification evidences produced by the different individual classifiers. The proposed system provides at least the same or a higher classification accuracy result than the classification accuracy produced by the most excellent individual classifiers before they are combined using the proposed system. Also, the proposed system produces classification accuracy result that is higher than the classification accuracy results produced by the popular combination systems tested in this work.

For further research, the method proposed here can be used to additional classification problem domains with a different dimensionality, the number of classes and attributes ranges. The proposed system can be tested to perform the combination process on other different predictive or descriptive data mining tasks such as clustering, regression, and prediction. Also can be considered further extension of our approach to combine other different level classifiers such as rank level.

References

[1] Z.-G. Liu, Q. Pan, J. Dezert, and A. Martin, "Combination of classifiers with optimal weight based on evidential reasoning," IEEE Transactions on Fuzzy Systems, vol. 26, no. 3, pp. 1217–1230, 2018.

[2] M. Woźniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," Information Fusion, vol. 16, no. 1, pp. 3–17, 2014.

[3] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," Information Fusion, vol. 46, pp. 147–170, 2019.

[4] S. Karakatič and V. Podgorelec, "Improved classification with allocation method and multiple classifiers," Information Fusion, vol. 31, pp. 26–42, 2016.

[5] S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann, "Review of classifier combination methods," in Machine learning in document analysis and recognition, Springer, 2008, pp. 361–386.

[6] R. Polikar, "Ensemble based systems in decision making," IEEE Circuits and systems magazine, vol. 6, no. 3, pp. 21–45, 2006.

[7] Š. Raudys and F. Roli, "The behavior knowledge space fusion method: Analysis of generalization error and strategies for performance improvement," in International Workshop on Multiple Classifier Systems, 2003, pp. 55–64.

[8] A. Ferreira et al., "Behavior knowledge space-based fusion for copy--move forgery detection," IEEE Transactions On Image Processing, vol. 25, no. 10, pp. 4729–4742, 2016.

[9] K. Sentz and S. Ferson, Combination of evidence in Dempster-Shafer theory, vol. 4015. Citeseer, 2002.

[10] Y. Bi, J. Guan, and D. Bell, "The combination of multiple classifiers using an evidential reasoning approach," Artificial Intelligence, vol. 172, no. 15, pp. 1731–1751, 2008.

[11] Y. Bi, D. Bell, H. Wang, G. Guo, and J. Guan, "Combining multiple classifiers using dempster's rule for text categorization," Applied Artificial Intelligence, vol. 21, no. 3, pp. 211–239, 2007.

[12] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 3, pp. 226–239, 1998.

[13] A. Al-Ani and M. Deriche, "A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence," Journal of Artificial Intelligence Research, vol. 17, pp. 333–361, 2002.

[14] T. Denœux, “A k-nearest neighbor classification rule based on,” in Classic works of the Dempster-Shafer theory of belief functions, vol. 25, no. 05, Springer, 1995, pp. 737–760.

[15] L. Xu, A. Krzyzak, and C. Y. Suen, “Methods of combining multiple classifiers and their applications to handwriting recognition,” IEEE transactions on systems, man, and cybernetics, vol. 22, no. 3, pp. 418–435, 1992.

[16] G. Rogova, “Combining the results of several neural network classifiers,” in Classic Works of the Dempster-Shafer Theory of Belief Functions, Springer, 2008, pp. 683–692.

[17] T. Denoeux, “A neural network classifier based on Dempster-Shafer theory,” IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 30, no. 2, pp. 131–150, 2000.

[18] F. Elmasri, “Multi-classifiers approach using Dempster-Shafer theory for data mining.,” 2006.

[19] “UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).

[20] G. Jiang and W. Wang, “Error estimation based on variance analysis of k-fold cross-validation,” Pattern Recognition, vol. 69, pp. 94–106, 2017.

[21] “Weka 3 - Data Mining with Open Source Machine Learning Software in Java.” [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/.