On Combining Multiple Classifiers Using an Evidential Approach

Yaxin Bi
School of Computing and Mathematics
University of Ulster at Jordanstown
Co. Antrim, BT37 0QB, UK
y.bi@ulster.ac.uk

Sally McClean
School of Comp and Info Engineering
University of Ulster
Co. Londonderry, BT52 1SA, UK
si.mcclean@ulster.ac.uk

Terry Anderson
School of Computing and Mathematics
University of Ulster at Jordanstown
Co. Antrim, BT37 0QB, UK
tj.anderson@ulster.ac.uk

Abstract

Combining multiple classifiers via combining schemes or meta-learners has led to substantial improvements in many classification problems. One of the challenging tasks is to choose appropriate combining schemes and classifiers involved in an ensemble of classifiers. In this paper we propose a novel evidential approach to combining decisions given by multiple classifiers. We develop a novel evidence structure – a focal triplet, examine its theoretical properties and establish computational formulations for representing classifier outputs as pieces of evidence to be combined. The evaluations on the effectiveness of the established formalism have been carried out over the data sets of 20-newsgroup and *Reuters*-21578, demonstrating the advantage of this novel approach in combining classifiers.

Introduction

Ensemble research has shown the choice or design of the method for combining classifier decisions is one of the challenging tasks in constructing an effective ensemble (Dietterich, 1997) and various approaches have been developed in the past decade. The combination methods can roughly be characterized, based on the forms of classifier outputs, into two categories (Kuncheva, 2001). In the first category, the combination of decisions is performed on single class labels, such as majority voting and Bayesian probability, which have extensively been examined in the ensemble literature (Xu, *et al*. 1992; Kittler, *et al*. 1998; Sebastiani, 2002). The second category is concerned with the utilization of continuous values corresponding to class labels. One typical method is based on the same class labels from different classifiers in calculating the support for class labels, regardless of what the support for the other classes is. We refer to this method as the *class-aligned* method. Another is to use continuous values of class labels as a set of features to learn a combining function in addition to the ensemble of classifiers in terms of *meta-learning*. An alternative group of methods is to make use of as much information as possible obtained from single and sets of classes in calculating the support for each class, called *class-indifferent* methods (Kuncheva, 2001).

Considerable effort has been devoted to studying the combination of decisions in the form of continuous values of class-aligned labels, such as linear sum and order statistics (Yang, *et al*. 2000; Tumer, *et al*. 2002), and meta-learning, such as stacking (Dzeroski, *et al*. 2004). In contrast several works related to class-indifferent methods utilizes single classes and sets of classes (Xu, *et al*. 1992; Denoeux, 2000). However class-indifferent methods for combining decisions in the form of lists of ordered decisions have not been so intensively studied and therefore are poorly understood. In particular, little is known about evidential reasoning methods for combining truncated lists of ordered decisions in the context of text categorization.

Concentrating on combining lists of ordered decisions made by a classifier committee, we propose a novel evidential approach to modeling the process of combining multiple classifiers and examine the effect of different sizes of decision list on both efficiency and accuracy. We do so by modeling each output given by classifiers on new instances as a list of decisions that are quantitatively prioritized by continuous values and each list of decisions is ranked and partitioned to subsets of 2, 3 and 4 decisions which are represented by novel evidential structures in terms of *triplets*, *quartets* and *quintets* (Bi, *et al*. 2004). Resulting triplets or quartets or quintets are combined by using Demspter's rule to constrain the final decision.

The advantages of our approach are summarized as follows. The first is that our method makes use of a wider range of pieces of evidence in classification to make the final decision. The idea is inspired by the observation that if only 'best' single class labels are selected on the basis of their values, valuable information contained in the discarded labels may be lost. Arguably, the potential loss of support derived from the other classes could be avoided if such support information is utilized in the decision making process. The evidence structures, such as triplet, are able to incorporate the best-supported class, the second best-supported class, and undeterministical status in terms of ignorance into the process of decision making. The second is that these evidence structures provide an efficient way for combining more pieces of evidence since they break down a large list of decisions into smaller and tractable subsets. Like the dichotomous structure (Barnett, 1981), our method deals well with a long-standing criticism saying that the evidence theory does not translate easily into practical applications due to the computational complexity of combining multiple pieces of evidence. In fact, our method generalizes the idea of the dichotomous structure and has the advantage over this structure in *distinguishing trivial focal elements from important ones* and in representing uncertainty associated with lists of decisions.

