Advances in Computer Science Research, volume 92
International Conference on Mathematics, Big Data Analysis and Simulation and Modeling (MBDASM 2019)

A Novel Feature Selection Method Based on Category Distribution Ratio in Text Classification

Pujian Zong[1] and Jian Bian[1,2,*]
[1]School of Data and Computer Science Sun Yat-Sen University, China
[2]Guangdong Province Key Laboratory of Computational Science Guangdong, China
[*]Corresponding author

***Abstract*—In text classification, texts are represented as a high-dimensional and sparse matrix, whose dimension is the same as the total number of terms of all texts. Using all terms for text classification tasks will affect the accuracy and efficiency. Feature selection algorithm can select some features most relevant to text category and reduce the dimension of text representation vector. In this paper, we propose a new feature ranking metric as category distribution ratio (CDR) which takes the true positive rate and false positive rate and their difference of a term into account while estimating the significance of a term. To prove the effectiveness of the proposed feature selection algorithm, we compare its performance against six metrics (balanced accuracy measure (ACC), odds ratio (OR), Gini index (GI), max-min Ratio (MMR), normalized difference measure(NDM)，chi - square (CHI)) on three benchmark data sets (20newsgropus, Ohsumed, Reuters 21578) using multinomial naive Bayes, support vector machines and k-nearest neighbor classifiers. The experimental results show that the classification evaluation index macro F1 based on CDR feature selection is higher than the other six algorithms.**

Keywords—text classification; feature selection; feature reduce

I. INTRODUCTION

Text categorization predicts the category of documents based on the content of the document [1]. Documents with similar contents are assigned to the same category for easy management and search. Text classification is an important method for people to manage and use text data effectively. It has been used in many applications such as SMS spam filtering [2], spam e-mail filtering [3], sentiment analysis [4], topic detection [5], document classification [6]. As one of the technical bases of natural language processing, text categorization has a wide application prospect. However, with the rapid growth of data, the management and utilization of text information has become a huge challenge. Time-consuming and laborious manual management can't meet the demand of information society for current data management. Therefore, automatic text categorization technology which can effectively manage and quickly locate large-scale data information has been widely used in various fields and achieved rapid development in recent years [7-8].

With the growth of text data, the dimensions of text representation vector become larger and the vector is sparse. Since many dimensions do not contain useful classification information, they may only contain information irrelevant to classification, or redundant information. The existence of such term is equivalent to noise, which only reduces accuracy of classification. Feature selection is an indispensable pre-processing step in text classification, which can significantly reduce the text dimension, remove redundant terms and irrelevant terms that have adverse effects on the classification effect from the original vector space, and retain only the terms with strong relevance to categories [9-11].

II. RELATED WORKS

A. Overview of Feature Selection Methods for Text Classification

According to the combination of feature selection algorithm and classifier, feature selection algorithm can be divided into three categories: filters, wrappers and embedded methods [14]. Embedded, is to embed feature selection algorithm in the classifier algorithm [15]; Filter method uses an independent evaluation function to calculate the importance of every item [16]; Wrappers method uses a certain classifier algorithm as an evaluation metric to select a feature subset [17].

Filter methods are widely used due to fast calculation, low storage cost and good classification effect. Some well-known filtering methods include chi-square (CHI) [18], mutual information (MI) [19], information gain (IG) [20], odds ratio (OR) [21], etc. In addition, some new feature selection algorithms have emerged in recent years, such as the Measure using Poisson distribution [22], the bi-test based on binomial hypothesis Test [23], and the CMFS based on comprehensive measures [24].

One of the challenges is that the feature selection algorithm need consider the dependency of the feature, but calculation of feature selection can't become time-consuming. Therefore, two-stage algorithms have been proposed recently. Uguz uses IG to select some relevant features from the original feature space, and then uses the genetic algorithm or principal component analysis to eliminate redundant features from subsets in the second stage [25]. Similarly, the combination of distinguishing feature selection (DFS) and latent semantic index (LSI) in two stages [26]. More recently, a multivariate algorithm called MRDC was proposed. It selects features with maximal-relevancy in first stage and removes redundant features by minimal-redundancy in second stage [27].

