Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 15, no. 4, April 2019, pp. 1131-1140
DOI: 10.23940/ijpe.19.04.p8.11311140

Feature Selection Combined Feature Resolution with Attribute Reduction based on Correlation Matrix of Equivalence Classes

Zhifeng Zhang* and Junxia Ma

School of Software, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

Abstract

Feature selection is one of the key steps in text classification. To some extent, it can affect the performance of text classification. In this paper, we firstly proposed an optimized document frequency-based word frequency and document frequency and then presented the feature resolution based on the optimized document frequency. Meanwhile, we introduced rough set into feature selection and provided an attribute reduction algorithm based on the correlation matrix of equivalence classes. We finally put forward a feature selection method combining the presented feature resolution with the provided attribute reduction algorithm. The proposed feature selection method firstly employs the presented feature resolution to select some valuable text features and filter out useless terms to reduce the sparsely of text feature spaces, and then it uses the provided attribute reduction algorithm to eliminate redundant features. The comparative experimental results show that the proposed feature selection method has certain advantages in consumed time, macro-average, micro-average, and average classification accuracy.

Keywords: feature selection; text classification; rough set; attribute reduction; equivalence class

(Submitted on November 8, 2018; Revised on December 10, 2018; Accepted on January 12, 2019)

1. Introduction

The 40th China Internet Network Development State Statistic report, which was released by the China Internet Network Information Centre on August 4, 2017, shows that in June 2017, there were 751 million Chinese netizens and the internet network penetration rate reached 54.3%. This was 4.6 percentage points above the global average, and it constantly updates and increases each day [1]. This also indicates that the information in internet network has exploded with the rapid development and massive popularization of Internet networks. Needless to say, the rapid increase in the amount of information in Internet networks has been expanding people's horizons [2]. However, information is producing far faster than people can gather and process information, which makes it impossible to quickly and effectively find the information that they are really interested in and creates a huge waste of time, money, and energy.

As long as information resources exist, they are likely to contain valuable knowledge and be used in traditional data mining. However, the most natural form of network information is text, and the network information is often represented as texts or is translated into texts. Many studies show that more than 80 percent of information on Internet networks is contained in text documents [3]. Because text data is unlabelled, semi-structural or non-structural, higher dimensional, non-uniform, and dynamic [4], traditional data mining cannot directly process them. This results in the phenomenon that information explodes but knowledge is relatively less, which greatly discourages people's enthusiasm to fully use the information resources in Internet networks [5]. Therefore, people are excited to be able to obtain such rich resources of information, but they are also deeply saddened by the inability to use these resources effectively. Faced with this challenging problem, how to efficiently organize, process, and manage these massive amounts of text information and rapidly, accurately, and comprehensively obtain the needed information has become the focus of academia and business circles [6]. In this context, text mining emerged and has become a research hotspot.

* Corresponding author.
E-mail address: zhuhaodong80@163.com

At present, text classification has become one of the main tasks of text mining [7]. Text classification is basically based on text features or word strings, and its premise is that there is a very close relationship between text features or word strings and text classes [8]. In text classification, a document is generally represented by a vector space model and is automatically distributed to one or more categories according to relevant contents and attributes [9]. The text feature set is formed by word segmentation, which ignores the semantic content of text features. Its scale is larger, so that the dimensions of text vectors often reach up to tens of thousands of dimensions when the space vector model is used to represent texts [10]. In theory, the more text features of text there are, the better the text is represented. However, this is not always the case; only a few text features play a positive role in text classification [11]. Those negative features not only occupy a lot of space, but also greatly increase classification time and thus degrade text classification performance [12]. Meanwhile, the text feature set may contain a large number of redundant text features [13]; these redundant text features will greatly increase the computational overhead of text classification, degrade the efficiency of text classification, and potentially produce classification results that are similar to the classification results of a smaller text feature subset. A larger text feature space also leads to higher consumption of time and space resources in the text classification process. Therefore, in order to ensure the text classification to be effectively executed quickly, we should purify the text feature set to eliminate negative features and filter out redundant features. On the basis of keeping the original meaning, text feature selection can find the text feature subset that more concisely and more representatively reflects the text content [14]. To a certain extent, text feature selection can eliminate negative features and redundant features and make the similarity between texts more accurate, which can improve the similarity between semantically related texts and at the same time reduce the similarity between semantically unrelated texts [15].