Representation of Classifier Outputs

In supervised machine learning, a learning algorithm is provided with training instances of the form $D \times C = \{\langle d_1, c_1 \rangle, \langle d_2, c_2 \rangle, \ldots, \langle d_{|D|}, c_q \rangle\}$ $(1 \leq q \leq |C|)$ for inducing some unknown function f such that $c = f(d)$. The d_i values are typically vectors of the form $(w_{i1}, w_{i2}, \ldots, w_{in})$ whose components are symbolic or numeric values, and the c_i values are typically drawn from a set of categorical classes in terms of class label. Given a set of training data, a learning algorithm is aimed at learning a function φ in terms of classifier from the training data, where classifier φ is an approximation to an unknown function f.

Given a new instance d, a classification task is to make the decision for d using φ about whether instance d belongs to class c_i. Instead of single-class assignment, we denote such a process as a mapping:

$$\varphi : D \to C \times [0,1] \tag{1}$$

where $C \times [0,1] = \{(c_i, s_i) \mid c_i \in C, 0 \leq s_i \leq 1\}$, s_i can be in different forms, such as similarity scores or posterior probabilities, depending on the types of learning algorithms. It represents the support or degree of confidence about the proposition that instance d is assigned to class c_i.

Given an ensemble of classifiers, $\varphi_1, \varphi_2, \ldots, \varphi_M$, each classifier output on instance d can be represented by a distinct score vector (list) as illustrated in formula (2).

$$\varphi_i(d) = \{s_{ij} \mid 1 \leq j \leq |C|\} \text{ where } 1 \leq i \leq M \tag{2}$$

Based on the above formula, the combination of decisions can be carried out in different ways. The *class-aligned* method calculates the support for class c_j using only $s_{1j}, s_{2j}, \ldots, s_{Mj}$, such as *minimum*, *maximum*, *average* and *sum*, etc. regardless of what the support from the other classes is. The *class-indifferent* methods, as an alternative group of methods, make use of all the vectors $\varphi_1(d), \varphi_2(d), \ldots, \varphi_M(d)$ or the selected information to calculate the support for c_j, which is used to constrain the final decision. In this paper, we propose an evidential reasoning method that is similar to a class-indifferent method, in which it truncates each score vector and restructures them into the novel evidence structures of triplet, quartet or quintet.

Preliminaries

The Dempster-Shafer (DS) theory of evidence has been recognized as an effective method for coping with uncertainty or imprecision in reasoning processes. It is often viewed as a generalization of the Bayesian probability theory, by providing a coherent representation for *ignorance* (lack of evidence) and also by discarding *the insufficient reasoning principle*. We briefly present the DS framework as follows (Shafer, 1976).

Definition 1 Let Θ be a finite nonempty set, called the *frame of discernment.* Let [0,1] be an interval of numeric values. A mapping function $m: 2^\Theta \to [0,1]$ is defined as a *mass function* if it satisfies:

1) $m(\phi)=0$
2) $\sum_{H \subseteq \Theta} m(H) = 1$ (3)

A mass function is a *basic probability assignment* (*bpa*) to all subsets X of Θ. A subset $A \subseteq \Theta$ is called a *focal element* of a mass function m over Θ if $m(A)>0$ and A is called a singleton if it is a one-element subset. Given the general representation of classifier outputs in formula (2), we define an application-specific mass function below.

Definition 2 Let C be a *frame of discernment*, where each choice $c_i \in C$ is a proposition that instance d is classified in category c_i. Let $\varphi(d) = \{s_1, s_2, \ldots, s_{|C|}\}$ be a list of scores, an application-specific mass function is defined a mapping, m: $2^C \to [0,1]$, i.e. a *bpa* to $c_i \in C$ for $1 \leq i \leq |C|$ as follows:

$$m(\{c_i\}) = s_i \Big/ \sum_{j=1}^{|C|} s_j \text{, where } 1 \leq i \leq |C| \tag{4}$$

This mass function expresses the degrees of belief with regard to the choices of classes to which a given instance could belong. By equation (4), we can rewrite $\varphi(d)$ as $\varphi(d) = \{m(\{c_1\}), m(\{c_2\}), \ldots, m(\{c_{|C|}\})\}$, referred to as a list of decisions – a piece of evidence.