Copyright © 2019, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Another challenge is because of the high imbalance or skew between text data categories, which will result in feature selection selects more features related to larger classes but less discriminating features for smaller classes [28]. Uysal proposes an improved global feature selection scheme (IGFSS) that marks features based on their ability to distinguish between classes and uses these marks when generating feature sets [29]. In 2018, Rehman proposed the MMR (Max-Min Ratio) algorithm, which performs feature selection based on the product of max-min ratio of the true positives and false positives and their difference [30].

B. Filter Methods for Feature Selection

Most of the feature ranking algorithms are based on document frequency. There are four document frequencies for a term. Ture positive (*tp*) is the number of documents containing the term and belong to the positive class. True negative (*tn*) is the number of documents that containing the term and do not belong to the positive class. False positive (*fp*) is the number of documents that containing the term and belong to the negative class. False negative (*fn*) is the number of documents that do not belong to the negative class and do not contain the term.

ACC is the simplest metric to estimate the value of a term. The mathematical formula is as follows:

$$ACC = tp - fp \tag{1}$$

Since ACC only considers the difference in the number of documents, it is more likely to favor categories with large amounts of text. Therefore, ACC is improved to the difference of absolute values of true positive rate (tpr) and false positive rate (fpr). For a term, $tpr = tp/(tp + fn)$, $fpr = fp/(fp + tn)$.

$$ACC2 = |tpr - fpr| \tag{2}$$

NDM algorithm is improved for the deficiency of ACC2. Since ACC cannot distinguish the importance of features of the same difference between *tpr* and *fpr*. when the ACC2 values of the two features are equal, we expect the smaller $min\{tpr, fpr\}$ is, the bigger the score is. The NDM algorithm takes $min\{tpr, fpr\}$ as the denominator, and forms the following formula:

$$NDM = \frac{|tpr - fpr|}{min\{tpr, fpr\}} \tag{3}$$

However, NDM algorithm also has its limitations, that is, terms are sparse for large and highly skewed data. Terms are absent in most of document, so their discrimination power is poor but NDM will give assign higher value for them. To address this problem, Rehman propose a new metric name MMR (max-min ratio). Its Formula is as follows:

$$MMR = \frac{max\{tpr, fpr\}}{min\{tpr, fpr\}} \times |tpr - fpr| \tag{4}$$

When $min\{tpr, fpr\}$ is equal to 0, the denominator is replaced by a small value ε or $1/N$ (N is the num of all document) to avoid infinite values. The MMR algorithm optimizes the NDM algorithm and solves the problem that the NDM algorithm cannot deal with large and highly skewed dataset reasonably. The main weaknesses of MMR algorithm is that it does not solve the first problem of NDM algorithm: how to choose the denominator when $min\{tpr, fpr\}$ is 0.Usually using $1/N$ doesn't seem to go terribly wrong, but as N gets bigger, this defect becomes more apparent.

III. OUR WORKS

A. Category Distribution Ratio

The distribution of items in inter-category and intra-category has influence on classification in the following aspects:

- A term that occur frequently within a class is more important than a term that rarely appear, and the higher the frequency, the higher the score should be.
- A term occurs frequently in a single class and are often found in the other classes. Its score should be lower. Namely, we need to consider the frequency of occurrence of term between different classes.
- A term t, the difference between the true positive rate (*tpr*) and false positive rate (*fpr*) is the same as the other term t'. In this case, to distinguish the worth of t and t', it is necessary to take both *tpr* and *fpr* into account.