There are two main kinds of text feature selection methods. One is the independent evaluation method, which firstly constructs a text feature evaluation function to independently assess each text feature to let each text feature obtain the related weight, ranks text features in order of weight from large to small, and selects the text feature subset according to the weight threshold or a predetermined number of text features [16]. The other is the comprehensive evaluation method; this kind of method finds a comprehensive index for describing these text features from the original text feature set [17]. At present, the independent evaluation method is the main method [18], and it contains information gain [19], mutual information [20], x^2 statistics [21], cross entropy [22], word frequency [23], and document frequency [24]. However, after careful analysis, we find that the abovementioned text feature selection methods have the following deficiencies: they rely solely on weights as the criteria for selecting features and do not adequately consider the distribution characteristics of text features in between-class documents, the distribution characteristics of text features in within-class documents, or the potential relationships between text features. Therefore, the selected text feature subset contains a large amount of negative features and redundant features and is not more representative.

Under the circumstances, this paper firstly simply analyzes two commonly used text feature selection methods, word frequency and document frequency, presents an optimized document frequency combined word frequency with document frequency, defines the feature resolution based on the optimized document frequency to eliminate negative features, subsequently introduces rough set theory, and proposes an attribute reduction algorithm based on correlation matrix of equivalence class to remove redundant features.

2. Presented Feature Resolution

2.1. Word Frequency

When the word frequency method evaluates a text feature, it only considers the occurrence number of the text feature in the text document set [23]. For a text feature *f*, if the occurrence number of *f* in a text document set is equal or greater than a predetermined threshold, then *f* is still retained in the original text feature set; otherwise, *f* is deleted from the original text feature set. The probability formulas for this method are as follows:

$$P(f) = \frac{WF(f, D)}{|D|} \tag{1}$$

$$P(\overline{f}) = \frac{WF(\overline{f}, D)}{|D|} \tag{2}$$

$$P(c|f) = \frac{WF(f, c)}{WF(f, D)} \tag{3}$$

$$P(c|\overline{f}) = \frac{WF(\overline{f},c)}{WF(\overline{f},D)} \tag{4}$$

Among Equations (1) to (4), *f* represents a text feature, *c* represents the set of text documents that belong to a class, *D* represents the set of all text documents, |*D*| represents the number of text documents in *D*, *WF*(*f*, *c*) represents the occurrence number of *f* in *c*, *WF*($\overline{f}$, *c*) represents the occurrence number of non *f* in *c*, *WF*(*f*, *D*) represents the occurrence number of *f* in *D*, and *WF*($\overline{f}$, *D*) represents the occurrence number of non *f* in *D*.

The disadvantage of the word frequency method is that it only selects the words with high occurrence numbers in the text document set as text features and ignores the number of text documents that include the words in the text document set.

2.2. Document Frequency

When the document frequency method evaluates a text feature, it only considers the number of text documents that include the text feature in the text document set [24]. For a text feature *f*, if the number of text documents that include *f* in a text document set is equal or greater than a predetermined threshold, then *f* is still retained in the original text feature set; otherwise, *f* is deleted from original text feature set. The probability formulas for this method are as follows:

$$P(f) = \frac{DF(f,D)}{|D|} \tag{5}$$

$$P(\overline{f}) = \frac{DF(\overline{f},D)}{|D|} \tag{6}$$

$$P(c|f) = \frac{DF(f,c)}{DF(f,D)} \tag{7}$$

$$P(c|\overline{f}) = \frac{DF(\overline{f},c)}{DF(\overline{f},D)} \tag{8}$$

Among Equations (5) to (8), *f* represents a text feature, *c* represents the set of text documents that belong to a class, *D* represents the set of all text documents, |*D*| represents the number of text documents in *D*, *DF*(*f*, *c*) represents the number of text documents that include *f* in *c*, *DF*($\overline{f}$, *c*) represents the number of text documents that do not include *f* in *c*, *DF*(*f*, *D*) represents the number of text documents that include *f* in *D*, and *DF*($\overline{f}$, *D*) represents the number of text documents that do not include *f* in *D*.

The disadvantage of the document frequency method is that it only considers whether words are contained in the text document set and ignores the occurrence number of words in the text document set. This results in a problem: for the word *a* and the word *b*, if they have the same document frequency, then they are considered to have the same contribution to classify text documents, and the occurrence numbers of them in the text document set are ignored. However, the words with lower occurrence numbers in the text document set are usually noise terms.

