Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 10 (13): 1243-1254, 2010
ISSN 1812-5654

Tournament Structure Ranking Techniques for Bayesian Text Classification with Highly Similar Categories

[1]L.H. Lee, [2]D. Isa, [1]W.O. Choo and [3]W.Y. Chue
[1]Faculty of Information and Communication Technology-Perak Campus,
Universiti Tunku Abdul Rahman, Jalan University, Bandar Barat, 31900 Kampar, Perak, Malaysia
[2]Department of Electrical and Electronic Engineering, Faculty of Engineering,
University of Nottingham, Malaysia Campus, Jalan Broga, 43500 Semenyih, Selangor, Malaysia
[3]Faculty of Business and Finance, Universiti Tunku Abdul Rahman, Jalan University,
Bandar Barat, 31900 Kampar, Perak, Malaysia

Abstract: This study implements a series of tournament structure ranking technique to improve the classification accuracy of conventional Bayesian classification, especially in handling classification tasks with highly similar categories. Bayesian classification approach has been widely implemented in many real-world text categorization applications due to its simplicity, low cost training and classifying algorithms and ability in handling raw text data directly without needing extensive pre-processes. However, Bayesian classification has been reported as one of the poor-performing classification approaches. The poor performance of the Bayesian classification is critical especially in handling text classification tasks with multiple highly similar categories. In this study, we introduce a series of tournament structure based ranking classification techniques to overcome the low accuracy of conventional Bayesian classification which implements the flat ranking technique. Experiments that have been conducted in this research to show that the proposed Bayesian classifier embedded with tournament structure ranking techniques is able to ensure promising performance while dealing with knowledge domains with highly similar categories. This is due to the enhanced Bayesian classifier performs its classification tasks based on the implementation of multiple, iterative and isolated binary classifications and thus guarantee a low-error-rate Bayesian classification. As the result, an enhanced Bayesian classifier which is applicable to different types of domains of varying characteristics is introduced to handle the real world text classification problems effectively and efficiently.

Key words: Bayesian theorem, multi-class classification, round robin tournament, single-elimination tournament, swiss system tournament

INTRODUCTION

Document classification is defined as the task of learning methods for categorizing collections of electronic documents into their annotated classes, based on its contents. For several decades now, document classification in the form of text classification systems have been widely implemented in numerous applications such as spam e-mails filtering (Cunningham *et al*., 2003), (O'Brien and Vogel, 2002; Provost, 1999; Sahami *et al*., 1998), knowledge repositories (Hartley *et al*., 2006) and ontology mapping (Su, 2002), contributed by the extensive and active research works. An increasing number of statistical approaches have been developed for document classification, such as k-nearest-neighbor classification (Han *et al*., 1999), Bayesian classification (McCallum and Nigam, 2003), support vector machines (SVMs) (Burges, 1998; Joachims, 1998; Mohammadi and Gharehpetian, 2008; Sani *et al*., 2009) maximum entropy (Nigam *et al*., 1999), decision rule (Apte *et al*., 1994) and artificial neural networks (Abd Alla, 2006; Soltanizadeh and Shokouhi, 2008).

Among the classification approaches as mentioned above, Bayesian classification has been widely used due to its simplicity in both the training and classifying stage. Some research works have proven that Bayesian classification approach provides intuitively simple text generation models and performs surprisingly well in many other domains, under some specific conditions (Domingos and Pazzani, 1997; Kim *et al*., 2002; McCallum and Nigam,

Corresponding Author: Lam Hong Lee, Faculty of Information and Communication Technology-Perak Campus,
Universiti Tunku Abdul Rahman, Jalan University, Bandar Barat, 31900 Kampar, Perak, Malaysia
Tel: +605-4688888 Fax: +605-4661672

2003; Rish *et al*., 2001). Depending on the precise nature of the probability model, Bayesian classifiers can be trained very efficiently by requiring a small amount of training data to estimate the parameters necessary for classification. As the tradeoff of its simplicity, Bayesian classification approach has been reported to be less accurate than the discriminative methods such as SVMs (Chakrabarti *et al*., 2003; Joachims, 1998).

An optimum text classification system which is widely applicable in many real word applications should be equipped with the ability in handling different datasets of varying characteristics. This issue is addressed through the application of domain specific text classification algorithms. In this study, we propose the tournament structure ranking techniques, such as Round Robin tournament structure, Single-Elimination tournament Structure and Swiss System tournament structure, in order to enhance the conventional Bayesian classification. Our aim for introducing the tournament structure based ranking techniques to Bayesian classification is to deal with the datasets of multiple highly similar categories more effectively than the conventional version, which implements the flat ranking technique. The reason of introducing the tournament structure ranking techniques for conventional Bayesian classification is that the tournament structure ranking techniques perform the iterative binary classification for the computation of the posterior probability for the input document to be annotated to each category, Pr(Category|Document), in multiple categories classification tasks. Based on the our investigation, the binary classification results to a pure computation of Pr(Category|Document) between two categories without any distortion by the information from other available categories in the classification tasks. This will lead to an effective calculation with minimum noise, especially in the sensitive classification tasks with highly similar categories, where the posterior probability values for each category are very similar to each other. On the other hand, the conventional Bayesian classification performs the computation of Pr(Category|Document) in a single round by taking all the available categories into account. We found that in the classification tasks with multiple highly similar categories, the flat ranking technique is less effective than the tournament structure ranking techniques due to the all-in computation for the posterior probability for each category.

BAYESIAN CLASSIFICATION ALGORITHM

The conventional Bayesian classification approach performs its classification tasks starting with the initial step of analyzing text documents by extracting words which are contained in the document to generate a list of words. The list of words is constructed with the assumption that input document contains words w_1, w_2, w_3, ………, w_{n-1}, w_n, where the length of the document (in terms of number of words) is n.

