Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 15(1): 53-61, Mar. 2017 **Regular paper**

New Feature Selection Method for Text Categorization

Xingfeng Wang[1] and **Hee-Cheol Kim**[2*], ***Member, KIICE***

[1]Information Engineering College, Eastern Liaoning University, Dandong 118000, China
[2]Department of Computer Engineering/Institute of Digital Anti-Aging Healthcare (IDA), Inje University, Gimhae 50834, Korea

Abstract

The preferred feature selection methods for text classification are filter-based. In a common filter-based feature selection scheme, unique scores are assigned to features; then, these features are sorted according to their scores. The last step is to add the top-N features to the feature set. In this paper, we propose an improved global feature selection scheme wherein its last step is modified to obtain a more representative feature set. The proposed method aims to improve the classification performance of global feature selection methods by creating a feature set representing all classes almost equally. For this purpose, a local feature selection method is used in the proposed method to label features according to their discriminative power on classes; these labels are used while producing the feature sets. Experimental results obtained using the well-known 20 Newsgroups and Reuters-21578 datasets with the k-nearest neighbor algorithm and a support vector machine indicate that the proposed method improves the classification performance in terms of a widely known metric (F_1).

Index Terms: Global feature selection, Local feature selection, Pattern recognition, Text classification

I. INTRODUCTION

The rapid development of Internet technologies has led to an increase in the number of electronic documents worldwide. Consequently, a hierarchical organization of these documents has become necessary. This situation has in turn increased the importance of text classification, the goal of which is to classify texts into appropriate classes according to their contents. Text classification is applied to numerous domains such as topic detection [1], spam e-mail filtering [2], SMS spam filtering [3], author identification [4], web page classification [5], and sentiment analysis [6]. Text classification tasks can be realized with schemes having different settings. A fundamental text classification scheme, as in many different pattern recognition problems, consists of the feature extraction and classification stages. Because of the nature of the problem, the feature extraction mechanism needs to extract numerical information from raw text documents. Then, any classifier can be used for finalizing the text classification process by predicting the document label. However, preprocessing and feature selection are known as very important stages besides feature extraction and classification. Researchers in this field are still working on enhancing the performance of text classification by incorporating various preprocessing [7], feature extraction [8], feature selection [9], and classification [10, 11] methods.

Although there exist some recent studies on the improvement of feature extraction with the contribution of Wikipedia or similar resources, the bag-of-words approach [12] is the commonly used technique for the feature extraction stage. In this approach, the orders of terms are neglected and text documents are represented with weighted frequencies (i.e., term frequency-inverse document frequency

Received 09 March 2017, **Revised** 10 March 2017, **Accepted** 17 March 2017
***Corresponding Author** Hee-Cheol Kim (E-mail: heeki@inje.ac.kr, Tel: +82-55-320-3720)
Department of Computer Engineering/Institute of Digital Anti-Aging Healthcare (IDA), Inje University, 197, Inje-ro, Gimhae 50834, Korea.

Open Access **http://doi.org/10.6109/jicce.2017.15.1.53** print ISSN: 2234-8255 online ISSN: 2234-8883

[TF-IDF]) of the unique terms in the collection. As each unique term is used in the construction of the feature set, even a collection including a small number of documents may be expressed with thousands of features. An excessive number of features may have negative effects on both the classification accuracy and the computational time. Therefore, most of the researchers are concerned with the feature selection stage in order to overcome such negative effects.

Feature selection techniques are generally categorized as filters, wrappers, and embedded methods. While wrappers and embedded methods require a frequent classifier interaction in their flow, filters do not need any classifier interaction during the construction of the feature set. The requirement of a classifier interaction may increase the run time and adapt the feature selection method to a specific learning model. Therefore, filter-based methods are preferred to wrappers and embedded methods.