The advantage of the document frequency method is that its time complexity increases linearly with the increase in the scale of the text document set, which is very suitable for text feature selection of the text document set with super-large scale.

2.3. Related Optimization

Through careful analysis, we find that there is a complementary relationship between the word frequency method and the document frequency method: for the text feature *f*, the word frequency method only counts the occurrence number of *f* in the text document set but ignores the number of text documents that include *f* in the text document set, while the document frequency method only considers the number of text documents that include *f* in the text document set but ignores the occurrence number of *f* in the text document set. Therefore, if the two methods are combined, they can complement each

other very well and potentially can obtain a better effect.

Definition 1 The optimized document frequency for the text feature *f* and the text class *c* refers to the number of text documents in which the occurrence number of *f* meets the given threshold *n* in the text document set of *c*. It is recorded as *Optimized-DF*$_n$ (*f*, *c*), where *n* is the minimal occurrence number of *f* in the text document set of *c*.

If the text feature *f* contributes significantly to *c*, then the text feature *f* should be concentrated in the text document set of *c*, rather than being spread across text document sets of various text classes. Therefore, we define the feature resolution to reflect the importance degree of a text feature for a text class.

Definition 2 Feature resolution for the text feature f_i and the text class c_j refers to the resolving power of f_i to c_j and is represented by Feature-Resolution (f_i, c_j). The formula is shown below:

$$\text{Feature-Resolution}(f_i, c_j) = \sum_{k=1 \wedge k \neq j}^{m} \left(\frac{Optimized\text{-}DF_n(f_i, c_j) - Optimized\text{-}DF_n(f_i, c_k)}{\sum_i Optimized\text{-}DF_n(f_i, c_j)} \right)^2 \tag{9}$$

In Equation (9), *m* refers to the quantity of classes and *Optimized-DF*$_n$ (f_i, c_j) is defined in definition 1. The greater the Feature-Resolution (f_i, c_j), the more concentrated the text feature f_i in the text class c_j, the better the categorization ability of the text feature f_i, and the more important the text feature f_i.

Feature resolution can be used effectively to select text features and eliminate negative features. However, there are many redundant features in the selected text feature subset by means of feature resolution, which may decrease the text classification performance. Therefore, we need to further optimize the selected text feature set. Attribute reduction is one of research contents in rough set theory and has a strong ability to eliminate redundancy [25]. In this paper, it is introduced to further optimize selected text features on the basis of feature resolution.

3. Proposed Attribute Reduction Algorithm

Rough set was proposed by Polish scholar Pawlak in 1982 [26] and has been widely used in many fields, such as machine learning, data mining, fault diagnosis [27], and fuzzy control [28]. Its core research content includes attribute reduction [29] and classification rule reduction [30]. At present, domestic and foreign scholars mainly study it in terms of attribute reduction [31]. Correlation matrix is a useful matrix and has been widely used in many fields [32]. This paper introduces correlation matrix into rough set and puts forward an attribute reduction algorithm based on the correlation matrix of equivalence classes.

3.1. Basics Theory

Given a decision information system $S =< U, R = C \cup D, V, f >$, U is a non-empty finite set of data objects, $R = C \bigcup D$ is an attribute set, *C* and *D* are the condition attribute set and the decision attribute set respectively, and $f: U \times R \to V$ is an information function. The basic knowledge related to this paper is as follows [33]:

Definition 3 In the information system $S =< U, R = C \cup D, V, f >, P \subseteq R$, and the equivalence relation of *P* in *U* is defined as below:

$$\text{IND}(P) = \{(x, y) \mid (x, y) \in U^2, \forall p \in P, f(x, p) = f(y, p)\} \tag{10}$$

In Equation (10), IND(*P*) = {X_1, X_2, ⋯, X_n} indicates that *U* is divided into several equivalence classes. It also can be represented as *U*/IND(*P*) = {X_1, X_2, ⋯, X_n}, where X_i refers to an equivalence class, and *i* = 1, 2, ⋯, *n*.