Based on the list of words, the trained Bayesian classifier calculates the posterior probability of the particular word from the document being annotated to a particular category by using Bayes formula which is shown in Eq. 1, since each word in the input document contributes to the document's categorical probability.

$$\text{Pr (Category|Word)} = \frac{\text{Pr(Word | Category).Pr(Category)}}{\text{Pr(Word)}} \quad (1)$$

The derived equation above shows that by observing the value of a particular word, w_j, the prior probability of a particular category, C_i, $Pr(C_i)$ can be converted to the posterior probability, $Pr(C_i \mid w_j)$, which represents the probability of w_j being annotated to C_i. The prior probability, $Pr(C_i)$ can be compute from Eq. 2, 3:

$$Pr(C_i) = \frac{Total\_of\_Words\_in\_Ci}{Total\_of\_Words\_in\_Training\_Dataset} \quad (2)$$

$$= \frac{Size\_of\_Ci}{Size\_of\_Training\_Dataset} \quad (3)$$

Meanwhile, the evidence, which we call the normalizing constant of a particular word, w_j, $Pr(w_j)$ is calculated by using Eq. 4:

$$Pr(w_j) = \frac{\Sigma\, occurrence\_of\_wj\_in\_all\_categories}{\Sigma\, occurrence\_of\_all\_words\_in\_all\_categories} \quad (4)$$

The total occurrence of a particular word in every category can be calculated by searching the training data base, which is composed from the lists of word occurrences for every category. As previously mentioned, the list of word occurrence for a category is generated from the analysis of all the training documents in that particular category during the initial training stage. The same method can be used to retrieve the sum of occurrence of all words in every category in the training data base.

To calculate the likelihood of a particular category, C_i with respect to a particular word, w_j, the lists of word occurrence from the training data base is searched to retrieve the occurrence of w_j in C_i and the sum of all words in C_i. These information will contribute to the value of $Pr(w_j|C_i)$ given in Eq. 5:

Table 1: Table of words occurrence and probabilities

Words	Probability category 1	Probability category 2	Probability category 3	....	Probability category k-1	Probability category k
w_1	$Pr(C_1\| w_1)$	$Pr(C_2\| w_1)$	$Pr(C_3\| w_1)$	....	$Pr(C_{k-1}\| w_1)$	$Pr(C_k\| w_1)$
w_2	$Pr(C_1\| w_2)$	$Pr(C_2\| w_2)$	$Pr(C_3\| w_2)$	....	$Pr(C_{k-1}\| w_2)$	$Pr(C_k \| w_2)$
w_3	$Pr(C_1\| w_3)$	$Pr(C_2\| w_3)$	$Pr(C_3\| w_3)$	....	$Pr(C_{k-1}\| w_3)$	$Pr(C_k \| w_3)$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
w_{n-1}	$Pr(C_1\| w_{n-1})$	$Pr(C_2\| w_{n-1})$	$Pr(C_3\| w_{n-1})$	....	$Pr(C_{k-1}\| w_{n-1})$	$Pr(C_k \| w_{n-1})$
w_n	$Pr(C_1\| w_n)$	$Pr(C_2\| w_n)$	$Pr(C_3\| w_n)$	....	$Pr(C_{k-1}\| w_n)$	$Pr(C_k \| w_n)$
Total	$\Sigma Pr(C_1\|W)$	$\Sigma Pr(C_2\|W)$	$\Sigma Pr(C_3\|W)$	....	$\Sigma Pr(C_{k-1}\|W)$	$\Sigma Pr(C_k \|W)$
Probability of input document	$\frac{\Sigma Pr(C_1\|W)}{n}$	$\frac{\Sigma Pr(C_2\|W)}{n}$	$\frac{\Sigma Pr(C_3\|W)}{n}$	....	$\frac{\Sigma Pr(C_{k-1}\|W)}{n}$	$\frac{\Sigma Pr(C_k\|W)}{n}$

$$Pr(w_j \mid C_i) = \frac{occurrence\_of\_wj\_in\_Ci}{\Sigma\, occurrence\_of\_all\_words\_in\_Ci} \quad (5)$$

Based on the derived Bayes formula for text classification and the value of the prior probability Pr(Category), the likelihood Pr(Word|Category) and the evidence Pr(Word), along with the posterior probability, Pr(Category|Word) of each word in the input document being annotated to each category can be measured.

The posterior probability value of each word being annotated to a category is then filled in at the appointed cells in a table respectively as shown in Table 1. After all the Probability cells have been filled, the overall probability for an input document to be annotated to a particular category, C_i is calculated by dividing the sum of each of the Probability column with the length of the document (total number of words), n, which is shown in Eq. 6:

$$Pr(C_i \mid Document) = \frac{Pr(C_i \mid w_1, w_2, w_3, ..., w_{n-1}, w_n)}{n} \quad (6)$$

where, w_1, w_2, w_3, ..., w_{n-1}, w_n, are words that extracted from the input document.

The conventional Bayesian classifier is able to determine the right category of an input document by referring to the highest probability value calculated by the trained classifier based on Bayes formula. The right category is represented by the category which has the highest posterior probability value, Pr(Category|Document), as stated in bayes decision rule.

IMPLEMENTATION OF TOURNAMENT STRUCTURED BAYESIAN CLASSIFIER

The basic structure of the proposed tournament structured Bayesian classification approach is shown in Fig. 1. The important features of the classification system are the input analysis facility, the Bayesian classifier equipped with tournament structure ranking techniques, which work together to provide a method of guaranteeing an optimum classification for multiple highly similar classification tasks. By introducing the tournament structure classification ranking techniques, the enhanced Bayesian classifier is able to handle the classification tasks effectively and efficiently for datasets with highly similar categories, beyond the performance of the conventional Bayesian classifier.