Base on the three point, we propose a scoring framework as:

$$CDR = \frac{\alpha}{\beta \times |tpr - fpr| + \alpha \times |1 - |tpr - fpr||} \times |tpr - fpr| \tag{5}$$

where $a = max\{tpr, fpr\}$, $\beta = min\{fpr, fpr\}$. CDR is the value of a term t to class C_i. For the whole data set, the score of term should be represented by the maximum score between term and classes:

$$CDR_{\max}(t) = \max_{i=1}^{m}\{CDR(t, C_i)\} \quad (0 \le i \le m) \tag{6}$$

tpr is the proportion of the document containing the feature t in the positive class, and *fpr* is the proportion of the document containing the feature t in the negative class. |*tpr-fpr*| is the relative document frequency. Once the scores of all terms in a data set are attained, we can select the Top-*N* terms as feature subset. The denominator of CDR never be equal to 0, so CDR can avoid the weakness of MMR.

B. Compare CDR and MMR

CDR algorithm is an improvement of NDM and MMR algorithms. Its advantage is that it inherits the advantages of ACC2, NDM, MMR:

(1) when there is a high imbalance or skew among the categories of text data, CDR can correctly select features with strong prediction without bias to larger classes.

(2) the relative document frequency (|*tpr*-*fpr*|) of the two term are the same, and CDR can also distinguish their importance.

(3) when calculating the formula, even if *tpr* or *fpr* is 0, the denominator will not be 0. So it can avoid the failure to correctly sort the features in such cases.

(4) The importance of term in the prediction of classification depends on the one class with the greatest correlation with the term, rather than the whole data set, so the selected terms are more representative (representing one class).

To compare CDR and MMR, let us consider the example shown in Table 1.

TABLE I. EXAMPLE TO SHOW ISSUES WITH MMR

Terms	+ class(100)	- class(3000)	*tpr*	*fpr*	*MMR*	*CDR*
t1	2	2800	0.02	0.93	42.3	8.305
t2	10	1	0.1	0.0003	33.3	0.1107
t3	10	0	0.1	0	100	0.1112
t4	2	1	0.02	0.0003	1.31	0.02
t5	80	2	0.8	0.00067	954.4	3.97

CDR ranking list of these five terms is {t1, t5, t3, t2, t4} and MMR ranking list is {t5, t3, t1, t2, t4}. First, we compare term t1 and t5. Term t1 is present in 93% documents of negative class, while it only occurs in 2% documents of the positive class. Term t2 is present in 80% documents of positive class while it occurs in 0.067% document of the negative class. Intuitively, t1 discriminates between the classes the best as it is present in 93% documents of negative class but only 2% in positive class. CDR assigns the highest score to t1 but MMR assigns t5 higher score than t1 due to the denominator of MMR of t1($min\{tpr,fpr\}$ = 0.00067) is too small. Then we compare t3 and t1. Term appears in 10% documents of positive classes, which is far less than t1 presenting in 93% documents of negative class. CDR correctly puts t1 ahead of t5, However, MMR puts t3 before t1, which is not practical. In addition, the scores of t2 and t3 should be close to each other, while the scores given by MMR varies greatly, while CDR is very close. All these indicate that CDR is better than MMR in ranking features when data categories are imbalanced.

IV. EMPIRICAL EVALUATION

A. Data Sets and Pre-Processing

There are three data sets we used in this experiment to verify the validity of the new feature selection algorithm. The first dataset is a popular text collection namely 20 Newsgroups (20NG), which contains18846 documents and 20 classes. The Ohsumed Corpus comes from the MEDLINE database, which is a bibliographic database of medical literature. Each document belongs to one or more of the 23 categories. Our task is single-label text classification, so only 7400 documents belonging to only one category are remained. R8 is a subset of the Reuters 21578 dataset. R8 has 8 categories and 7674 documents.

B. Data Sets Pre-processing

First, we stem the data set and remove the stop words in the documents. Then remove low frequency words with fewer than 5. Data set is split into 75% training set and 25% test set. This paper uses the vector space model (VSM) to represent text vectors and uses TF-IDF weight to represent the weight of each feature. Since all the datasets used have multiple classes, there are two strategies: one-versus-rest and one-versus-one can handle classification problems. We use the first strategy because it is more widely used.