Definition 4 In the information system $S =< U, R = C \cup D, V, f >$, $P \subseteq C$, $X = \{X_1, X_2, \cdots, X_n\}$ is the equivalence class set of *P* in *U*. Assuming that $U^* = \{(X_i, X_j) \mid X_i, X_j \in X, 1 \le i < j \le n\}$, the correlation matrix **M** of *S* is defined as: the value of element m_{ij} in Row *i* and Column *j* in the matrix is $\forall u = (X^{\#}, X^*) \in U^*$ and $\forall p \in P$:

$$m_{ij} = \begin{cases} 1, if \cdots f(p,x) \neq f(p,y) \wedge \forall x \in X^{\#} \wedge \forall y \in X^{*} \wedge D(x) \neq D(y) \\ 0, \text{else} \end{cases} \tag{11}$$

In Equation (11), if a row vector of **M** is zero, it is redundant and can be deleted from **M**. In such cases, all non-zero vectors form a simplified correlation matrix **M***.

3.2. Proposed Algorithm Description

In correlation matrix **M***, the higher the number of element 1 in a column, the more important the corresponding attribute of this column; the lower the number of element 1 in a row, the more important the corresponding attribute of this row [33]. Based on this greedy strategy, this paper designs an attribute reduction algorithm, and the relevant steps are as follows:

Input A decision information system $S =< U, C \cup D, V, f >$.

Output An attribute reduction set *RedSet* of *S*.

Step 1 $RedSet = \varnothing$.

Step 2 Generate the correlation matrix **M** = (m_{ij}) of S according to formula 11.

Step 3 Delete the "zero row vectors" in **M** to generate the simplified correlation matrix **M***.

Step 4 Find columns that contain the maximum number of element 1 in **M***. If more than one column contains the same number of element 1, then the corresponding row with the least number of elements 1 is selected. Given the attribute c_j corresponds to the column *j* with the maximum number of element 1 and its corresponding row with the least number of element 1, carry out $RedSet = RedSet + \{c_j\}$.

Step 5 Delete the rows corresponding to the element 1 in Column *j* from **M***.

Step 6 If **M*** = ∅, then the algorithm ends and the attribute reduction set *RedSet* is output; otherwise, go to Step 4.

For this algorithm, if there are *m* condition attributes and *n* objects in a decision information system, the time complexity of each cycle is $O(n^2 \times m)$ and the whole algorithm can obtain the final attribute reduction set through *m* cycles at most, so the time complexity of the whole algorithm is $O(n^2 \times m^2)$, which is acceptable.

3.3. Example Verification

The known information system is shown in Table 1, and the corresponding correlation matrix **M*** is shown in Table 2.

According to definition 3, the equivalence class set $X = \{C_1, C_2, C_3, C_4, C_5\}$ can be obtained from Table 1, where $C_1 = \{x1, x5, x7\}$, $C_2 = \{x2, x6\}$, $C_3 = \{x3, x9\}$, $C_4 = \{x4\}$, and $C_5 = \{x8\}$.

Table 1. The information system

U	*a*	*b*	*c*	*d*
*x*1	0	1	1	0
*x*2	2	1	0	1
*x*3	0	0	0	1
*x*4	1	0	2	0
*x*5	0	1	1	0
*x*6	2	1	0	1
*x*7	0	1	1	0
*x*8	1	0	2	1
*x*9	0	0	0	1

According to step 4 of section 3.2, we can know that attribute *a* should be firstly merged into *RedSet*. After the correlation matrix is adjusted according to step 5 of section 3.2, as shown in Table 3, *d* is also merged into *RedSet*. After the correlation matrix is adjusted according to step 5 of section 3.2, the correlation matrix is empty and the final attribute reduction set $RedSet = \{a, d\}$. Its reduction result is the same as the reduction result of the classical attribute reduction

algorithm in literature [33], but its consumed time is much less than the consumed time of the classical attribute reduction algorithm in literature [33].