As shown in Fig. 1, the task of our proposed classifier is divided into two stages, the training phase and the classifying phase. Hence, the dataset which has been acquired for the classification task have to be separated into two sets, one set contains the training documents which have been well-labeled by domain experts and another set contains the documents for testing and evaluation purposes. During the training phase, the training documents are entered into the classifier and the classifier will organize all the documents into different groups by the learning facility according to their label/category. The grouped documents are then stored into a domain database to build the training set. Before the classifier is used to perform its classification task, it needs to be trained with the training set. The training data generation facility will generate the training data, in the form of list of words and the frequency of their occurrence for each category. The lists of words and their frequency of occurrence for each category are generated by using a simple data structure algorithm and they are used to train the Bayesian classifier to equip the classifier ready for the classifying stage.

After the training phase, the classifier is ready to perform its task of classifying unlabeled documents. However, for the ease of evaluation of the performance of the classifier, all the testing documents have also been labeled by the domain experts but this information is reserved to be compared with the results generated by the classifier after processing the testing documents. At the first step of the classifying phase, the testing document is entered into the classifier and the input analysis facility

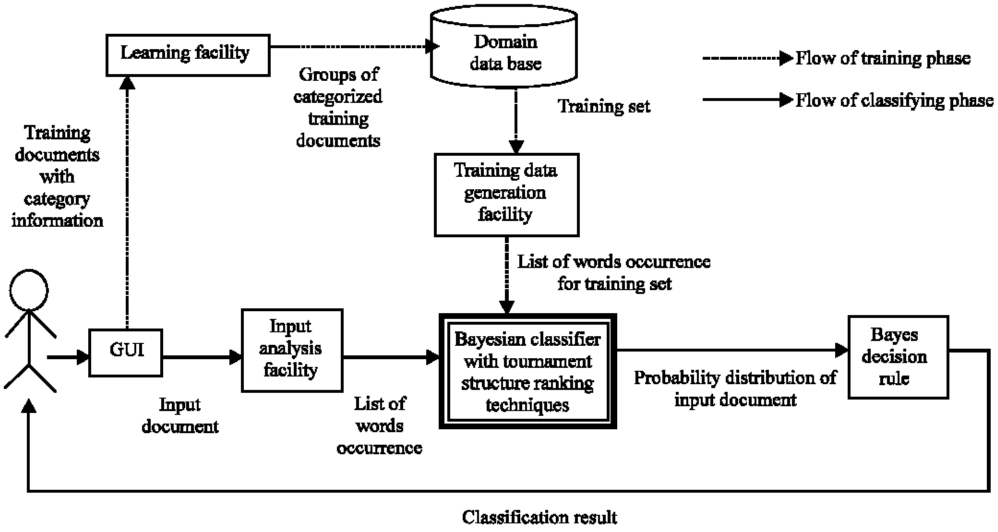
Learning facility
Domain data base
Groups of categorized training documents
Training set
Flow of training phase
Flow of classifying phase
Training documents with category information
Training data generation facility
List of words occurrence for training set
GUI
Input analysis facility
Bayesian classifier with tournament structure ranking techniques
Bayes decision rule
Input document
List of words occurrence
Probability distribution of input document
Classification result

Fig. 1: Block diagram of proposed Bayesian classifier with tournament structure ranking techniques

will analyze the document and generate the list of words associated with their frequency of occurrence for the document. The classifier will process the list of words occurrence in order to compute the posterior probability values of each testing document to be annotated to each of the categories. After that, the probability distribution of the testing document in the vector space is generated and the classifier will generate the classification result as the right category of the input document by referring to the category which has the highest posterior probability value among the other available categories. The iterative classifying process will continue until all the testing documents have been categorized.

As most of the Artificial Intelligence (AI) learning machines need to be trained before they can perform their tasks, our proposed tournament structured Bayesian classifier requires a proper initial training to perform its classification tasks effectively and efficiently. For initial training purposes, training documents have been acquired from the well-organized knowledge domain, or human experts to construct a repository of training data. In order to build the initial training set, human experts must manually label the training documents and ensure they are well-categorized. For each category, a reasonable number of training documents need to be acquired. Since, we have implemented Bayesian classification approach as our probabilistic classifier, a training set with large number of entries is not required due to the individual attribute values are assumed as statistically independent. Bayesian classifier can be very robust to violations of its independent assumption. In other words, the classifier is robust enough to ignore serious deficiencies in its underlying naïve probability model, for example, the assumption of the independence of probable events. This robustness is encapsulated in the relative magnitudes of the probabilities (relatively constant for the same category).

The training documents are in the form of raw text documents to fit our system's requirements. This is an advantage of Bayesian classification that able to handle raw text data, without requiring any transformation step to transform text data into representation suitable format, typically in numerical format, as requiring by most of the classification approaches such as the support vector machines and the k-nearest neighbor. At the basic level, Bayesian classifier examines a set of training documents that have been well-organized and categorized and then compares the content of all categories in order to build a database of words and their occurrences by using a simple data structure algorithm. The database is used to identify or predict the membership of future documents to their right category, according to the probability of certain words occurring more frequently for certain categories. It overcomes the obstacles faced by many static technologies, such as blacklist checking and word-to-word comparisons using static databases filled with pre-defined keywords. The training set can be updated by human experts to upgrade the quality of training data, which will directly contribute to the performance of the probabilistic classifier. As the future work, the updates of training set of our classification system will be enhanced with an automatic intelligent domain database update facility which will greatly improve the ability and the performance of the system.

The input to the classification system is in the form of text document. The system analyses the input text data before proceeding to the probabilistic classifier for data identification and classification of the input document. The input analysis facility uses a simple word extraction algorithm to extract each individual word from the input text document prior to generating a list of words.