Filter-based methods can be divided into two categories (global and local), depending on whether they assign a unique score or multiple class-based scores for any feature. In the case of local feature selection methods, a globalization policy is necessary to convert multiple local scores into a unique global score. On the other hand, in the case of global feature selection methods, the scores can be directly used for feature ranking. Features are ranked in the descending order, and the top-N features are included in the feature set, where N is usually an empirically determined number. Some examples of global feature selection methods for text classification are document frequency [13], information gain [14], improved Gini index [15], and distinguishing feature selector [16]. Another categorization of the characteristics of filter-based feature selection methods is whether they are one-sided or two-sided [17]. In one-sided metrics, while features indicating membership to classes have a score greater than or equal to 0, features indicating non-membership to classes have a score less than 0. As features are ranked in the descending order and the features having the highest scores are included in the feature set, negative features are not used in case there is no candidate positive feature. However, scores of two-sided methods are greater than or equal to 0. They implicitly combine positive and negative features, which indicate the membership and non-membership to any class, respectively. In this case, considering the one-against-all strategy in feature selection, positive features attain higher scores than negative ones. Thus, negative features are rarely added to the feature set in two-sided metrics.

In spite of numerous approaches reported in the literature, feature selection for text classification is still an ongoing research topic. In this paper, inspired by some of the abovementioned studies, we propose a new method that has some similarities with the characteristics of the other approaches discussed thus far in the literature. These similarities can be listed as a hybrid approach combining the power of two feature selection methods, benefiting from the power of negative features and proposing a generic solution for all the filter-based global feature selection methods. This method aims to improve the classification performance of the global feature selection method by creating a feature set representing all classes nearly equally. For this purpose, a one-sided local feature selection method assigns a class label to each feature with a positive or negative membership degree. Therefore, positive and negative features mentioned in the previous works are used as a part of the new method. The odds ratio was employed as the one-sided local feature selection method during experiments. Instead of adding the top-N features having the highest global feature selection scores to the feature set, an equal number of features representing each class equally with a certain membership and non-membership degree were included in the final feature set. In order to analyze the classification performance, two common metrics for text classification were employed in the experiments.

The rest of this paper is organized as follows: The feature selection methods used in this study are briefly described in Section II. Section III introduces the details of the proposed method. In Section IV, the classifiers used in the experiments are explained in detail. Section V presents the experimental study and accuracy results for each dataset, classifier, and success measure. Finally, some concluding remarks are presented in Section VI.

II. FEATURE SELECTION METHODS

Global feature selection methods and one-sided local feature selection methods are within the scope of this study. As was pointed out in the previous section, some of the widely known global feature selection methods are document frequency, information gain, Gini index, and distinguishing feature selector. Document frequency is not a part of this study because it does not seem to be as successful as the other methods. Odds ratio and correlation coefficient can be listed in the category of one-sided local feature selection methods. In this study, the odds ratio was utilized as it produces an excessive number of negative features. Therefore, the efficacy of the proposed method was assessed using the information gain, Gini index, and distinguishing feature selector. Mathematical backgrounds of the existing feature selection methods used in this study are provided in the following subsections.

A. Information Gain Ratio

In decision tree learning, IG is a ratio of the information

gain to the intrinsic information. It is used for reducing a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute. Information gain is also known as mutual information and is a global feature selection metric. It produces only one score for any term *t*, and this score is calculated as follows:

$$IG(t) = -\sum_{i=1}^{M} P(C_i)log + P(t)\sum_{i=1}^{M} P(C_i|t)logP(C_i|t) + P(\bar{t})\sum_{i=1}^{M} P(C_i|\bar{t})logP(C_i|\bar{t}) \ , \quad (1)$$

where *M* denotes the number of classes, $P(C_i)$ represents the probability of class C_i, $P(t)$ and $P(\bar{t})$ indicate the probabilities of presence and absence of term *t*, respectively, and $P(C_i|t)$ and $P(C_i|\bar{t})$ refer to the conditional probabilities of class C_i given the presence and absence of term *t*, respectively.

B. Gini Coefficient

GI (Gini coefficient, sometimes called the Gini index) measures the inequality among the values of a frequency distribution. A GI of zero expresses perfect equality, where all values are the same. A GI of 1 (or 100%) expresses maximal inequality among values. However, a value greater than 1 may occur if some persons represent a negative contribution to the total. GI can be calculated as follows:

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2 P(C_i|t)^2, \quad (2)$$

where $P(t|C_i)$ denotes the probability of term *t* given the presence of class C_i, and $P(C_i|t)$ represents the probability of class C_i given the presence of term *t*.

