Cluster Comput (2017) 20:1425–1435
DOI 10.1007/s10586-017-0812-9

Feature selection method based on multiple centrifuge models

Qinghu Wang[1] · Lisha Liu[2] · Jingqing Jiang[1] · Mingyang Jiang[1] · Yinan Lu[3] · Zhili Pei[1]

Received: 30 October 2016 / Revised: 9 February 2017 / Accepted: 25 February 2017 / Published online: 6 March 2017

Abstract High-dimension of feature space in text classification is a major problem of it. Feature selection is an effective method for feature reduction. A multiple centrifuge models based feature selection method is put forward in the view of the hypothesis that the same documents have core feature set in the text classification and the classes of the same high-frequency feature words of document have affinity. The proposed feature selection algorithm made a lot of innovation ideas in the field of feature reduction which improve the values of the low-frequency features in classification meanwhile ensuring the classification effect. The experiments in the Reuters-21578 corpus show that this method has better classification effect, and effectively improves the utilization of medium or low frequency features which have strong classification ability.

Keywords Centrifuge model · Qrderly whole class feature vector · Centroid feature set · Centrifuge matrix · Torque adjoint matrix

✉ Zhili Pei
zhilipei@sina.com

Qinghu Wang
stigerkingdom@126.com

[1] College of Computer Science and Technology, Inner Mongolia University for the Nationalities, Tongliao 028000, China

[2] Logistics University of PAP, Tianjin 300309, China

[3] College of Computer Science and Technology, Jilin University, Changchun 134000, China

1 Introduction

The so-called text mining is a method and tool for discovering hidden knowledge and pattern from large number of documents. It evolved from data mining, while it is different from traditional data mining. Text classification is a typical application of text mining field [1]. It discovers concealed knowledge and pattern from a large number of already classified sample set, and determines the category belongings of test documents by it [2]. Owing to the immaturity of text semantic knowledge expression technology, current classic method of text classification reduces text semantics to the feature words semantics level, which makes the feature set as the classification standard [3]. If we take all the words as feature item, the dimension of feature vectors would be too massive. This unprocessed text vector would bring huge computational overhead for following work, which would even reduce the accuracy of classification and clustering algorithm and cannot get the satisfying classification result [4]. In summary, high-dimensional problems of characteristics will be the main obstacle to text classification [5]. Feature selection is typical dimension reduction method, which focuses on selecting a feature subset in sample set to make the text classification performance best based on this set [6].

Generally, the methods of feature selection can be divided into four categories. 1. Transform the original feature set to the new feature set with smaller dimension using the method of mapping or transformation; 2. Select some features with highest value to classification from original feature; 3. Select the most influential feature based on expert knowledge engineering; 4. Find out the feature with most classification information using mathematics method [7]. Basing on the knowledge engineering method and with the help of professionals, plenty of inference rules are defined for each category [8]. If a document can meet all these inference rules, it can be

judged that it belongs to the category. Because human judgment factors are put into the system, its accuracy improved greater than that of word matching method. But the flaws of the method are still very obvious, such as long classification period, high cost, and low efficiency [9]. Other kinds of methods are widely applied to methods basing on statistics and text classification methods basing on machine learning, which are present mainstream feature selection methods. Many feature selection methods based on statistics and machine learning have achieve good effect by improving classification efficiency and precision while saving a lot of effort meanwhile [10]. This paper will analysis two typical feature selection methods and explore their advantages and disadvantages.

X^2testing method is a method based on testing the independence of two variables in mathematical statistics which is to select feature words by judging the independence of feature words and classification [11]. The experimental data shows that X^2testing has better classification effect and its classification effect is less affected by the training set, relatively stable. The formula of X^2testing is

$$\begin{aligned} CHI(t) &= \sum_{i=1}^{m} Chi(t, C_i) \\ &= \frac{N[P(t, C_i) * P(\bar{t}, \bar{C}_i) - P(t, \bar{C}_i) * P(\bar{t}, C_i)]^2}{P(t) * P(C_i) * P(\bar{t}) * P(\bar{C}_i)} \end{aligned} \tag{1}$$

From Formula 1, we can see that X^2 testing integrally considers the case of feature appearance and disappearance while it does not consider the frequency characteristic of feature words but pay more attention to the document number of feature words appearance. The defect of low-frequency feature words of X^2 testing seriously neglects the classification value of words frequency, while in fact it plays an important role in classification [12].