V. RESULTS AND DISCUSSION

We compare CDR with other six feature ranking algorithms namely ACC, OR, GI, NDM, MMR and CHI over three data sets in our experiments.

A. 20newsgroups Data Set

Figure.1 compare the Micro F1 and Macro F1 value of the methods for MNB, SVM and KNN classifiers, respectively. From the results we can see that for 20newsgroups dataset, the proposed method outperforms others in all cases. CDR with each classifier attains the best performance when 500 features selected. The highest macro F1 value with SVM is 0.808 and the highest micro F1 value is 0.7923 in the same case. For MNB classifier, CDR results in the highest performance for at least 83.3% of the trials. While for KNN classifier, CDR results in the highest performance for 66.7% case, the better than SVM classifier.

B. R8 Data Set

Fig.2 is the macro F1 and micro F1 values of CDR compared with other methods over R8 data set. The experimental results show that the performance on R8 is not as good as 20NG, but the CDR is still optimal in 60% of cases for SVM classifier. And CDR obtains the highest F1 value when 10 ,20 and 50 features are selected, this indicates that the topmost part of the CDR selection is strongly predictive features. In addition, CDR often performs second best micro F1 value, only a little worse than the first one when 200 or 500 features selected, such as the macro F1 of CDR with SVM classifier is 0.9358, the best is 0.9434 by GI.

The R8 data set is a highly skewed data set. The difference between macro and micro F1values in R8 data set is easily observed. For highly skewed data sets, documents of rare classes are more easily misclassified. For example, when 10 features selected by CDR, the highest micro F1 value is 0.7759 while the macro F1 value is 0.3729 over R8 with SVM classifier. Other methods also have observed difference between macro and micro F1 values over R8. While the number of features selected is small, such as less than 200, the macro F1 values attained by CDR is higher than others. This shows that the features selected by the CDR has better performance than others in the imbalanced data set.

C. Ohsumed Data Set

From figure 3, it is clear that CDR has achieved the best results for at least 83.3% of the trials when 20-200 features are selected, such as the best micro F1 for KNN classifier is 0.518

and 0.5662 for SVM. This shows that the CDR performs much better than other methods in the case of low dimensions. When 500 features selected, CDR also result in second best performance in most trails. Ohsumed is also an imbalanced data set, and the macro F1 value of the CDR is much higher than other algorithms.

D. Discussion of Result

Table 3 counts the number of times each method achieves highest macro F1 values and micro F1 values over different datasets, respectively. CDR result in highest micro F1 value in 59.3% of the trials, win the first place. GI is at second place with 27.8% of the cases. According to the number of times each method reaches the highest macro F1 value, CDR still ranks first, accounting for 61% of the cases. The second and third place are GI and CHI with 22.2% and 12.9% of the cases respectively.

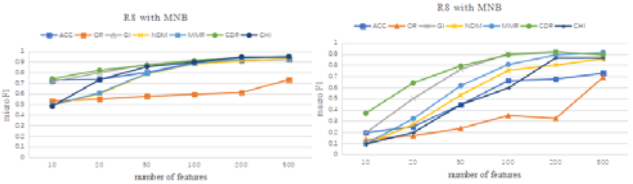
Comparing different data sets, we can find that CDR performs best on 20newsgroups dataset since CDR attains the highest F1 values 24 times. Ohsumed after 20newsgroups and R8 is least with 19 times. The R8 dataset is a highly skewed dataset that has an adverse effect on the effectiveness of all methods. Through the results, we find that the performance of CDR is the best, indicating that the imbalanced data set has the least adverse effect. Experiments have also shown that CDRs tend to have the best performance compared to other methods when the number of selected features is small. This indicates that the CDR can find relatively more relevant features.