C. Distinguishing Feature Selector

Distinguishing feature selector (DFS) is one of the recent successful feature selection methods for text classification and is a global feature selection metric. The idea behind DFS is to select distinctive features while eliminating uninformative ones considering some pre-determined criteria. DFS can be expressed as follows:

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i)+P(t|\bar{C}_i)+1} \ , \quad (3)$$

where *M* denotes the number of classes, $P(C_i|t)$ represents the conditional probability of class C_i given the presence of term *t*, $P(\bar{t}|C_i)$ indicates the conditional probability of the absence of term *t* given class C_i, and $P(t|\bar{C}_i)$ refers to the conditional probability of term *t* given all the classes except C_i.

D. Odds Radio

In statistics, the odds ratio (OR) is one of the three main ways to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population. If each individual in a population either does or does not have property A, and either does or does not have property B, where both properties are appropriately defined, then a ratio can be formed. In other words, this ratio quantitatively describes the association between the presence/absence of A and the presence/absence of B for individuals in the population. OR can be calculated as follows:

$$OR(t, C_i) = log\frac{P(t|C_i)[1-P(t|\bar{C}_i)]}{[1-P(t|C_i)]P(t|\bar{C}_i)} \ , \quad (4)$$

where $P(t|C_i)$ denotes the probability of term *t* given the presence of class C_i, and $P(t|\bar{C}_i)$ represents the conditional probability of term *t* given all the classes except C_i. In this study, a simple smoothing method was applied to avoid division by zero errors and prevent the nominator from becoming zero. These situations are valuable as they represent maximum membership and non-membership. Therefore, a small value such as 0.01 was added to both the nominator and the denominator of the fraction.

III. PROPOSED METHOD

In a classical global feature selection scheme for text classification, first, feature selection scores indicating the discriminatory powers of all terms in a given collection are calculated. Then, these terms are sorted according to their feature selection scores in the descending order. After this sorting process, the top-N features are included in the feature set as a final step, where N is usually an empirically determined number. The proposed method aims to improve the classification performance by modifying the above-mentioned global feature selection process. Therefore, a one-sided local feature selection method is integrated into the feature selection process in addition to an existing global feature selection method. Thus, the proposed method can be considered an ensemble method where the power of a global feature selection method and a one-sided local feature selection method are combined in a different manner. The flow of the proposed method is as follows:

Stage 1: Feature labeling

1) Calculate the one-sided local feature selection scores of features for each class.

2) Create a label set for features including the m*2 class labels, where m is the number of classes. While the first set of m class labels represent membership, the second set of m

labels represent non-membership to these classes.

3) For each feature, determine the highest local feature selection score with respect to its absolute value and assign the associated class label from the label set to the feature.

Stage 2: Common global feature selection process

1) Calculate the feature selection scores for features using one of the global feature selection metrics.

2) Sort the features in the descending order according to the scores.

Stage 3: Construction of the new feature set

1) Suppose that the size of the final feature set was given as fs and a set of negative feature ratios was determined as nfrs. The values in nfrs may change from 0 to 1 with a specified predetermined interval such as 0.1.

2) Iterate over the sorted list obtained in the previous stage and put the appropriate features in the final feature set ffs. Make the ffs equally representative for each class by using the feature labels determined in stage 1. At the end of this stage, ffs must contain an equal number of features for each class considering a specific negative membership ratio value nfr inside nfrs.

Stage 4: Conditional part

If the number of features in ffs is less than fs, finalize the feature selection process by adding the missing number of disregarded features having the highest global feature selection scores to ffs.

Thus, in the worst case, all features need to be traversed once and some of them may be traversed twice during the construction of the candidate feature set. Apart from the explanations above, a sample collection is provided in Table 1 to illustrate how the proposed method works.

The sample collection contains six documents consisting of 11 distinct terms. The features sorted according to their GI values and their corresponding OR scores are presented in Table 2.