The tournament structured Bayesian classifier plays an important role when the classification process advances to the following stage. For instance, a particular word has its probability occurring in a particular category, then, the tournament structured Bayesian classifier computes the posterior probability of words being annotated to each category based on their frequency of occurrence in the particular category, using Bayes formula. The algorithms for the computation of posterior probability by each of the tournament structure ranking techniques are discussed in the sub-sections below. The overall probability of an input document being annotated in a particular category is computed based on the sum of the posterior probability values of every word in the document, divided by the length of the document (in terms of total number of words). The probability distribution of the particular input document in the vector space can be obtained by integrating all the posterior probability of an input document being annotated in each category, individually and sequentially, according to the order of categories in the vector space.

The output from the Bayesian classifier is the probability distribution of input document in the vector space. The probability distribution of input document is not only carrying vectorized data of the document in the format of numerical values, but also information about the right category of the input document. As we are implementing Bayesian classification as the core of our classifier, the right category of the input document can be determine by referring to the category which has the highest posterior probability value, Pr(Category|Document) among the other available categories in the probability distribution, as stated in Bayes Decision Rule.

Flat ranking technique: The flat classification ranking technique is widely-implemented in the conventional Bayesian probability calculation in order to support multi-dimensional classification task. The probability value for a document D to be annotated to a category C is computed as Pr(C|D). With the assumption that we have a category list as C_1, C_2, C_3, C_4 C_5 ……., C_n thus, each document has n-associated probability values, where document D will have $Pr(C_1|D)$, $Pr(C_2|D)$, $Pr(C_3|D)$, $Pr(C_4|D)$, $Pr(C_5|D)$, ……, $Pr(C_n|D)$. The n-associated probability values represent the probability distribution for document D in the vector space. By implementing the flat ranking technique, the probability value of an input document being annotated to a particular category is calculated by considering all the available categories together in a single round. During the calculation of the prior probability, Pr(Category), the normalization constant, Pr(Word) and also the likelihood of a particular category with respect to that particular word, Pr(Word|Category), all the available categories in the classification task are involved to produce a complete calculation of the probability distribution in the vector space. The flat classification ranking technique selects the category with the highest posterior probability value, Pr(C|D), as the right category being annotated for the input document, based on Bayes Decision Rule.

Based on our investigation, in the multi-dimensional classification tasks which involve highly similar categories, the flat ranking technique suffers from the distorted calculation of the posterior probability for the input document, D, to be annotated to each category, $Pr(C_i|D)$. Since all the categories are highly similar to each other, the information (list of keywords and their occurrence) from each category is also highly similar. This will lead to a distortion in computation while calculating the posterior probability for a particular input document to be annotated to a particular category by taking all the available categories together in a single round. Thus, in this study, we introduce the tournament structure ranking techniques, which perform the calculation of posterior probabilities based on the iterative binary classification, to overcome the problem faced by the flat ranking technique.

Round robin tournament ranking technique: Tournament structures are possible to be implemented in classification ranking algorithm in order to handle multi-categories classification. In round robin tournament ranking technique, the calculation of the posterior probability values of an input document being annotated to each available category is performed only between two categories. It is an iterative binary classification algorithm and each competitor plays against all the others an equal number of times, typically once. The round robin tournament ranking technique contributes to a relatively complete and pure competition among all the categories, as compared to the flat ranking technique. The process iterates until every category has competed against all the others once.

The structure of the round robin tournament ranking technique that implemented in our proposed system is based on the Host and Guest concept. Firstly, all the categories are randomly sorted. The first category will act as the host of the initial round and plays against all the

Table 2: Structure of the round robin tournament ranking technique

▨	Category 1	Category 2	Category 3	Category 4
Category 1	▨			
Category 2		▨		
Category 3			▨	
Category 4				▨
Total				

others which are ranked below it as guests, individually and sequentially, in the matches conducted in the first round. In the second round, the second category will become the host and those categories which are ranked below it will compete against the host. This process iterates until all categories have competed against the others an equal number of time, typically once. The illustration of structure of the round robin tournament ranking technique that implemented in our proposed classification system is as shown in Table 2.

There are several methods available to determine the final winner after the iterative calculation processes have completed. One of the methods is awarding the winning category of each match with a score, typically 1 and the loser is not awarded with any score, or in other words, score 0 is awarded. The scores gained from each match by each category are then added together after the competition until the calculation has completed. The category with the highest final score will be the overall winner, which represents the right category of the input document.

There is a situation where dilemma occurs in determining the right category of an input document when two or more categories have the same highest final score. This situation can be avoided by awarding the two competing categories of each match with the score which is equal to their posterior probability value calculated based on Bayes formula from the binary classification. With this method, the final highest scores for each category are rarely to be the same.

As a result, the round robin tournament ranking technique has an improved ability in distinguishing highly similar categories, since it performs an iterative binary classification algorithm. The binary classification algorithm has a greater ability in differentiating highly similar categories as compared to the flat ranking technique, which involves all the available categories together in a single round of classification. The binary classification algorithm is smart enough to isolate two categories temporarily and perform the posterior probability calculation without considering the other parties. However, the iterative binary classification process consumes a relatively long time as compared to other algorithms, such as the flat ranking technique. Therefore, the round robin classification technique is not efficient for the implementation in the classification tasks on datasets with a high dimensionality of categories.

Single elimination tournament ranking technique: By comparing with other classification ranking techniques, the single elimination tournament ranking technique has some restrictions. Firstly, most often the number of categories is fixed as a power of two. Somehow, in a given situation where the number of categories is not a power of two, typically the highest-rated categories from the previous accomplishment will be advanced to the second round without joining any match in the first round. Besides, seeding is recommended as a pre-process to prevent the highest-rated categories from being scheduled to face each other in the early stages of competition. The seeds ranking process can be executed by the classifier equipped with the classification ranking technique in either the flat ranking technique or the round robin tournament ranking technique. Therefore, the single elimination algorithm is more suitable to be implemented at the back-end of a hybrid ranking technique system.