Definition 3 Let m_1 and m_2 be two mass functions on the frame of discernment C, and for any subset $A \subseteq C$, the *orthogonal sum* $\oplus$ of two mass functions on A is defined as:

$$(m_1 \oplus m_2)(A) = \sum_{X \cap Y = A} m_1(X) * m_2(Y) \Big/ N \tag{5}$$

where $N = 1 - \sum_{X \cap Y = \phi} m_1(X) * m_2(Y)$ and $K = 1/N$ is called the normalization constant of the orthogonal sum $m_1 \oplus m_2$. The orthogonal sum is often called Dempster's rule of combination. There are two conditions to ensure the orthogonal sum exists: 1) $N \neq 0$ – if $N = 0$, then two mass functions m_1 and m_2 are totally contradictory; 2) two mass functions must be independent of each other – represent independent opinions relative to the same frame of discernment.

Partitioning a List of Decisions – Triplet

Starting by analyzing the computational complexity of combining multiple pieces of evidence, we consider how a more efficient method for combining evidence can be established. Given M pieces of evidence represented by formula (2), the computational complexity of combining these pieces of evidence using Equation (5) is dominated by the number of elements in C and the number of classifiers M. In the worst case, the time complexity of combining M pieces of evidence is $O(|C|^{M-1})$. One way of reducing the computational complexity is to reduce the pieces of evidence being combined so that the combination of evidence is carried by a partition ϑ of the frame of discernment *C*, where ϑ has less focal elements than *C* and includes possible answers to the proposition of interest. The partition ϑ can thus be used in place of *C* when the computations of the orthogonal sum are carried out (Shafer, 1976). For example, Barnett (1981) proposed a dichotomous structure which can partition the frame of discernment *C* into two subsets ϑ_1 and ϑ_2, where there are a number of mass functions that represent evidence in favor of ϑ_1 and against ϑ_2, along with the lack evidence – ignorance. It has been shown that Dempster's rule can be implemented such that the number of computations increases only linearly with the number of elements in *C if* the mass functions being combined are focused on the subsets where ϑ_1 is singleton and ϑ_2 is the complement of ϑ_1, i.e. $O(|C|)$.

The partitioning technique is powerful because it enables to partition a large problem space into a number of smaller and more tractable problems. However a fundamental issue in applying this technique is how to select elements that contain the possibly correct answers to the propositions corresponding to *C*.

An intuitive way is to select the element with the highest strength of confidence. Indeed, since the classifier outputs approximate class posteriori probabilities, selecting the maximum probability reduces to selecting the output that is the most 'certain' of the decisions. This could be justified from two perspectives. First, on the one hand, the probability assignments given in formula (2) represent quantitatively judgments made by classifiers on the propositions, where the greater their values, the more likely these decisions are correct. Thus selecting the maximum *distinguishes the trivial from the important ones.* Second, on the other hand, the combination of decisions with the lower degrees of confidence may not contribute to the increase of combined performance of the classifiers, but only make the combination of decisions more complicate (Tumer, *et al.* 2002). The drawback of selecting the maximum, however, is that the combined performance can be dropped down by a single classifier that repeatedly provides high confidence values because the decisions with the higher values are always chosen as the final classification decisions, but some of them may not always be correct.

To cope with the deficiency resulting from a single classifier, we propose to take the second maximum decision into account in combining classifiers. Its inclusion not only provides valuable information contained in the discarded class labels by the maximal selection for combining classifiers, but this also avoids the deterioration of the combined performance to some extent, which is caused by the errors resulting from a single classifier that repeatedly produces high confidence values. The following details a novel structure – a *triplet* – partitioning a list of decisions $\varphi(d)$ into three subsets.

Definition 4 Let *C* be a frame of discernment and $\varphi(d) = \{m(\{c_1\}), m(\{c_2\}), \ldots, m(\{c_{|C|}\})\}$, where $|\varphi(d)| \geq 2$, a *triplet* is defined as an expression of the form $Y = \langle A_1, A_2, A_3\rangle$, where $A_1, A_2 \subseteq C$ are singletons, and A_3 is the whole set *C*.