Table 1. Sample collection

Document name	Content	Class
Doc 1	Apple, orange, banana,	C1
Doc 2	Apple, orange, watermelon, lemon	C2
Doc 3	Apple, orange, watermelon, strawberry, lemon	C2
Doc 4	Tomato, pineapple, peach, lemon, cherry	C3
Doc 5	Tomato, pineapple, lemon, cherry	C3
Doc 6	Tomato, pineapple, pear, lemon, strawberry	C3

In Table 2, feature labels and their associated membership degrees are also given. We used Eqs. (1), (2), and (3) to compute the IG, GI, and DFS scores of each feature, respectively, and ranked the features according to the GI scores. Then, we used Eq. (4) to compute the OR scores of each feature's classes. We selected the maximum absolute value of each OR score as the OR label. Then, we compared each feature with its OR label. If this feature belonged to the corresponding class, we considered this feature to be positive. Otherwise, we considered this feature to be negative.

Note that if the final feature set contains the same number of positive features and negative features, the set is considered to be well balanced.

Therefore, we selected the same number of positive and negative features as far as possible. In other words, we first selected different class features from the positive feature set and then, different class features from the negative feature set. We tried our best to select the same number of positive and negative features in order to obtain as balanced a set as possible as the final feature set to achieve better evaluation results.

Table 3 presents the feature sets obtained by using the GI method and the GI + the proposed method.

The size of the feature set was determined to be 6 because 6 was half of the raw feature number. For the GI method, we

Table 2. Feature selection scores and membership degrees

Feature	IG scores	GI scores	DFS scores	OR scores (C1, C2, C3)	OR label	Positive/negative
Tomato	0.6931	1	1	−4.0943, −4.3175, 4.6052	C3	Positive
Pineapple	0.6931	1	1	−4.0943, −4.3175, 4.6052	C3	Positive
Watermelon	0.6365	1	1	−3.6889, 4.6052, −4.1997	C2	Positive
Banana	0.4506	1	1	4.6052, −3.2189, −3.5066	C1	Positive
Orange	0.6931	0.5556	0.7714	4.0943, 4.3175, −4.6052	C3	Negative
Apple	0.6931	0.5556	0.7714	4.0943, 4.3175, −4.6052	C3	Negative
Lemon	0.4506	0.5200	0.5886	−4.6052, 3.2189, 3.5066	C1	Negative
Cherry	0.3183	0.4444	0.75	−3.6889, −3.9120, 4.1997	C3	Positive
Peach	0.1323	0.1111	0.6	−2.9957, −3.2189, 3.5066	C3	Positive
Pear	0.1323	0.1111	0.6	−2.9957, −3.2189, 3.5066	C3	Positive
Strawberry	0.0872	0.0903	0.5557	−3.6889, 1.0986, 0	C1	Negative

Table 3. Final feature sets obtained with two different methods

Method	Final feature set	Distributions of class labels
GI	Tomato, pineapple, watermelon, banana, orange, apple	C1(1), C2(1), C3(4)
GI + proposed method	Tomato, watermelon, banana, orange, lemon	C1(2), C2(1), C3(2)

only considered the GI score; we selected the maximum 6 GI scores of the feature. However, in the proposed method, we first selected the feature of each class from the positive features according to descending order of the GI score, and then, we selected the feature of each class from the negative features depending on the descending order of the GI score. The proposed method showed better distributions of the class labels.

In this sample scenario, two main points draw attention to global feature selection methods. The first is that the classes may not be represented almost equally in the final feature set. According to the sample scenario, while 6 features having higher GI scores are selected, each class is represented with 1, 2, or 4 features. The second point is that most of the feature selection methods do not consider negative features. The term 'banana' representing membership to class C1 was added to the feature set. On the other hand, the term 'lemon,' which is a good indicator of non-membership to class C1 was not included in the feature set. However, if we analyze the discriminative power of the terms 'banana' and 'lemon' manually, we can infer that they have similar discriminative powers with respect to C1. According to the GI score, the term 'banana' is nearly twice as important as 'lemon.' As some studies in the literature state, negative features are also valuable and a portion of the negative feature set must be included in the final feature set. In the proposed method, both classes are represented almost equally and negative features such as the term 'lemon' can be added to the final feature set. The sample collection and the final feature sets are provided to show how the proposed method works.