TABLE II. SUMMARY OF DATASETS

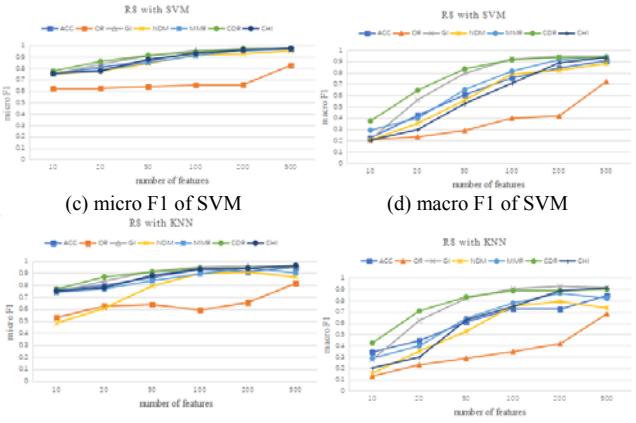
Data set	Total docs	Total term	Smallest class size	Largest class size	Class num	Doc Average Length
20NG	18846	30079	628	999	20	221.26
R8	7674	5291	51	3923	8	65.72
Oh	7400	9008	50	1175	23	153.82

TABLE III. NUMBER OF TIMES A METHOD ATTAINS HIGHEST F1 VALUES OVER DIFFERENT DATASETS WITH THREE CLASSIFIERS.

Method	Micro F1				Macro F1			
	20NG	R8	Oh	all	20NG	R8	Oh	all
ACC	0	0	0	0	0	0	0	0
OR	0	0	0	0	0	0	0	0
GI	0	4	2	6	0	7	5	12
NDM	0	0	0	0	0	0	0	0
MMR	0	1	0	1	0	2	0	2
CDR	12	10	10	32	12	9	12	33
CHI	6	3	6	15	6	0	1	7



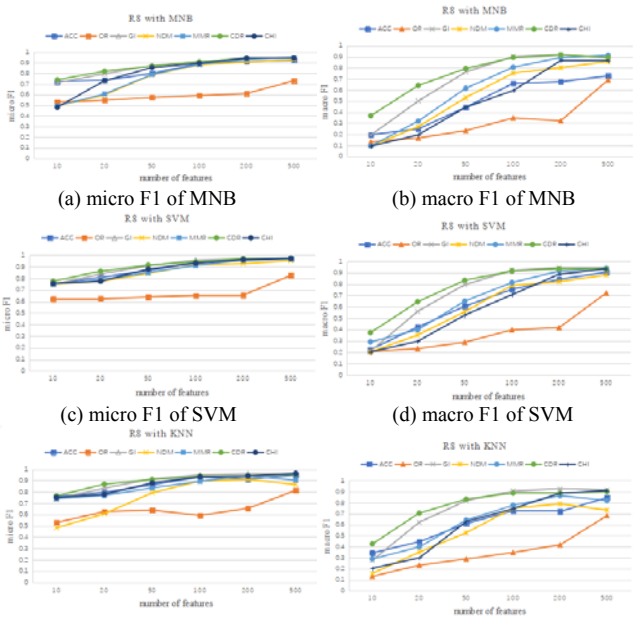
(a) micro F1 of MNB (b) macro F1 of MNB



(c) micro F1 of SVM (d) macro F1 of SVM

(e) micro F1 of KNN (f) macro F1 of KNN

FIGURE I. MICRO F1 AND MACRO OF THE METHODS OBTAINED OVER 20NEWSGROUPS WHEN DIFFERENT NUMBERS OF FEATURES ARE SELECTED FOR CLASSIFICATION USING (A,B) MNB, (C,D) SVM, AND (E,F) KNN CLASSIFIERS