Definition 5 Given a frame of discernment *C*, a mass function *m* is called a *triplet mass function* if it has no focal elements other than two singletons $\{x\}$, $\{y\}$ and the whole set *C*, where $x, y \in C$ such that

$$m(\{x\}) + m(\{y\}) + m(C) = 1$$

To obtain triplet mass functions, we define a focusing operation in a general context, called the *outstanding rule*.

Definition 6 Let *C* be a frame of discernment, and let *m* be a mass function with focal elements $\{x_1\}, \{x_2\}, \ldots, \{x_n\} \subseteq C$, $n \geq 2$, and $n \leq |C|$, then an outstanding rule is defined as a focusing operation σ on *m*, denoted by m^σ as follows:

$$m^\sigma(\{u\}) + m^\sigma(\{v\}) + m^\sigma(\Theta) = 1$$

where

$$\{u\} = \arg\max(\{m(\{x_1\}), m(\{x_2\}), \ldots, m(\{x_n\})\}) \quad (6)$$

$$\{v\} = \arg\max(\{m(\{x\}) \mid x \in \{x_1, x_2, \ldots, x_n\} - \{u\}\}) \quad (7)$$

$$m^\sigma(C) = 1 - m^\sigma(\{u\}) - m^\sigma(\{v\}) \quad (8)$$

and formula (2) is rewritten formula (9).

$$\varphi_i(d) = \{m^\sigma(\{u\}), m^\sigma(\{v\}), m^\sigma(C)\}\ ,\ 1 \leq i \leq M \quad (9)$$

and $A_1 = \{u\}$, $A_2 = \{v\}$ and $A_3 = C$.

Computing Two Triplet Mass Functions

Based on the number of singleton decisions, we also refer to a triplet as a structure of *two-point focuses*, and call the associated mass function a *two-point mass function*.

Suppose we are given two triplets $\langle\{x_1\}, \{y_1\}, C\rangle$ and $\langle\{x_2\}, \{y_2\}, C\rangle$ where $\{x_i\} \subseteq C$, $\{y_i\} \subseteq C$ $(i = 1, 2)$, and the associated triplet mass functions m_1 and m_2. The enumerative relationships between any two pairs of focal elements $\{x_1\}$, $\{y_1\}$ and $\{x_2\}$, $\{y_2\}$ are illustrated below:

1) if $\{x_1\} = \{x_2\}$ and $\{y_1\} = \{y_2\}$, then $\{x_1\} \cap \{y_2\} = \phi$ and $\{y_1\} \cap \{x_2\} = \phi$, so the combination of two triplet functions involves three different focal elements (two focal points equal).
2) if $\{x_1\} = \{x_2\}$ and $\{y_1\} \neq \{y_2\}$ then $\{x_1\} \cap \{y_2\} = \phi$, $\{y_1\} \cap \{x_2\} = \phi$ and $\{y_1\} \cap \{y_2\} = \phi$ or if $\{x_1\} \neq \{x_2\}$ and $\{y_1\} = \{y_2\}$, then $\{x_1\} \cap \{y_2\} = \phi$, $\{x_2\} \cap \{y_1\} = \phi$ and $\{x_1\} \cap \{x_2\} = \phi$, so the combination of two triplet functions involves four different focal elements (one focal point equal).
3) if $\{x_1\} \neq \{x_2\}$, $\{y_1\} \neq \{y_2\}$, $\{x_1\} \neq \{y_2\}$ and $\{y_1\} \neq \{x_2\}$, then $\{x_1\} \cap \{x_2\} = \phi$, $\{y_1\} \cap \{y_2\} = \phi$, $\{x_1\} \cap \{y_2\} = \phi$ and $\{y_1\} \cap \{x_2\} = \phi$, so the combination involves five different focal elements (totally different focal points).