IV. CLASSIFICATION ALGORITHMS

In order to prove the efficacy of the proposed method, it was necessary to employ the classifiers commonly used for text classification research in the literature and proven to be significantly successful. For this purpose, k-nearest neighbor (kNN) classifiers and support vector machine (SVM) classifiers were utilized. A brief explanation of these methods is given in the following subsections.

A. k-Nearest Neighbor Classifiers

The kNN method was first described in the early 1950s. The method is labor intensive when given large training sets and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition.

The nearest neighbor classifiers are based on learning by analogy, i.e., by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. Thus, all the training tuples are stored in an n-dimensional pattern space. When given an unknown tuple, a kNN classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k 'nearest neighbors' of the unknown tuple.

'Closeness' is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, $X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, is as follows:

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \quad . \tag{5}$$

A good value for k, the number of nearest neighbors, can be determined experimentally. Starting with k = 1, we use a test set to estimate the error rate of the classifier.

This process can be repeated each time by incrementing k to allow for one more neighbor. The k value that gives the minimum error rate may be selected. In general, the larger the number of training tuples is, the larger is the value of k. As the number of training tuples approaches infinity and k = 1, the error rate can be no worse than twice the Bayes error rate. If k also approaches infinity, the error rate approaches the Bayes error rate.

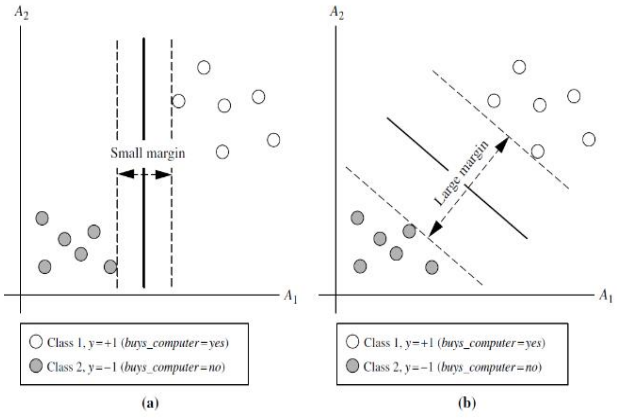

Fig. 1. Two possible separating hyperplanes and their associated margins. The one with the larger margin (b) should have greater generalization accuracy.

B. Support Vector Machines

SVMs, a method for the classification of both linear and nonlinear data. In a nutshell, an SVM is an algorithm that works as follows: it uses nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane. With appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane by using support vectors and margins, as shown in Fig. 1.

V. EXPERIMENTAL WORK

In this section, an in-depth investigation was carried out to measure the performance of the proposed method against the individual performance of the three global feature selection methods. While the one-sided local feature selection method utilized in the flow of the proposed method was OR, the global feature selection methods employed in the experiments were IG, GI, and DFS. Note that stop-word removal and stemming [18] were used as the two pre-processing steps in addition to weighting terms with TF-IDF. In order to validate the performance of the proposed method, two different datasets with varying characteristics and success measures were utilized to observe the effectiveness of the proposed method under different circumstances. In the following subsections, the utilized datasets and success measures are briefly described. Then, the characteristics of the feature sets produced by the global feature selection methods are analyzed to show that the classes are not equally represented in the feature set. Finally, the experimental results are presented.

Table 4. Reuters-21578 dataset (top 8 classes)

Class	Training docs	Test docs	Total docs
acq	1596	696	2292
crude	253	121	374
earn	2840	1083	3923
grain	41	10	51
interest	190	81	271
money-fx	206	87	293
ship	108	36	144
trade	251	75	326
Total	5485	2189	7674

A. Datasets

In this study, three distinct datasets with varying characteristics were used for the assessment. The first dataset consists of the top 8 classes of Reuters-21578 [19]. The second dataset is another popular benchmark collection, namely 20 Newsgroups [20], we selected the top 10 classes to use. Both these datasets are widely used benchmark collections for text classification. Detailed information regarding these datasets is provided in Tables 4 and 5.