As similar to the round robin tournament ranking technique, the single elimination ranking technique also performs the probability values calculation in the form of binary competition for every match. In the first round, the best category is played against the worst and the second best against the second worst and so on for the rest. Brackets are set up, so that the top two seeds could not possibly meet until the final round, none of the top four can meet before the semifinals and so on. The winner of each match will proceed to the next round while the looser of each match will be eliminated from the tournament. This concept is applicable in the following rounds until the overall winner is represented by the winner of the final round. The structure of the single elimination tournament ranking technique that implemented in our proposed system is as shown in Fig. 2. In Fig. 2, a simple assumption has been made that the classification task consists of 16 categories.

In contrast to the round robin tournament ranking technique, as rounds progress, the successive rounds of the single elimination tournament ranking technique halve the number of remaining categories. This is due to the single elimination tournament ranking technique progresses the winners from the previous round to the next round and eliminates the losers. The single elimination tournament ranking technique is suitable for implementation in the domains which have a large number of categories. Since this ranking technique is also a binary classification-based algorithm, it has a great ability in handling the classification tasks which involve categories with a high degree of similarities.

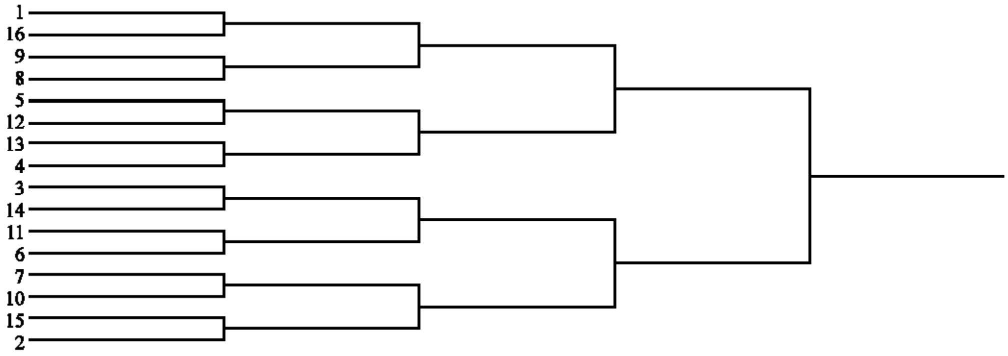

Fig. 2: Structure of the single elimination tournament ranking technique

Table 3: Structure of the swiss system tournament ranking technique

Round 1	Round 2	Round 3
#1 plays #5, #1 wins	#1 plays #3, #1 wins	#1 plays #2, #1 wins
#2 plays #6, #2 wins	#2 plays #4, #2 wins	#3 plays #4, #3 wins
#3 plays #7, #3 wins	#5 plays #7, #5 wins	#5 plays #6, #5 wins
#4 plays #8, #4 wins	#6 plays #8, #6 wins	#7 plays #8, #7 wins
After one rounds, the standing are	After two rounds, the standing are	After three rounds, the standing are
#1 1-0	#1 2-0	#1 3-0
#2 1-0	#2 2-0	#2 2-0
#3 1-0	#3 1-1	#3 2-1
#4 1-0	#4 1-1	#4 1-2
#5 0-1	#5 1-1	#5 2-1
#6 0-1	#6 1-1	#6 1-2
#7 0-1	#7 0-2	#7 1-2
#8 0-1	#8 0-2	#8 0-3

Swiss system tournament ranking technique: The Swiss system tournament ranking technique can be implemented independently or at the back-end of a hybrid ranking technique system. The initial seeding of a Swiss system tournament ranking technique is not compulsory, but is recommended. The competing categories are then divided into two parts, the top half is paired with the bottom half. For instance, if there are eight categories in the classifier, the first category is paired with the fifth category; the second is paired with the sixth, the third is paired with the seventh and so on. The structure of the Swiss system tournament ranking technique that is implemented in our proposed system is shown in Table 3.

After the first round of the competition, the winners from the first round will play against the winners and the losers will play against the losers. Similar to the round robin tournament ranking technique, the winning category of each match is awarded with a score, typically 1 and the loser is not awarded with any score. In further rounds, each competitor will be pitted against another competitor who has the same score. Modifications are then made to prevent competitors from meeting each other twice.

In contrast to the round robin tournament ranking technique, the Swiss system tournament ranking technique can determine the top rank and bottom rank competitors with fewer rounds, although the middle rankings are rather unreliable. On the other hand, in our classification tasks, we are only interested in the final overall winner, which represents the right category of input document. Therefore, the Swiss system tournament ranking technique is applicable in our classification system with large number of categories. Besides, similar to other binary classification based ranking techniques, it is suitable to be implemented in classification tasks where the domain contains categories of high degree of similarities.

However, the number of competing categories has become the biggest obstruction for the Swiss system tournament ranking technique. Likened to the round robin tournament ranking technique, the Swiss system tournament ranking technique has the potential in facing a dilemma in determining the right category of input documents, or the final winner. There is a possibility that two or more categories have the same highest and perfect score, winning all the matches but have never faced each other. Therefore, the conventional ranking technique needs to be slightly modified in terms of the number of rounds played. To determine a clear overall winner, we have applied the same concept with the single elimination tournament ranking technique in terms of number of rounds that is the base 2 logarithm and the number of competitors is rounded up by the power of two.

EXPERIMENTS, EVALUATIONS AND DISCUSSION

In order to evaluate the performance of the tournament structure ranking classification techniques to be implemented to classification tasks on datasets with high degree of similarity between their available categories, we have acquired four datasets of varying characteristics. Each of these datasets has its own unique characteristics in terms of the degree of similarity between categories and the dimensionality of categories, as shown in Table 4.

The degree of similarity can be defined by the designers or organizers of the dataset. There is a way to measure the degree of similarity between categories for a particular dataset and that is by comparing the words in the vocabulary of every document from each category in a dataset. If most of the categories from a dataset share a majority of the keywords, say more than 80% of the vocabulary size in each category, the dataset is considered to have a high degree of similarity between categories. A dataset with a low degree of similarity between categories is built from categories which are easily differentiated from each other by a set of specific keywords where the size of the specific keywords set is typically 80% of the vocabulary size of each category.