Table 2. The corresponding correlation matrix **M***

*U**	*a*	*b*	*c*	*d*
*u*1 = (*C*1, *C*2)	1	0	1	1
*u*2 = (*C*1, *C*3)	0	1	1	1
*u*3 = (*C*1, *C*4)	1	1	1	0
*u*4 = (*C*1, *C*5)	1	1	1	1
*u*5 = (*C*2, *C*3)	1	1	0	0
*u*6 = (*C*2, *C*4)	1	1	1	1
*u*7 = (*C*2, *C*5)	1	1	1	0
*u*8 = (*C*3, *C*4)	1	0	1	1
*u*9 = (*C*3, *C*5)	1	0	1	0
*u*10 = (*C*4, *C*5)	0	0	0	1

Table 3. The adjusted correlation matrix M* after selecting *a*

*U**	*a*	*b*	*c*	*d*
*u*2 = (*C*1, *C*3)	0	1	1	1
*u*10 = (*C*4, *C*5)	0	0	0	1

4. Description of Provided Feature Selection Method

Given *T* is the original text feature set and *C* is the class set, the minimum frequency threshold is *n* and the weight threshold is ω. The feature selection process is as follows:

Step 1 For $\forall c_j \in C$, the corresponding training document set is DS_j, and its original text feature set $T_j = T$;

Step 2 For $f_i \in T_j$, calculate Feature-Resolution(f_i, c_j) of f_i according to formula 9;

Step 3 If Feature-Resolution(f_i, c_j) $< \omega$, the f_i is removed from the T_j; otherwise, f_i is reserved;

Step 4 If there are unexamined text features in T_j, go to step 2;

Step 5 If there are unexamined classes in *C*, go to step 1;

Step 6 The selected text feature set of each class is merged as a single text feature set;

Step 7 Combine the merged text feature set and the class set with the training document set to obtain a decision table: $S = <U, C \cup D, V, f>$, and adopt the proposed attribute reduction algorithm to reduce text features.

Step 8 Adjust slightly the final text feature subset to highlight text features with a relatively large contribution to text classification.

5. Experimental Verification of Provided Feature Selection Method

5.1. Experimental Data

In this paper, we select the Chinese text classification corpus of Fudan University as the experimental corpus, which was constructed by a team within the Natural Language Processing of Computer Information and Technology Department and can be downloaded from http://www.nlp.org.cn/categories/default.php? cat_id = 16. The experimental corpus contains 20 categories and is divided into the training set and the test set. Each part contains 20 subdirectories. The same categories of documents are stored in a corresponding subdirectory. Meanwhile, each storage file contains only a document. All documents are numbered uniquely with the file name. After removing some repeated and damaged documents, only 14378 documents remain, wherein the training set contains 8214 articles and the test set contains 6164 articles without the repeated cross-category documents, i.e., one document only belongs to one category. The distribution of documents of the corpus is uneven. Among them, there are 1369 training documents for the class Economy of the training set, which has the largest number of documents, and 25 training documents for the minimal class. Meanwhile, there are 11 categories whose number of training documents of categories is less than 100, and the training set and the test set are not overlapped.

5.2. Experiment Settings

We employ the Chinese lexical analysis system (ICTCLAS), which was developed by the Institute of Computing Technology of Chinese Academy of Sciences, to carry out word segmentation. We select the Weka tool, which was developed at the University of Waikato, as the experiment platform. It includes a range of machine learning algorithms in the field of data mining, such as data preprocessing, classification and regression analysis, clustering, association rules, and visualization, and it can be downloaded from the following url: http://www.cs.waikato.ac.nz/ml/weka/. We also adopt MATLAB 7.0 to implement numerical calculation.

The presented method is mainly compared with the following six feature selection methods: Min-WF [16], MI-RS [17], CC-RBF [18], information gain (IG) [19], mutual information (MI) [20], and x^2 statistics (CHI) [21]. The parameter settings in the proposed method are $n = 4$ and $\omega = 0.095$. We use the KNN classifier (K is set to 20, and cosine distance is adopted to calculate similarity) to implement classification experiment. Meanwhile, we select the consumed time (unit: *s*), the micro average F_1, the macro average F_1, and the classification precision as the performance evaluation standards.

5.3. Experimental Results and Analysis

5.3.1. Experimental Results and Analysis on Consumed Time

In the process of experiments, each method is performed ten times, and we take the average time of the ten records as the experimental results on consumed time. The experimental results are shown in Table 4.

Table 4. The experimental results on consumed time (unit: *s*)

The presented method	CHI	IG	MI
1392	1538	1427	1496
Min-WF	MI-RS	CC-RBF	
1172	1421	1414	

After careful analysis, we can see that Min-WF only calculates the documentary frequency, which meets the given word frequency threshold. The presented method, MI-RS, and CC-RBF need to further select features based on Min-WF. CHI, MI, and IG need to adopt complex statistical calculations to select features, so Min-WF is better than the other six methods based on consumed time. Since the presented method in this paper reduces the operating scale by pruning the correlation matrix continuously in the process of eliminating redundancy, it is better than MI-RS, CC-RBF, CHI, MI, and IG based on consumed time.