Table 5. 20 Newsgroups dataset (top 10 classes)

Class	Training docs	Test docs	Total docs
alt.atheism	480	319	799
comp.graphics	584	389	973
comp.os.ms-windows.misc	572	394	966
comp.sys.ibm.pc.hardware	590	392	982
comp.sys.mac.hardware	578	385	963
comp.windows.x	593	392	985
misc.forsale	585	390	975
rec.autos	594	395	989
rec.motorcycles	598	398	996
rec.sport.baseball	597	397	994
Total	5771	3851	9622

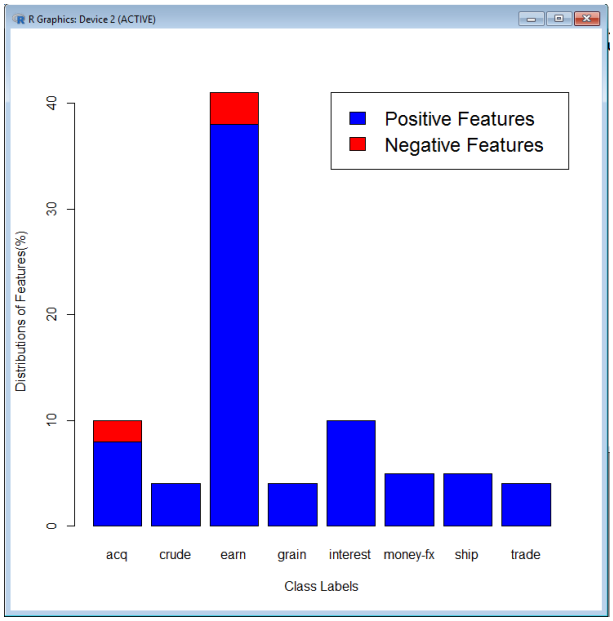

Fig. 2. Reuters-21578: class distribution of features.

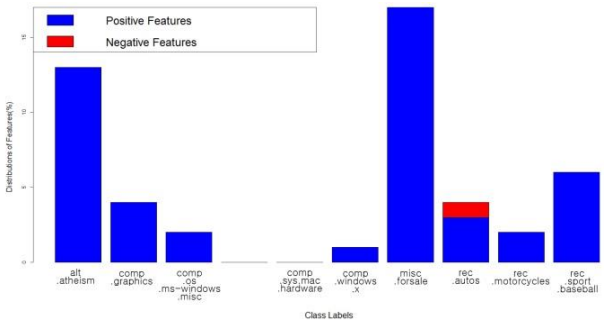

Fig. 3. 20 Newsgroups: class distribution of features.

B. Success Measures

In a statistical analysis of the binary classification, the F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score. Here, p denotes the number of correct positive results divided by the number of all positive results, and r represents the number of correct positive results divided by the number of positive results that should have been returned. The F_1 score can be interpreted as a weighted average of the precision and recall, where the score reaches its best value at 1 and worst at 0.

The traditional F-measure or balanced F-score (F_1 score) is the harmonic mean of precision and recall, multiplying with the constant 2 scales the score to 1 when both the recall and the precision are 1:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \ . \quad (6)$$

C. Analysis of the Feature Sets Produced by Global Feature Selection Methods

As pointed out in Section 3, the feature sets constructed using the global feature methods may not represent all classes almost equally. In this part, the distribution of features to classes is analyzed for the benchmark datasets. The profiles of feature sets obtained from the Reuters-21578 and 20 Newsgroup datasets are illustrated in Figs. 2 and 3. The labels that OR assigns to the features are used for this analysis. The numbers of positive and negative features, which show the membership and non-membership to classes, respectively, are also presented in the figures. Term-document matrices tend to get very big for normal-sized datasets. Therefore, we provide a method to remove sparse terms, i.e., terms occurring only in very few documents. Normally, this reduces the matrix dramatically without losing significant relations inherent to the matrix. Figs. 2 and 3 show the class distributions of features obtained from the considered datasets. It is clear that negative features are considerably fewer than positive features; therefore, we need to select as many negative features as possible for the final feature set.