There exists another method to measure the degree of similarity between categories for a particular dataset. This method firstly computes the probability distribution in the vector space for the documents in each of the available categories in the dataset, which represented by the n-associated posterior probability values for the documents to be annotated to each of the n-categories. After that, the probability distributions of the documents in the vector space are analyzed to measure the similarity between the categories. This could be done by comparing the values of each of the posterior probability values for the documents to be annotated to each of the available categories among the others and a threshold is set for the difference between each of the n-associated posterior probability values for the documents. If the n-associated posterior probability values for the documents to be annotated to each of the available categories, $Pr(C_1|D)$, $Pr(C_2|D)$, $Pr(C_3|D)$, $Pr(C_4|D)$, $Pr(C_5|D)$, ……, $Pr(C_n|D)$, are close to each other, in other words, the difference between each of the n-associated posterior probability values for the documents are very small, or less than the threshold, the dataset is considered as a dataset with highly similar categories.

As for the number of categories of a dataset, it represents the total number of available categories in a dataset which has been pre-defined and can be easily measured.

Table 4: Four datasets and their characteristics which have been used for experiments and evaluations

Dataset (source)	Degree of similarity between categories	Dimensionality of categories
Featured articles (wikipedia)	Normal	23
Vehicles (wikipedia)	Low	4
Mathematics (arxiv.org)	High	8
Automobiles (wikipedia)	High	9

Table 5: List of categories of the featured articles dataset

No.	List
1	Art, Architecture and Archaeology
2	Biology and medicine
3	Business, economics and finance
4	Chemistry and mineralogy
5	Computing
6	Culture and society
7	Engineering and technology
8	Geography and places
9	Geology, geophysics and meteorology
10	History
11	Language and linguistics
12	Law
13	Literature
14	Mathematics
15	Media
16	Music
17	Physics and astronomy
18	Politics and government
19	Religion and beliefs
20	Royalty, nobility and heraldry
21	Sport and games
22	Transport
23	War

Experiments on the featured articles dataset: The Featured articles dataset was designed and organized by our research group by extracting different types of articles from the wikipedia website. These articles were acquired from 23 randomly selected categories and were converted into plain text documents for the ease of accessing by our prototype text classifier. The list of selected categories for the Featured Articles dataset is shown in Table 5.

All the 23 categories from the featured articles dataset are generally unrelated to each other though some of the categories such as history, religion and beliefs, War are slightly related. Other categories such as engineering and technology and transport can also be considered as similar categories. However, the degree of similarity between categories of this dataset is considered as normal because each category still has its own set of specific keywords that differentiates itself from other categories.

Table 6 shows the classification accuracies of the classifiers built from different classification ranking techniques on the featured articles dataset.

The experimental results shown in Table 6 show that the flat ranking classifier performs the best among the other classifiers while handling the classification task on the Featured Articles dataset. This classifier has achieved

Table 6: Experimental results of the classifiers with different classification ranking techniques on the featured articles dataset

Classification ranking techniques	Classification accuracy (%)
Flat ranking	71.15
Round robin tournament ranking	70.51
Single elimination tournament ranking	70.83
Swiss system tournament ranking	70.51

Dataset: Featured articles, Training set: 230 documents, Testing Set: 929 documents

Table 7: List of categories in the vehicles dataset

No.	List
1	Aircrafts
2	Boats
3	Cars
4	Trains

Table 8: Experimental results of the classifiers with different classification ranking techniques on the vehicles dataset

Classification ranking techniques	Classification accuracy (%)
Flat ranking	88.64
Round robin tournament ranking	80.00
Single elimination tournament ranking	77.27
Swiss System Tournament Ranking	77.50

Dataset: Vehicles, Training set: 200 documents, Testing set: 440 documents

the highest accuracy rate of 71.15%. The baseline performance is 70.51%, achieved by both the round robin tournament ranking classifier and the Swiss system tournament ranking classifier. As we can observe from Table 6, the classifiers equipped with the flat ranking technique have outperformed the classifiers equipped with the tournament structure ranking techniques.

Experiments on the vehicles dataset: The Vehicles dataset was built by extracting vehicle related articles from the Wikipedia website. This dataset was acquired by extracting articles from four sub categories in the Vehicles category, which are Aircrafts, Boats, Cars and Trains. All four categories are easily differentiated and each category has a set of unique keywords. The list of categories of the Vehicles dataset is shown in Table 7.

The experimental results of the classification task on the Vehicles dataset using different classifiers, built from different classification ranking techniques, are shown in Table 8.

The experimental results shown in Table 8 show that the flat ranking classifier has achieved the highest classification accuracy rate of 88.64% as compared to other classifiers. The baseline performance is 77.27%, produced by the single elimination tournament ranking classifier.

Experiments on the mathematics dataset: A dataset containing articles about mathematical topics has been acquired from arxiv.org and it is one of the datasets used to evaluate the performance of our prototype text classifier. This dataset contains eight mathematical sub-categories. The list of categories of the Mathematics dataset is shown in Table 9.

Table 9: List of categories of the Mathematics dataset

No.	List
1	Algebraic geometry
2	Analysis of PDEs
3	Combinatorics
4	Differential geometry
5	Mathematical physics
6	Number theory
7	Probability
8	Statistics

Table 10: Experimental results of the classifiers with different classification ranking techniques on the mathematics dataset

Classification ranking techniques	Classification accuracy (%)
Flat ranking	80.42
Round robin tournament ranking	81.25
Single elimination tournament ranking	81.25
Swiss system tournament ranking	81.25

Dataset: Mathematics, Training set: 80 documents, Testing Set: 240 documents

All eight categories in the mathematics dataset have high degree of similarity. Due to the fact that the eight categories of this dataset are the sub-topics in mathematics, the documents in all the categories share a lot of common mathematical keywords. There are only a few specific and unique keywords that differentiate each topic. Therefore, the frequency in which those specific keywords occur in that certain category is the main key in differentiating these topics.