5.3.2. Experimental Results and Analysis on Classification Performance

The experiments are divided into two groups, and the details are as follows:

The first group: the presented method is mainly compared with the three feature selection methods information gain (IG), mutual information (MI), and x^2 statistics (CHI) on macro average F_1 and micro average F_1. In the process of experiments, each method is performed ten times under different numbers of features, and we take the average macro average F_1 and the average micro average F_1 of the ten records as the experimental results on the macro average F_1 and micro average F_1. The experimental results are shown in Figures 1 and 2 respectively. Note that in Figures 1 and 2, the numbers from 1 to 15 represent the number of features respectively, i.e., 50, 100, 200, 500, 800, 900, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000. The *y*-axis represents the corresponding macro average F_1 and micro average F_1 in unit %.

The second group: the presented method is mainly compared with the three feature selection methods Min-WF, MI-RS, and CC-RBF on classification precision. In the process of experiments, each method is performed ten times, and we take the average classification precision of the ten records as the experimental results of classification precision. The experimental results are shown in Table 5.

From Figure 1, Figure 2, and Table 5 we can see that the presented method is better than the other six methods in overall classification performance. After careful analysis, the reasons are as follows: the presented method not only adopts the presented feature resolution to examine the word frequency and the document frequency of features synchronously to eliminate negative features, but also uses the proposed attribute reduction algorithm to further eliminate redundant features, and it is not sensitive to data sample distribution. MI only checks the existence situation of text features in the text document

set, CHI checks the existence situation and the nonexistence situation of text features in the text document set, but they do not further eliminate redundant features. Therefore, the presented method is better than MI and CHI, and CHI is better than MI. Since IG is extremely sensitive to data sample distribution, under the condition of uneven data sample distribution in the experiments, its classification performance is the worst. Min-WF only evaluates the documentary frequency that meets the minimum frequency threshold, and the presented method, MI-RS, and CC-RBF all optimize the selected text feature set based on Min-WF and have similar processes. Therefore, the overall classification precision of Min-WF is the worst, while the overall classification precisions of the presented method, MI-RS, and CC-RBF have no significant difference.

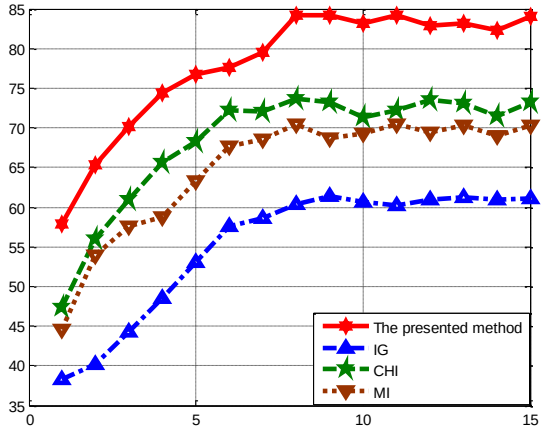

Figure 1. The experimental results on macro average F_1

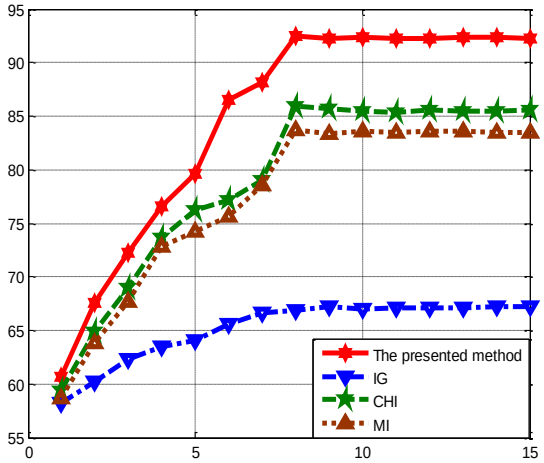

Figure 2. The experimental results on micro average F_1

Table 5. The experimental results on classification precision

Methods	The presented method	Min-WF	MI-RS	CC-RBF
Classification precision	98.27%	65.05%	97.56%	98.21%

6. Conclusions

On the basis of the analysis of word frequency and document frequency, this paper firstly proposed an optimized document frequency, put forward feature resolution based on the optimized document frequency to eliminate negative features, and subsequently introduced rough set to present an attribute reduction algorithm based on the correlation matrix of equivalence classes to further eliminate redundant features. From the experimental results, the proposed feature selection method has certain advantages in classification performance and time performance, and it has a certain practical value.