D. Accuracy Analysis

In this section, the individual performance of the global feature selection methods and the proposed method were compared. This comparison was carried out according to the maximum micro-F_1 values that these methods achieved. Some numbers of features, selected by each selection method, were fed into the kNN and SVM classifiers. Tables 6 and 7 show the micro-F_1 scores obtained on two different datasets with these two classifiers.

According to Tables 6 and 7, the proposed method surpasses the individual performance of three different global feature selection methods in terms of micro-F_1. The improvement on the Reuters-21578 dataset seems to be more impressive than that on the other dataset. Further, the relatively low improvement on the 20 Newsgroups dataset could be attributed to its structure. As pointed out in the previous subsection, it is a more balanced dataset than the other and thus, may be more effective. Further, the SVM classifier shows better improvement than the kNN classifier; therefore, we can conclude that SVM is a better classifier in this case. Moreover, the GI and DFS scores are higher than the IG scores; hence, we can infer that GI and DFS are the better global feature selection methods.

Table 6. Micro-F_1 scores obtained using kNN and SVM on the Reuters-21578 dataset

Method	Micro-F_1 (%)	
	kNN	SVM
IG	36.094	46.412
IG + proposed method	59.083	64.390
GI	65.836	72.881
GI + proposed method	70.451	73.267
DFS	65.984	66.070
DFS + proposed method	66.243	71.589

Table 7. Micro-F_1 scores obtained using KNN and SVM on the 20 Newsgroups dataset

Method	Micro-F_1 (%)	
	kNN	SVM
IG	40.382	43.075
IG + proposed method	54.623	60.246
GI	58.522	68.267
GI + proposed method	62.329	70.442
DFS	57.567	60.684
DFS + proposed method	65.594	69.268

VI. CONCLUSIONS

The main contribution of this study is the introduction of an improved global feature selection scheme for text classification. The proposed method is a generic solution for all the filter-based global feature selection methods unlike most of the other approaches in the literature. As pointed out before, most of the previous studies are focused on providing some improvements on specific feature selection methods rather than providing a new generic scheme. The proposed method is an ensemble method that combines the power of a filter-based global feature selection method and a one-sided local feature selection method. The idea behind the proposed method is to make the feature set represent each class in the dataset almost equally. For this purpose, efficient feature ranking skills of global feature selection methods were combined with the class membership and non-membership detection capability of the one-sided local feature selection methods in a different manner. Using well-known benchmark datasets, classification algorithms, and success measures, we investigated the effectiveness of the proposed method and compared it against the individual performance of filter-based global feature selection methods. The results of a thorough experimental analysis clearly indicate that the proposed method improved the classification performance in terms of micro-F_1.

Despite its significant contribution, the proposed scheme has some limitations. Firstly, before we compute each feature's score, we need to reduce the feature's number to an acceptable range; this may decrease the capability of the proposed method. Secondly, there are considerably fewer negative features than positive features in our final dataset; this makes the final set unbalanced. Finally, for different datasets, the proposed method has different performance and hence, this method cannot be considered to be very stable.

Based on the limitations of this study and the computational results, some potential directions for future research might be proposed. For example, heuristic approaches may be integrated into the proposed method in order to detect a more appropriate ratio for negative features. Correspondingly, the impact of using a varying negative feature ratio for classes may be examined. Apart from these, the integration of other global feature selection methods in the literature into the proposed method and the ratio of the probable performance improvement still remain as an interesting future work.

REFERENCES

[1] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, "PoliTwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 24-33, 2014.

[2] A. S. Ghareb, A. B. Bakar, and A. R. Hamdan, "Hybrid feature selection based on enhanced genetic algorithm for text categorization," *Expert Systems with Applications*, vol. 49, pp. 31-47, 2016.

[3] H. Elghazel, A. Aussem, O. Gharroudi, and W. Saadaoui, "Ensemble multi-label text categorization based on rotation forest and latent semantic indexing," *Expert Systems with Applications*, vol. 57, pp. 1-11, 2016.