Table 10 shows the experimental results of the classification task on the Mathematics dataset by classifiers with different classification ranking techniques.

The pattern of the experimental results in Table 10 is different from the previous experiments as discussed earlier. The flat ranking classifier scores lower classification accuracy than the classifiers equipped with tournament structure classification ranking techniques. The flat ranking classifier scores an accuracy of 80.42%, which is the baseline performance in this experiment. All the three classifiers equipped with the tournament structure classification ranking techniques achieved the best performance in this experiment with the highest classification accuracy of 81.25%.

Experiments on the automobiles dataset: The Automobiles dataset is a dataset which was designed and organized by collecting articles about automobiles from the Wikipedia website. This dataset contains nine categories of automobiles, differentiated in terms of geographical regions and types. Table 11 shows the list of categories in the Automobiles dataset.

All nine categories in the Automobiles dataset are considered highly similar to each other since the documents from these categories describe the

Table 11: List of categories of the mathematics dataset

No.	List
1	American mini vans
2	American Sports cars
3	American SUVs
4	Asian mini vans
5	Asian sports cars
6	Asian SUVs
7	European mini vans
8	European sports cars
9	European SUVs

Table 12: Experimental results of the classifiers with different classification ranking techniques on the automobiles dataset

Classification ranking techniques	Classification accuracy (%)
Flat ranking	77.78
Round robin tournament ranking	86.67
Single elimination tournament ranking	86.67
Swiss system tournament ranking	86.67

Dataset: Automobiles, Training set: 180 documents, Testing set: 90 documents

characteristics of automobiles. However, each category has a minor set of unique keywords that describes the types of the automobiles and their geographical regions to differentiate their characteristics from other categories.

The experimental results of the classification task on the automobiles dataset using classifiers of different classification ranking techniques are shown in Table 12.

The results shown in Table 12 show that all the three classifiers equipped with tournament structure ranking techniques achieve the greatest classification accuracy with a rate of 86.67%. The baseline performance is 77.78%, achieved by the classifier equipped with the flat ranking technique. The pattern of the experimental results as shown in Table 12 is similar with the pattern of the experimental results shown in Table 10, where all the three classifiers with tournament structure ranking techniques outperform the classifier equipped with flat ranking technique. The similar pattern of the experimental results of the mathematics dataset and the experimental results of the automobiles dataset is due to the fact that both of these datasets have high degree of similarity between their available categories.

Discussion of experimental results: Based on the experimental results that we have obtained from our experiments on datasets of varying characteristics, we can conclude that for datasets with low and normal degree of similarity between the available categories, the classifier with flat ranking technique has achieved the best classification performance among all the classifiers which equipped with the tournament structure ranking techniques. However, for the datasets with highly similar categories in terms of their contents, the classifiers equipped with the tournament structure ranking techniques outperform the flat ranking classifier.

Bayesian classification has been reported by many research works as one of the poor performing classification approaches in general. As the results, many active research works have been carried out to clarify the reasons that Bayesian classifier fails in classification tasks and improve the performance of conventional approach by implementing some enhancing techniques (Domingos and Pazzani, 1997; Eyheramendy *et al*., 2003; Kim *et al*., 2002; McCallum and Nigam, 2003; Rish *et al*., 2001). From the experiments that we conducted in our previous works, we found that conventional Bayesian classifier fails to achieve high classification accuracy in some of the classification domains, as reported by the research works mentioned above. Based on our experiences, the poor performance of the conventional Bayesian classification is distinct when the classifier deals with the datasets with highly similar categories. One of the reasons that we found from our investigation for the poor performance of conventional Bayesian classification in such a situation is that it suffers from the confusion in differentiating highly similar categories in the classification task. This is due to the conventional Bayesian classification takes all the available categories into the computation of posterior probability values for a document to be annotated to each available category in a single round. In the datasets with highly similar categories, all the categories share a lot of the same keywords. By considering this kind of categories in a single round of computation, it would be difficult to obtain good classification results. As the solution for handling domains with highly similar categories in a classification task, we have proposed a series of tournament structure classification ranking techniques, which perform the iterative binary classification to overcome this problem. The iterative binary classification algorithm is able to solve this problem by isolating two categories temporarily for the computation of the posterior probability values for a document to be annotated to the competing categories, without considering the other categories.

For both of the classification tasks on the featured articles dataset and the vehicles dataset, the flat ranking technique outperforms all three tournament structure based ranking techniques. This is probably caused by the appropriately divided categories of the input documents which enables the flat ranking technique to differentiate all the other categories in a single round. However, in a tournament structure based ranking technique, the right category of an input document may not be able to win most of competitions in the tournament system. This situation is commonly found in tournament systems where the best competitor may not be able to win a majority of the competitions or is accidentally eliminated before the final round. Therefore, the flat ranking classification technique is more suitable for implementation rather than the tournament structure based ranking techniques in classification tasks where datasets with low and normal degree of similarity between categories are involved. The tournament structure based ranking techniques are more applicable in classification tasks which involve highly similar categories wherein the flat ranking technique is not as effective as the tournament structure ranking techniques in determining the right category of the input documents or queries.

As we expected, the flat classification ranking technique performed poorer than the tournament structure based classification ranking techniques in the classification task on the Mathematics dataset and the Automobiles dataset which have high degree of similarity among their available categories. This is due to the fact that flat ranking technique calculates the probability values of each category of the input data in a single round by taking into account all categories. This will lead to great confusion in the classification task since all categories share a lot of the same keywords, in other words, the contents of all the categories are very similar to each other. In such a situation, a binary classification technique is able to overcome this problem. Therefore, the tournament structure based ranking techniques that perform the iterative binary classification are more suitable for implementation in classification tasks of datasets that have high degree of similarity among categories, as compared to the flat ranking technique.