Acknowledgments

The authors would like to thank the editors and the anonymous reviewers for their helpful comments and suggestions, which have improved the presentation. This work was supported in part by the Key Science Research Project of Colleges and

Universities in Henan Province of China (No. 19A520009) and the National Science Foundation of China under (No. 81501548).

References

1. "The 40th China Internet Network Development State Statistic Report," (http://www.cac.gov.cn/2017-08/04/c_1121427728.htm, last accessed on September 15 2018)
2. X. F. Zhang and F. X. Kong, "Research on Scientific Literature Retrieval based on Semantic Concept Analysis," *Journal of Information Theory and Practice*, Vol. 39, No. 8, pp. 115-118, August 2016
3. D. T. Pele, O. E. Lazar, and A. Dufour, "Information Entropy and Measures of Market Risk," *Entropy*, Vol. 19, No. 5, pp. 226-244, May 2017
4. J. H. Liu, Y. J. Lin, M. L. Lin, S. X. Wu, and J. Zhang, "Feature Selection based on Quality of Information," *Neurocomputing*, Vol. 225, pp. 11-22, February 2017
5. B. Tang, S. Kay, and H. B. He, "Toward Optimal Feature Selection in Naive Bayes for Text Categorization," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 9, pp. 2508-2521, September 2016
6. T. Basu and C. A. Murthy, "A Supervised Term Selection Technique for Effective Text Categorization," *International Journal of Machine Learning and Cybernetics*, Vol. 7, No. 5, pp. 877-892, May 2016
7. R. H. W. Pinheiro, G. D. C. Cavalcanti, and I. R. sang, "Combining Dissimilarity Spaces for Text Categorization," *Information Sciences*, Vol. 406, pp. 87-101, September 2017
8. T. Sun, S. Y. Qian, and H. D. Zhu, "Feature Selection Method based on Category Correlation and Discernible Sets," *Journal of Computational Information Systems*, Vol. 11, No. 22, pp. 9687-9698, August 2014
9. H. D. Zhu, H. C. Li, D. Wu, D. S. Huang, and B. Wang, "Feature Selection Method based on Feature Distinguishability and Fractal Dimension," *Journal of Information and Computational Science*, Vol. 36, No. 5, pp. 6033-6041, May 2015
10. D. Oreski, S. Oreski, and B. Klicek, "Effects of Dataset Characteristics on the Performance of Feature Selection Techniques," *Applied Soft Computing*, Vol. 52, No. 3, pp. 109-119, March 2017
11. A. Katrutsa and V. Strijov, "Comprehensive Study of Feature Selection Methods to Solve Multicollinearity Problem According to Evaluation Criteria," *Expert Systems with Applications*, Vol. 76, No. 7, pp. 1-11, July 2017
12. Z. H. Zhang, L. Bai, and Y. H. Liang, "Joint Hypergraph Learning and Sparse Regression for Feature Selection," *Pattern Recognition*, Vol. 63, No. 3, pp. 291-309, March 2017
13. B. S. C. Wade, S. H. Joshi, and B. A. Gutman, "Machine Learning on High Dimensional Shape Data from Subcortical Brain Surfaces: A Comparison of Feature Selection and Classification Methods," *Pattern Recognition*, Vol. 63, No. 3, pp. 731-739, March 2017.
14. A. Khaled, W. H. Guo, and C. H. Yang, "Feature Selection based on Rough Sets and Minimal Attribute Reduction Algorithm," *International Journal of Hybrid Information Technology*, Vol. 9, No. 8, pp. 333-346, August 2016
15. S. Q. Wang and J. M. Wei, "Feature Selection based on Measurement of Ability to Classify Subproblems," *Neurocomputing*, Vol. 224, pp. 155-165, February 2017
16. I. A. Gheyas and L. S. Smith, "Feature Subset Selection in Large Dimensionality Domains," *Pattern Recognition*, Vol. 43, No. 1, pp. 5-13, January 2010
17. H. D. Zhu and H. C. Li, "Feature Selection based on Mutual Information and Rough Set Theory," *Computer Engineering*, Vol. 37, No. 15, pp. 181-183, August 2011
18. H. D. Zhu and H. C. Li, "Feature Selection Combined Classificatory Concentration with Improved RBF Neural Network," *China Journal of Microelectronics & Computer*, Vol. 28, No. 2, pp. 145-149, February 2011
19. M. H. Nguyen and F. D. L. Torre, "Optimal Feature Selection for Support Vector Machines," *Pattern Recognition*, Vol. 43, No. 3, pp. 584-591, March 2010
20. H. W. Liu, J. G. Sun, and L. Liu, "Feature Selection with Dynamic Mutual Information," *Pattern Recognition*, Vol. 42, No. 7, pp. 1330-1339, July 2009
21. A. Destrero, S. Mosci, C. D. Mol, A. Verri, and F. Odone, "Feature Selection for High-Dimensional Data," *Computational Management Science*, Vol. 6, No. 1, pp. 25-40, January 2009
22. X. Yan, "A Formal Study of Feature Selection in Text Categorization," *American Journal of Communication and Computer*, Vol. 6, No. 4, pp. 32-41, July 2009
23. A. Rehman, K. Javed, and H. A. Babri, "Feature Selection based on A Normalized Difference Measure for Text Classification," *Information Processing & Management*, Vol. 53, No. 2, pp. 473-489, February 2017
24. S. R. Y. Leela, V. Sucharita, B. Debnath, and H. J. Kim, "Performance Evaluation of Feature Selection Methods on Large Dimensional Databases," *International Journal of Database Theory and Application*, Vol. 9, No. 9, pp. 75-82, September 2016
25. J. Fan, Y. L. Jiang, and Y. Liu, "Quick Attribute Reduction with Generalized Indiscernibility Models," *Information Sciences*, Vol. 397, pp. 15-36, January 2017
26. X. X. Zhang, D. G. Chen, and E. C. C. Tsang, "Generalized Dominance Rough Set Models for the Dominance Intuitionistic Fuzzy Information Systems," *Information Sciences*, Vol. 378, pp. 1-25, January 2017
27. B. Yang and B. Q. Hu, "On Some Types of Fuzzy Covering-based Rough Sets," Fuzzy *Sets and Systems*, Vol. 312, pp. 36-65, April 2017
28. Y. H. She, X. L. He, and H. X. Shi, "A Multiple-Valued Logic Approach for Multigranulation Rough Set Model," *International Journal of Approximate Reasoning*, Vol. 82, No. 1, pp. 270-284, January 2017
29. J. Qian, C. Y. Dang, and X. D. Yue, "Attribute Reduction for Sequential Three-Way Decisions under Dynamic Granulation," *International Journal of Approximate Reasoning*, Vol. 85, pp. 196-216, June 2017