Two Focal Points Equal

Suppose we are given two triplet mass function m_1 and m_2, along with two pairs of two-point focal elements $\{x_1\}$, $\{y_1\}$ and $\{x_2\}$, $\{y_2\}$ and $x_1 = x_2$, $y_1 = y_2$ ($x_1 \neq y_1$) and we have

$$m_1(\{x_1\}) + m_1(\{y_1\}) + m_1(C) = 1$$
$$m_2(\{x_2\}) + m_2(\{y_2\}) + m_2(C) = 1$$

First, we need to show that in what condition the combination of $m_1 \oplus m_2$ exists, and then we give formulae for computing their combination.

Theorem 1 Let C be a frame of discernment, let m_1 and m_2 be two triplet mass functions on C, and also let $\{x\}$, $\{y\}$ and $\{x\}$, $\{y\}$ ($x \neq y$) be two pairs of two-point focal elements with the condition of,

$$m_1(\{x\}) + m_1(\{y\}) + m_1(C) = 1\,,\ 0 \le m_1(\{x\}), m_1(\{y\}), m_1(C) \le 1$$
$$m_2(\{x\}) + m_2(\{y\}) + m_2(C) = 1\,, 0 \le m_1(\{x\}), m_1(\{y\}), m_1(C) \le 1$$

Then

$$K = 1 - m_1(\{x\})m_2(\{y\}) - m_1(\{y\})m_2(\{x\})$$

and m_1, m_2 are combinable if and only if

$$0 \le m_1(\{x\})m_2(\{y\}) + m_1(\{y\})m_2(\{x\}) < 1$$

Under this theorem, we can obtain the formulae of combining the two triplet mass functions $m_1 \oplus m_2$ below:

$$m_1 \oplus m_2(\{x\}) = K(m_1(\{x\})m_2(\{x\}) + m_1(\{x\})m_2(C) + m_1(C)m_2(\{x\}) \qquad (10)$$

$$m_1 \oplus m_2(\{y\}) = K(m_1(\{y\})m_2(\{y\}) + m_1(\{y\})m_2(C) + m_1(C)m_2(\{y\}) \qquad (11)$$

$$m_1 \oplus m_2(\{C\}) = K(m_1(\Theta)m_2(C)) \qquad (12)$$

where

$$K = 1 - \sum_{X \cap Y = \phi} m_1(X)m_2(Y) = 1 - m_1(\{x\})m_2(\{y\}) - m_1(\{y\})m_2(\{x\})$$

One Focal Points Equal

Given two triplet mass functions m_1 and m_2, a focal element in one triplet is equal to one in another triplet. Theorem 2 reveals that the two mass functions are combinable.

Theorem 2 Let C be a frame of discernment, m_1 and m_2 be two triplet mass functions on C, and also let $\{x\}$, $\{y\}$ and $\{x\}$, $\{z\}$ ($y \neq z$) be two pairs of focal elements with the following condition:

$$m_1(\{x\}) + m_1(\{y\}) + m_1(\Theta) = 1\,, 0 \le m_1(\{x\}), m_1(\{y\}), m_1(\Theta) \le 1$$
$$m_2(\{x\}) + m_2(\{z\}) + m_2(\Theta) = 1\,,\ 0 \le m_1(\{x\}), m_1(\{z\}), m_1(\Theta) \le 1$$

Then

$$K = 1 - m_1(\{x\})m_2(\{y\}) - m_1(\{y\})m_2(\{z\}) - m_1(\{x\})m_2(\{z\})$$

and m_1, m_2 are combinable if and only if the following constraint is held:

$$m_1(\{x\})m_2(\{y\}) + m_1(\{y\})m_2(\{z\}) + m_1(\{x\})m_2(\{z\}) < 1$$

By Theorem 2 and the orthogonal sum, a new mass function can be obtained from the two triplet mass functions, the general formulae for computing a new mass function are given below:

$$m_1 \oplus m_2(\{x\}) = K(m_1(\{x\})m_2(\{x\}) + m_1(\{x\})m_2(C) + m_1(C)m_2(\{x\})) \qquad (13)$$

$$m_1 \oplus m_2(\{y\}) = K(m_1(\{y\})m_2(C)) \qquad (14)$$

$$m_1 \oplus m_2(\{z\}) = K(m_1(\Theta)m_2(\{z\})) \qquad (15)$$

where

$$K = 1 - \sum_{X \cap Y = \phi} m_1(X)m_2(Y) = 1 - m_1(\{x\})m_2(\{z\}) - m_1(\{y\})m_2(\{z\}) - m_1(\{y\})m_2(\{z\})$$

Totally Different Focal Points

Now let us consider the case where there are no focal points in common. As indicated previously, the combination of such two triplet mass functions will involve five different focal elements. We first provide a theorem to ensure that such two triplet functions are combinable.