Information gain is an assessment methodology based on entropy, which involves many mathematics theories and complex entropy theory formula. The core idea of information gain is to measure whether the existence of a certain feature item affects the class prediction. It computes the difference of information entropy before and after the emergence of it in document [13]. The bigger the value of information gain of certain feature item, the greater its contribution to the classification. The formula of information gain method is

$$\begin{aligned} IG(t) = {} & p(t) \sum_i P(C_i|t) \log \frac{P(C_i|t)}{P(C_i)} \\ & + p(\bar{t}) \sum_i P(C_i|\bar{t}) \log \frac{P(C_i|\bar{t})}{P(C_i)} \end{aligned} \tag{2}$$

The Formula 2 shows that the information gain also integrally considers the influence to the classification whether the appearance or disappearance of feature and its classification effect is very good [14]. While the biggest problem of information gain is that it only considers the contribution of feature to the whole classification system but not focuses to certain class. This makes it only suitable to be "whole "feature selection whose performance of classification is not good. Meanwhile, information gain considers the situation not happened, especially under the situation that class and feature distribute imbalanced [15] . Thus, the majority classes are negative classes and the majority features do not emerge. The function value is decided by feature that does not appear, so the effect of information gain will be greatly reduced [16].

In summary, the X^2testing and information gain are two effective feature selection methods that have better classification results [17]. While X^2testing neglects the feature word frequency, the factor having great reference value to classification. Information gain has natural defect in the aspects of "whole feature selection" and sample set uneven distribution [18]. Owing to the situation that statistical feature selection method is difficult to maintain the semantic categories, two methods mentioned above loss category semantic recognition in varying degrees [19]. The method that takes feature subset as classification basis leads the semantic level to reduce from document to word. The more reasonable method would be to maintain the original semantics of "rough" category as much as possible. In fact, feature word usually has category affinity which assumes in a closed semantic environment, we can find a convergent feature set to represent a category [20]. The semantics of this convergent feature set will be close to the semantics of category as much as possible. Now majority of feature selection methods do not consider the category affinity of feature words [21]. Basing the hypothesis that same class texts have core feature set and same text high frequency feature words have category affinity in text classification, the difference of feature among categories has been fully considered in this paper [22]. A centrifuge feature selection method is put forward which effectively circumvent the designing flaws of traditional feature selection methods basing statics and improve the classification effect further [23].

2 Centrifuge model

Assuming given a category with confirmed document numbers, the category represents a kind of closed semantic. Then a convergent feature set can be found in this category to represent the category. The semantics and power of this convergent feature set will be close to the semantics of category as much as possible. Each feature of this feature set has greater category affinity to the category. The feature selection method we