30. X. Y. Zhang and D. Q. Miao, "Three-Way Attribute Reducts," *International Journal of Approximate Reasoning*, Vol. 88, pp. 401-434, September 2017
31. U. Jamal, G. Rozaida, and M. M. Deris, "An Empirical Analysis of Rough Set Categorical Clustering Techniques," *Plos One*, Vol. 12, No. 1, pp. 1-22, January 2017
32. D. Hu, X. C. Yu, and J. Y. Wang, "Statistical Inference in Rough Set Theory based on Kolmogorov–Smirnov Goodness-of-Fit Test," *IEEE Transactions on Fuzzy Systems*, Vol. 25, No. 4, pp. 799-812, April 2017
33. S. T. Hu and Y. Q. He, "Rough Decision Theory and Application," Beihang University Press, Beijing, 2006

Zhifeng Zhang received his B.S. degree from Xi'an University of Electronic Science and Technology in 2001 and his M.S. degree from Xi'an University of Technology in 2006. He is currently an associate professor in the School of Software at Zhengzhou University of Light Industry. His major research interests include cloud computation, intelligence information processing, and data mining.

Junxia Ma received her B.S. degree from Henan Normal University in 1996 and her M.S. degree from Zhengzhou University in 2007. She is currently a lecturer in the School of Software at Zhengzhou University of Light Industry. Her major research interests include knowledge engineering and data mining.