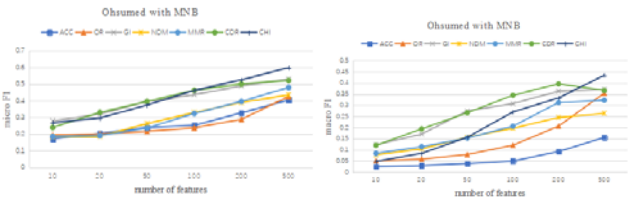


(a) micro F1 of MNB (b) macro F1 of MNB

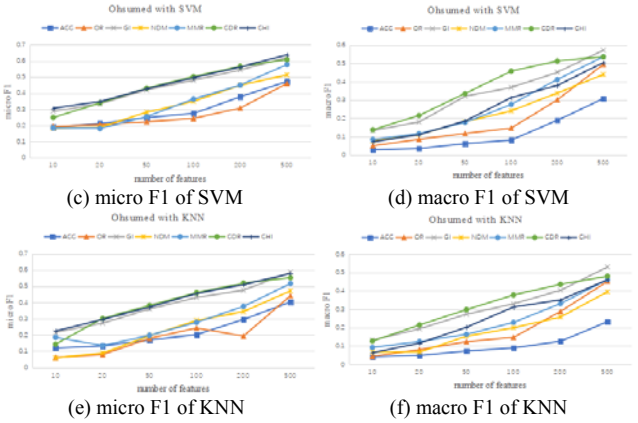
(c) micro F1 of SVM (d) macro F1 of SVM

(e) micro F1 of KNN (f) macro F1 of KNN

FIGURE II. MICRO F1 AND MACRO F1 OF THE METHODS OBTAINED OVER R8 WHEN DIFFERENT NUMBERS OF FEATURES ARE SELECTED FOR CLASSIFICATION USING (A,B) MNB, (C,D) SVM, AND (E,F) KNN CLASSIFIERS



(a) micro F1 of MNB (b) macro F1 of MNB



(c) micro F1 of SVM (d) macro F1 of SVM

(e) micro F1 of KNN (f) macro F1 of KNN

FIGURE III. MICRO F1 AND MACRO F1 OF THE METHODS OBTAINED OVER OHSUMED WHEN DIFFERENT NUMBERS OF FEATURES ARE SELECTED FOR CLASSIFICATION USING (A,B) MNB, (C,D) SVM, AND (E,F) KNN CLASSIFIERS

VI. CONCLUSIONS AND FEATURE WORKS

This paper proposes a new feature select method called Category Distribution Ratio (CDR), which considers the true positive rate of the feature and the false positive rate, and taking into account the positive and negative effects of the feature on the category. We have compared the performance of CDR with other six method namely ACC, OR, GI, NDM, MMR, CHI over 20newsgroups, Reuters(R8) and Ohsumed dataset.

we have repeated our experiments 5 times and averaged. CDR result in highest micro F1 value in 59.3% of the trials and 61.1% for macro F1. Besides, CDR attain better performance than others for nested subset with less than 200. CDR can find relatively more relevant features and rank these features in more front positions than competitors. When more terms added to subset, there are more weakly relevant terms affect the performance. So CDR performance becomes similar to others.

The CDR algorithm proposed in this paper can select the most relevant features, but it doesn't take the relationship among the selected features into account. There may be some redundant features in the feature subset, and some features are discriminated when combined with each other. We can design algorithms to eliminate redundant features in feature subsets.

ACKNOWLEDGMENT

This study was supported by Guangdong Province Science and Technology Funding Plan (2015A030401004) and Guangzhou City Science and Technology Funding Plan (201508010043).