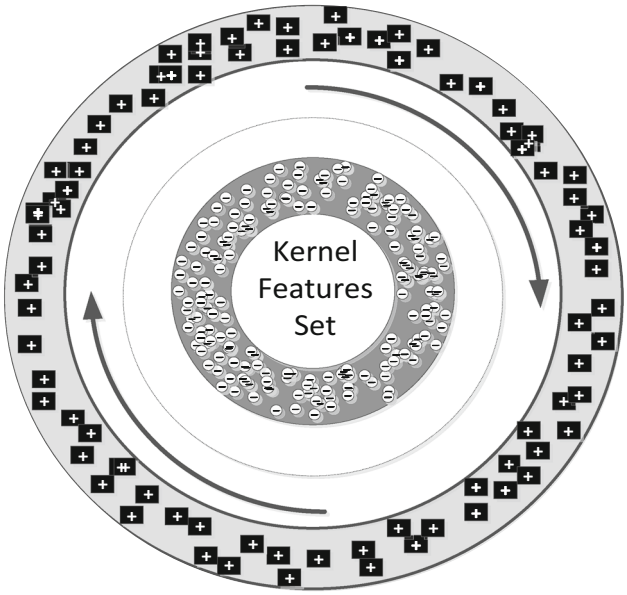

Fig. 1 Centrifuge model

put forward is based on the fact that same category documents in text classification have core feature set and the hypothesis that high frequency feature words of same category documents have category affinity. It means that there are core feature sets whose feature words have high class tendency. After pretreatment which deletes stop word and feature word that have high coverage percentages in the whole sample document and is obviously invalid for classification, having high frequency feature words in same class document means having high class tendency. These high frequency feature words have higher membership for actual category of this classification space.

The idea of centrifuge model for realizing feature selection is from a natural phenomenon. Imagine putting multiple small objects of different qualities on the center of a rolling wheel, as the wheel rolling accelerates, because of the difference of objects qualities, the objects of different qualities would distribute on the wheel disk area and objects with big qualities would be around the wheel axis area. Basing on the idea mentioned above, during the process of text feature selection, according to the difference of category attributes of feature words, a similar rolling wheel model is designed. Exert related necessary rolling movement to wheel model to distribute different feature words on the wheel model and ultimately extract feature words near the closest area of wheel axis. Thus, it realizes the aim of feature selection. The wheel model is exactly the centrifuge model put forward in this paper. The centrifuge model figure is as shown in the Fig. 1.

The model in Fig. 1 is the centrifuge model of category in classification sample set. In the whole classification sample space, the quantity of this classification model is equal to the quantity of text classification. The centroid of the model is the feature words set with weight of the category that is represented by the model. The centroid area is the magnetic field with positive pole. The outermost area of the circle is the magnetic field with negative pole. The default centroid set is all the features of the category. Initially, each feature does not carry electron (weight is 0) or carry negative electron (weight is negative). The dynamic driving centrifuge to rotate is the intensity adding positive electron to the outer area of centrifuge to reduce negative pole magnetic field, and the behavior "adding positive electron" is successively laying feature set (weight is positive number, with positive electron) of other documents not belonging to this centrifuge on the outer area of centrifuge. This can let the feature words, which was at the axes of centroid (with negative electron), depart from centroid step by step. When the centrifuge stops rotating, the feature words with classification affinity will separate surrounding axes of centrifuge and that realize the feature selection of category.

3 Centrifuge model operation mechanism

Basing on the description of centrifuge model operation mechanism above, this section sum up the key steps of model working process and guide the design and realization of feature selection algorithm depending on the thinking angle of centrifuge model.

3.1 Building centrifuge static model

The principle task of building centrifuge static model is to determine the initial centroid. That means building the feature set of category centroid with weight. The feature word of initial centroid feature concentration carry negative electron or does not carry electron. The weight determination of centroid feature concentration feature words depends on the calculation of each feature word weight of this category, and the weight should be negative or '0' value.

3.2 Key elements driving centrifuge rotate

The reason that centrifuge rotate is that adding positive electron to the outer area of centrifuge. The origin of this positive electron is the other documents whose category is different from that of this centrifuge. It can be seen as a mechanism for data injection. The data injected should meet certain requirements. So preprocess to the injected data should be done to meet the needs of model run. It improves positive pole magnetic of the outer area of centrifuge by adding positive electron, which forcing feature words near centroid moving towards the outer area of centrifuge.

3.3 Capture centroid core

When all the centrifuge representing each document category stop rotating, extract feature words of each category near centrifuge axes according to specified rules. The feature set is the final result of feature selection and the problem is solved.