[4] Y. Wang, Y. Liu, L. Feng, and X. Zhu, "Novel feature selection method based on harmony search for email classification," *Knowledge-Based Systems*, vol. 73, pp. 311-323, 2015.

[5] J. Yang, Z. Liu, and Z. Qu, "A novel feature selection based gravitation for text categorization," *International Journal of Database Theory and Application*, vol. 9, pp. 211-228, 2016.

[6] W. Medhat, A. Hassan, and H. Koashy, "Sentiment analysis algorithms and applications: a survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, 2014.

[7] M. Hadni, S. E. A. Ouatik, & A. Lachkar, "Word sense disambiguation for Arabic text categorization," *International Arab Journal of Information Technology*, vol. 13, no. 1A, pp. 215-222, 2016.

[8] A. H. Mohammad, T. Alwada'n, and O. Al-Momani, "Arabic text categorization using support vector machine, Naïve Bayes and Neural Network," *GSTF Journal on Computing*, vol. 5, no. 1, pp. 108-115, 2016.

[9] S. Gunal, "Hybrid feature selection for text classification," *Turkish Journal of Electrical Engineering Computer Sciences*, vol. 20, pp. 1296-1311, 2012.

[10] J. Yang, Z. Liu, and Z. Qu, "Text representation based on key terms of document for text categorization," *International Journal of Database Theory and Application*, vol. 9, no. 4, pp. 1-22, 2016.

[11] W. Zong, F. Wu, L. K. Chu, and D. Sculli, "A discriminative and semantic feature selection method for text categorization," *International Journal of Production Economics*, vol. 165, pp. 215-222, 2015.

[12] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in Naïve Bayes for text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2508-2521, 2016.

[13] W. Yang, Y. Fu, and D. Zhang, "an improved parallel algorithm for text categorization," in *Proceedings of International Symposium on Computer, Consumer and Control (IS3C)*, Xi'an, China, pp. 451-454, 2016.

[14] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article no. 27, 2011.

[15] I. Idris and A. Selamat, "Improved email spam detection model with negative selection algorithm and particle swarm optimization," *Applied Soft Computing*, vol. 22, pp. 11-27, 2014.

[16] L. Jiang, Z. Cai, H. Zhang, and D. Wang, "Naïve Bayes text classifiers: a locally weighted learning approach," *Journal of Experimental Theoretical Artificial Intelligence*, vol. 25, no. 2, pp.

273-286, 2013.

[17] H. Ogura, H. Amano, and M. Kondo, "Comparison of metrics for feature selection in imbalanced text classification," *Expert Systems with Applications*, vol. 38, no. 5, pp. 4978-4989, 2011.

[18] A. Pietramala, V. L. Policicchio, and P. Rullo, "Automatic filtering of valuable features for text categorization," in *Advanced Data Mining and Applications*. Heidelberg: Springer, pp. 284-295, 2012.

[19] R. H. Pinheiro, G. D. Cavalcanti, R. F. Correa, and T. I. Ren, "A global-ranking local feature selection method for text categorization," *Expert Systems with Applications*, vol. 39, no. 17, pp. 12851-12857, 2012.

[20] R. H. Pinheiro, G. D. Cavalcanti, and T. I. Ren, "Data-driven global-ranking local feature selection methods for text categorization," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1941-1949, 2015.

Xingfeng Wang

received his B.E. in Application of Electronic Technology from Liaoning Normal University, China, in 1998, and his M.E. in Computer Technology from Dalian University of Technology, China, in 2009. Currently, he teaches at Information Engineering College, Eastern Liaoning University, China. His research interests are in the areas of data mining, neural network, and deep learning. He has published more than 10 papers in these areas.

Hee-Cheol Kim

received his M.S. in Computer Science from Sogang University, Korea, in 1991, and his Ph.D. in C omputer Science from Stockholm University, Sweden, in 2001. He is currently a professor at Depart ment of Computer Engineering, Inje University, Korea. His research interests are in the areas of hu man computer Interaction, software engineering, and u-healthcare. He has published more than 100 papers in these areas.