Theorem 3 Let Θ be a frame of discernment, let m_1, m_2 be two triplet functions, and $\{x\}$, $\{y\}$ and $\{u\}$, $\{v\}$ ($x \neq y$, $x \neq u$ and $y \neq v$) be two pairs of focal elements along with the following conditions:

$$m_1(\{x\}) + m_1(\{y\}) + m_1(\Theta) = 1\,,\ 0 \le m_1(\{x\}), m_1(\{y\}), m_1(\Theta) \le 1$$
$$m_2(\{u\}) + m_2(\{v\}) + m_2(\Theta) = 1\,,\ 0 \le m_1(\{u\}), m_1(\{v\}), m_1(\Theta) \le 1$$

Then

$$K = 1 - m_1(\{x\})m_2(\{u\}) - m_1(\{x\})m_2(\{v\}) - m_1(\{y\})m_2(\{u\}) - m_1(\{y\})m_2(\{v\})$$

and m_1, m_2 are combinable if and only if the following constraint is held:

$$0 \le m_1(\{x\})m_2(\{u\}) + m_1(\{x\})m_2(\{v\}) + m_1(\{y\})m_2(\{u\}) + m_1(\{y\})m_2(\{v\}) < 1$$

Given this theorem, we look at how the combination of mass function can be made. Suppose we are given the following expression:

$$m_1 \oplus m_2(\{x\}) = K(m_1(\{x\})m_2(\Theta)) \qquad (16)$$

$$m_1 \oplus m_2(\{y\}) = K(m_1(\{y\})m_2(\Theta)) \qquad (17)$$

$$m_1 \oplus m_2(\{u\}) = K(m_1(\Theta)m_2(\{u\})) \qquad (18)$$

$$m_1 \oplus m_2(\{v\}) = K(m_1(\Theta)m_2(\{v\})) \qquad (19)$$

where

$$K = 1 - \sum_{X \cap Y = \phi} m_1(X)m_2(Y) = 1 - m_1(\{x\})m_2(\{u\}) - m_1(\{x\})m_2(\{v\}) - m_1(\{y\})m_2(\{u\}) - m_1(\{y\})m_2(\{v\})$$

So far we have validated that any two triplet mass functions are combinable and established the formulae for computing the combinations of triplet mass functions. By repeatedly applying the outstanding rule at each computational step of combining two triplet mass functions, the results can be transformed to a new triplet mass function. Suppose we are given M triplet mass functions m_1, $m_2, \ldots, m_M$, they can be combined in any order due to Dempster's rule being both commutative and associative. Formula (15) is a pairwise orthogonal sums for combining any number of triplets. Its time complexity is approcimately $O(2\times|C|)$ and the final decision is the maxium selection of supports (strengths of confidence) for all classes.

$$m = m^1 \oplus m^2 \oplus \ldots \oplus m^M = [\ldots[[m^1 \oplus m^2] \oplus \ldots \oplus m^M] \qquad (20)$$

Similarly, we can consider three-point focuses and four-point focuses in terms of quartet and quintet. The quartet and quintet are conceptually simple and they have added properties that can be used to handle the more separated decisions drawn from lists of ordered decisions. For both cases, the theoretical validity is more complicated than that of the triplet structure. Here we can not detail each of them due to the limited space. In fact these methods are built in the same way as the triplet method.

Experiments and Evaluation

To evaluate the proposed method, we have implemented an experimental system composed of the supervised learning methods of kNN (Nearest-Neighbour) and kNNM (kNN model-based model), Rocchio and SVM (Support Vector Machine), as well as the evidential algorithms (DS) and the majority voting (MV) method. It is noted that when voting an even number of classifiers, we take into account the performance of the classifiers to determine which category will be assigned to a given instance.