Therefore, based on the experimental results that have been obtained in this study, we can conclude that classifiers equipped with the tournament structure ranking techniques are more suitable for application in classification tasks with datasets with highly similar categories, whereas the conventional flat ranking technique is more suitable to be implemented in classification tasks with datasets with low and normal degree of similarities between their available categories.

CONCLUSIONS

In conclusion, the experimental results presented in this study show that Bayesian classifiers that handle classification tasks on datasets with a low or normal degree of similarity among categories perform better using the flat classification ranking technique. The tournament structure ranking techniques, such as the round robin tournament ranking technique, the single elimination tournament ranking technique and the Swiss system tournament ranking technique perform better than the flat ranking technique in classification tasks on datasets consisting of highly similar categories. This is due to the tournament structure ranking techniques perform the iterative binary classification algorithm for the computation of the posterior probability for the input document to be annotated to each category, Pr(Category|Document), in multi-dimensional classification tasks. On the other hand, the flat ranking classification technique performs the computation of Pr(Category|Document) in one round by taking all the available categories into account.

Based on our investigation, the binary classification algorithm from the tournament structure based raking techniques contributes to a pure computation of Pr(Category|Document) between two categories without any distortion by the information from other available categories in the classification tasks. This results to an effective calculation without any noise and distortion, especially in the sensitive classification with highly similar categories, where the posterior probability values for each category are very similar to each other. However, the major drawback of the tournament structure classification ranking techniques is that they consumes a longer processing time and requires a convoluted computing process to perform the classification task. In the future, our research emphasizes in enhancing the ability and performance of our existing prototype by introducing more facilities to develop simple, low consumptions and widely applicable Bayesian classifiers.

REFERENCES

Abd Alla, A.N., 2006. Three phase induction motor faults detection by using radial basis function neural network. J. Applied Sci., 6: 2817-2820.

Apte, C., F. Damerau and S.M. Weiss, 1994. Automated learning of decision rules for text categorization. ACM Trans. Inform. Syst., 12: 233-251.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowledge Discovery, 2: 121-167.

Chakrabarti, S., S. Roy and M.V. Soundalgekar, 2003. Fast and accurate text classification via multiple linear discriminant projection. VLDB J., 12: 170-185.

Cunningham, P., N. Nowlan, S.J. Delany and M. Haahr, 2003. A case-based approach to spam filtering that can track concept drift. In The ICCBR'03 Workshop on Long-Lived CBR Systems, Trondheim, Norway. http://www.scss.tcd.ie/publications/tech-reports/reports.03/TCD-CS-2003-16.pdf.

Domingos, P. and M. Pazzani, 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29: 103-130.

Eyheramendy, S., A. Genkin, W.H. Ju, D. Lewis and D. Madigan, 2003. Sparce bayesian classifiers for text categorization. Department of Statistics, Rutgers University. http://www.stat.rutgers.edu/~madigan/PAPERS/ jicrd-v13.pdf.

Han, E.H., G. Karypis and V. Kumar, 1999. Text categorization using weight adjusted k-nearest neighbour classification. Department of Computer Science and Engineering, Army HPC Research Center, University of Minnesota.

Hartley, M., D. Isa, V.P. Kallimani and L.H. Lee, 2006. A domain knowledge preserving in process engineering using self-organizing concept. Proceedings of the 3rd International Conference on Artificial Intelligence in Engineering and Technology, Nov. 22-24, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia, pp: 2-7.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Proceedings of 10th European Conference on Machine Learning, Chemnitz, Germany, Apr. 21-23, Springer Berlin/Heidelberg, pp: 137-142.

Kim, S.B., H.C. Rim, D.S. Yook and H.S. Lim, 2002. Effective methods for improving naïve bayes text classifiers. Lecture Notes Comput. Sci., 2417: 479-484.

McCallum, A. and K. Nigam, 2003. A comparison of event models for naïve bayes bayes text classification. J. Machine Learning Res., 3: 1265-1287.

Mohammadi, M. and G.B. Gharehpetian, 2008. Power system on-line static security assessment by using multi-class support vector machines. J. Applied Sci., 8: 2226-2233.

Nigam, K., J. Lafferty and A. McCallum, 1999. Using maximum entropy for text classification. Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering, August 1999, Stockholm, Sweden, pp: 61-67.

O'Brien, C. and C. Vogel, 2002. Spam filters: Bayes vs chisquared; letters vs words. Proceedings of the 1st International Symposium on Information and Communication Technologies. http://www.scss.tcd.ie /publications/tech-reports/reports.03/TCD-CS-2003-13.pdf.

Provost, J., 1999. Naïve-bayes vs rule-learning in classification of E-mail. Department of Computer Science. The University of Austin.

Rish, I., J. Hellerstein and J. Thathachar, 2001. An analysis of data characteristics that affect naïve bayes performance. IBM T.J. Watson Research Center 30 Saw Mill River Road, Hawthorne, NY 10532, USA.

Sahami, M., S. Dumais, D. Heckerman and E. Horvitz, 1998. A bayesian approach to filtering junk E-mail. Proceedings of the AAAI-98 Workshop on Learning for Text Categorization. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.1254&rep=rep1&type=pdf.

Sani, M.M., K.A. Ishak and S.A. Samad, 2009. Classification using adaptive multiscale retinex and support vector machine for face recognition system. J. Applied Sci., 10: 506-511.

Soltanizadeh, H. and B.S. Shahriar, 2008. Feature extraction and classification of objects in the rosette pattern using component analysis and neural network. J. Applied Sci., 8: 4088-4096.

Su, X., 2002. A text categorization perspective for ontology mapping. Department of Computer and Information Science, Norwegian University of Science and Technology, Norway.