REFERENCES

[1] Sebastiani, Fabrizio. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1):1-47.
[2] Idris I, Selamat A. Improved email spam detection model with negative selection algorithm and particle swarm optimization[J]. Applied Soft Computing, 2014, 22(Complete):11-27.
[3] Guzella T S, Caminhas W M. A review of machine learning approaches to Spam filtering[J]. Expert Systems with Applications, 2009, 36(7):10206-10222.
[4] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey[J]. Ain Shams Engineering Journal, 2014, 5(4):1093-1113.
[5] Zeng J, Zhang S. Variable space hidden Markov model for topic detection and analysis[J]. Knowledge-Based Systems, 2007, 20(7):607-613.
[6] Jiang L, Li C, Wang S, et al. Deep feature weighting for naive Bayes and its application to text classification[J]. Engineering Applications of Artificial Intelligence, 2016, 52:26-39.
[7] Harish B S, Guru D S, Manjunath S. Representation and Classification of Text Documents: A Brief Review[J]. International Journal of Computer Applications, 2010, 8(2):110-119.
[8] Yang J, Liu Y, Zhu X, et al. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization[J]. Information Processing & Management, 2012, 48(4): 741–754.
[9] Peng H, Long F, Ding C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy[M]. IEEE Computer Society, 2005.
[10] Javed K, Babri H A, Saeed M. Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3):465-477.
[11] Jieming Y, Zhaoyang Q, Zhiying L. Improved Feature-Selection Method Considering the Imbalance Problem in Text Categorization[J]. The Scientific World Journal, 2014, 2014:1-17.
[12] Agarwal B, Mittal N. Text classification using machine learning methods-a survey[C]//Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Springer, New Delhi, 2014: 701-709.
[13] Saleh S N, El-Sonbaty Y. A feature selection algorithm with redundancy reduction for text classification[C]//2007 22nd international symposium on computer and information sciences. IEEE, 2007: 1-6.
[14] Li Y, Li T, Liu H. Recent advances in feature selection and its applications[J]. Knowledge and Information Systems, 2017, 53(3): 551-577.
[15] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines[J]. Machine learning, 2002, 46(1-3): 389-422.
[16] Lal T N, Chapelle O, Weston J, et al. Embedded methods[M]//Feature extraction. Springer, Berlin, Heidelberg, 2006: 137-165.
[17] Kohavi R, John G H. Wrappers for feature subset selection[J]. Artificial intelligence, 1997, 97(1-2): 273-324.
[18] Li Y, Luo C, Chung S M. Text clustering with feature selection by using statistical data[J]. IEEE Transactions on knowledge and Data Engineering, 2008, 20(5): 641-652.
[19] Xu Y, Jones G J F, Li J T, et al. A study on mutual information-based feature selection for text categorization[J]. Journal of Computational Information Systems, 2007, 3(3): 1007-1012.
[20] Liu L, Kang J, Yu J, et al. A comparative study on unsupervised feature selection methods for text clustering[C]//2005 International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2005: 597-601.
[21] Mengle S S R, Goharian N. Ambiguity measure feature‐selection algorithm[J]. Journal of the American Society for Information Science and Technology, 2009, 60(5): 1037-1050.
[22] Ogura H, Amano H, Kondo M. Feature selection with a measure of deviations from Poisson in text categorization[J]. Expert Systems with Applications, 2009, 36(3): 6826-6832.
[23] Yang J, Liu Y, Liu Z, et al. A new feature selection algorithm based on binomial hypothesis testing for spam filtering [J]. Knowledge-Based Systems, 2011, 24(6): 904-914.
[24] Yang J, Liu Y, Zhu X, et al. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization [J]. Information Processing and Management, 2012, 48(4): 741-754.
[25] Uğuz H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm[J]. Knowledge-Based Systems, 2011, 24(7): 1024-1032.

[26] Uysal A K, Gunal S. Text classification using genetic algorithm oriented latent semantic features[J]. Expert Systems with Applications, 2014, 41(13): 5938-5947.
[27] Labani M, Moradi P, Ahmadizar F, et al. A novel multivariate filter method for feature selection in text classification problems[J]. Engineering Applications of Artificial Intelligence, 2018, 70: 25-37.
[28] Forman G. A pitfall and solution in multi-class feature selection for text classification[C]//Proceedings of the twenty-first international conference on Machine learning. ACM, 2004: 38.
[29] Uysal A K. An improved global feature selection scheme for text classification[J]. Expert systems with Applications, 2016, 43: 82-92.
[30] Rehman A, Javed K, Babri H A, et al. Selection of the most relevant terms based on a max-min ratio metric for text classification[J]. Expert Systems with Applications, 2018, 114: 78-96.