Based on the number of singletons in triplet, quartet and quintet, the combining triplets will be more efficient than the others since there is less computation involved. The empirical results verify this assertion and show the time required for combining quartet functions is 48.67% longer than that for combining triplets, and the time of combining quintets is 77.89% longer than that of triplets on average.

Evaluation on Text Categorization

To evaluate performance, we have chosen two public benchmark data sets. The first one is *20-newsgroup*, it consists of 20 categories, and each category has 1 000 documents (Usenet articles), so the data set contains 20 000 documents in total. We use *information gain* to optimally select 5 000 features on average after removing stopwords and applying stemming.

The second one is *Reuters*-21578, which consists of 135 potential topic categories. In this experiment, we only used the ten most frequent categories (*acq*, *corn*, *crude*, *earn*, *grain*, *interest*, *money-fx*, *trade*, *ship* and *wheat*) to form a new data set, denoted by ModApte-10, where each category contains a varying number of documents and its total number of documents is 9818. After stemming and stopword removal, we use *information gain* to optimally select 1000 features on average.

The experiments are performed on a three partition scheme using ten-fold cross validation to avoid overfitting to some extent. We divide the data set into 10 mutually exclusive sets. For each fold, after the test set removal, the training set is further subdivided into 70% for a new training set and a 30% validation set. Apart from the evaluation of the performance of individual classifiers and parameter optimization, the validation set is also used to select the best combination of classifiers. The performance of the combinations of selected classifiers using the DS and MV algorithms is evaluated on the testing set.

In our experiments, different values of parameters have been tried for each of *k*NN, *k*NNM and Rocchio on the validation sets, to ensure that the experimental results faithfully reflect the performance of these algorithms. For SVM, the default parameter setting has been used since the SVM can automatically find an optimal parameter. The experimental results reported here are based on the parameter settings of $k = 130$ (*k*NN), $\varepsilon = 5$ (*k*NNM) and $(\beta, \gamma) = (1, 0.2)$ (Rocchio), and the performance of the learning algorithms and their combinations was measured by the *micro-averaging* F_1 metric (Yang, *et al*. 2000).

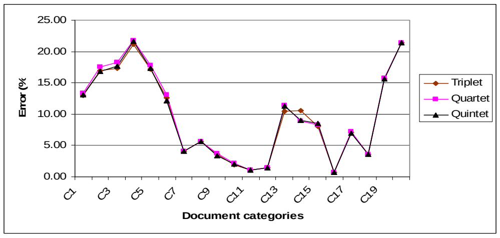

Figure 1: Performance comparison among triplet, quartet and quintet

Figure 1 presents a comparison on the performance of the best combination of classifiers SVM and Rocchio (SR) on triplet, quartet and quintet. The estimated performance of the combined classifiers on these structures are almost the same with the exception of categories C3 – C7 and C14 and C15. The performance of triplet is 0.18% better than that of quartet and 25.66% better than that of the full-list of decisions. There is almost no difference between the performance of triplet and quintet.

Table 1: Performance of individual classifiers and the best combined classifiers using DS and MV

	20-newsgroup (%)	ModApte-10 (%)
SVM (S)	87.88	87.27
kNNM (M)	83.34	82.39
kNN (N)	73.19	85.48
Rocchio (R)	84.84	78.96
SM (DS)	n/a	89.12
SR (DS)	90.32	n/a
SNR (MV)	n/a	87.02
SMNR (MV)	85.32	n/a

Table 1 presents the performance of four individual classifiers and the best combined classifiers, where SM (DS) stands for the combination of SVM and kNNM by using DS, SNR (MV) denotes the combination of classifiers SVM, kNN and Rocchio, etc. and n/a indicates that the combined classifier cannot achieve the best performance on the data sets. It can be seen that the best combined classifiers over the two data sets outperform any individual classifiers. The estimated performance of the best combination on 20-newsgroup is 90.32%, which is 2.44% better than the best individual classifier (SVM), whereas the best combined classifier on ModApte-10 is 1.85% better than the best individual SVM.

Table 1 also reports the performance of the combined classifiers by using majority voting on the data sets. Unfortunately, majority voting does not bring any benefits to the performance improvement of the combined classifiers in this occasion. The statistical significance analysis on the differences between categories, which are obtained via a ten-fold cross-validation, are also performed using a *t*-test with significance level of 95%, confirming that these performance differences are of statistical significance.

Discussion and Conclusions

We have presented novel evidence decision structures for representing classifier outputs – triplet, quartet and quintet – and the methods for obtaining them in terms of the outstanding rule and experimental results. Taking the best-evidence structure, the triplet, we performed a comparative analysis of Dempster's rule of combination and the majority voting method on combining decisions of classifiers on the 20-newsgroup and ModApte-10 datasets. We observed that in the case of DS, the best combination is that of two classifiers where one is the best and the other should have reasonable performance, whereas in MV case, the best combination consists of various numbers of classifiers. The combined classifiers using DS outperformed these using MV.

In the comparison of the best individual classifier and the best combined classifiers using DS and MV on the validation set with that on the test set, the first observation is that DS exhibits a higher overfitting than MV, the second is the overfitting of the combined classifiers is directly affected by the performance of the individual classifiers on the validation and test sets. If the performance of two individual classifiers is better than that of the other two classifiers, then the combination of the former two classifiers will outperform the combination of the later two.

To analyze the superiority of our method in text categorization, we compare our experimental results with the previous results of using the same benchmark datasets (Douglas, *et al*. 1997; Joachims, 1998; Sebastiani, 2002). In (Douglas, *et al*. 1997), the experimental results on the 20-newsgroup dataset can achieve 85.7% classification accuracy. Thus our experimental result is significantly better than it. In (Joachims, 1998), their experimental results on 20-newsgroup is 91.8% classification accuracy, which is 1.48% better than that of our DS method, and the estimated accuracy on ModApte-10 is comparable to the DS method. Turning the comparison to the former (on 20-newsgroup), the performance difference in the two results can be explained by two aspects. The first one is our evaluation method is a ten-fold cross validation whereas their work used a single hold out method (1/3-2/3 test-training splits). The second is that the classification accuracy of our best base classifier SVM is 87.88%, which 3.92% lower than that of their classifier. However, as discussed above, if the performance of our base classifier is comparable with that of his classifier, the best combined classifier using DS would outperform their classifier.

References

Dietterich T. G. 1997. Machine Learning Research: Four Current Directions. AI Magazine. 18 (4): 97-136.

Kuncheva L. 2001. Combining classifiers: Soft Computing Solutions. In Pal S.K. and Pal A. (Eds), Pattern Recognition: From Classical to Modern Approaches, 427-451.

Xu L., Krzyzak A. and Suen C. Y. 1992. Several Methods for Combining Multiple Classifiers and Their Applications in Handwritten Character Recognition. IEEE Trans. on System, Man and Cybernetics, 22 (3): 418-435.

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34 (1):1-47.

Yang, Y., Thomas Ault, Thomas Pierce. 2000. Combining multiple learning strategies for effective cross validation. In Proc of ICML'00, 1167-1182.

Kittler J., Hatef M., Duin R.P.W. and Matas J. 1998. On Combining Classifiers. IEEE Trans on pattern Analysis and Machine Intelligence. 20(3): 226-239.

Tumer K. and Ghosh J. Robust. 2002. Combining of Disparate Classifiers through Order Statistics. Pattern Analysis & Applications. 6 (1): 41 – 46.

Dzeroski, S. and Zenko, B. 2004. Is Combining Classifiers with Stacking Better than Selecting the Best One? Machine Learning. 54 (3): 255-273.

Denoeux T., 2000. A neural network classifier based on Dempster-Shafer theory. IEEE trans on Systems, Man and Cybernetics A. 30(2): 131-150.

Bi, Y., Bell, D, Guan, J.W. 2004. Combining Evidence from Classifiers in Text Categorization. In Proc of KES'04. 521-528.

Barnett, J A. 1981. Computational methods for a mathematical theory of evidence. In Proc of IJCAI'81. 868-875.

Shafer, G. 1976. A Mathematical Theory of Evidence, Princeton University Press, Princeton, New Jersey.

L. Baker and McCallum, A. 1998. A K. Distributional Clustering of Words for Text Classification. In Proc of 21st ACM International Conference on Research and Development in Information Retrieval, 96-103.

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proc of the 10th European Conference on Machine Learning, 137 – 142. Springer.