4 Centrifuge model algorithm design

Basing on the discussion of centrifuge model above, to realize text feature dimension reducing truly, the model operation mechanism must be turned to generalized algorithm realization. We will elaborate the key steps of centrifuge algorithm as follows:

Step 1 Build high-dimensional feature vector of whole category basing on sorting mechanism

After remove stopping words and clearly invalid feature words that have high coverage classification in the whole sample document, then sort the entire feature words of the whole sample space. The principle of feature words sorting is that the English word ordering feature words alphabet in ascending order, while the Chinese characteristics ordering the same as English word after being converted into the phonetic Chinese characteristics. The row vector, the characteristics after sorting is the feature vector of high-dimensional whole category. Of course, now the vector has no weight and the default weight value is 0. The aim of this vector is concerning the sequence of features. The feature vector of high-dimensional whole category T_g is represented as

$$T_g = (term_1, term_2, ...term_n) \tag{3}$$

Step 2 Build category centroid feature set and determine the nuclear vector of category

Build centroid feature set of every category basing on the high-dimensional total category vector feature words order built by step1. The centroid feature set of this category is a high-dimensional category vector feature row vector $T_{ig,}$

$$T_{ig} = (W_{i1}, W_{i2}, ...W_{ik}) \tag{4}$$

T_{ig} represents the centroid feature set of category i, W_{ik} represents the feature weight of number k in category i, weight of each feature word is calculated following as the formula below:

$$w_{ik} = -(\sum_{m=1}^{D_k} tf(m)) * \log(\frac{n_m}{N_D} + 0.01) \tag{5}$$

There into D_k represents the num of documents of category$i$. $tf(m)$ represents the feature frequency of m_{th} document of feature word k in category i. $\frac{n_k}{N_D}$ represents the ratio of document num of category i emergence in feature word k and the total document num in category i.

The nuclear vector of category is a bull vector, representing if the feature word in high-dimensional total category vector feature space appear in category i, the value of emergency is 1, otherwise the value is 0. As is shown in Formula 6.

$$K_i = (b_{i1}, b_{i2}, ..., b_{ik}) \tag{6}$$

There into $b_{ik} = \begin{cases} 1 & feature\ k\ appears\ in\ category\ i \\ 0 & feature\ k\ does\ not\ appear\ in\ category\ i \end{cases}$.

Step 3 Build centrifuge matrix

Centrifuge matrix is the matrix that built by the category centroid feature set as row vectors that built by the step2. The matrix produces ultimate feature set after dimension reduction by iterative calculating process. The centrifuge matrix CM represents as:

$$CM = \begin{bmatrix} w_{11} & w_{12} & w_{13} & ..... & w_{1k} \\ w_{21} & w_{22} & w_{23} & ..... & w_{2k} \\ w_{31} & w_{32} & w_{33} & ..... & w_{3k} \\ ..... & ..... & ..... & ..... & ..... \\ w_{n1} & w_{n2} & w_{n3} & ..... & w_{nk} \end{bmatrix} \tag{7}$$

There into, n is the number of category in sample space.

Step 4 Build torque adjoint matrix of sample document

The torque adjoint matrix designed by this paper is the "injecting data" that drive centrifuge rotating. Each document in sample space is corresponding to its respective torque adjoint matrix. The matrix structure of torque adjoint matrix is same as the one of centrifuge matrix CM, and is only different on the calculation of weights. The weight of row vectors of this document in categories are all 0. The torque adjoint matrix is a redundant matrix and it is very necessary. The redundant of matrix data is because of the parallel computing of "injecting data". The torque adjoint matrix $AM_{i,j}$ of document j in category i represents as:

$$AM_{i,j} = \begin{bmatrix} w_{j,1} & w_{j,2} & w_{j,3} & ..... & w_{j,k} \\ ..... & ..... & ..... & ..... & ..... \\ w_{j,1} & w_{j,2} & w_{j,3} & ..... & w_{j,k} \\ 0 & 0 & 0 & ..... & 0 \\ w_{j,1} & w_{j,2} & w_{j,3} & ..... & w_{j,k} \\ ..... & ..... & ..... & ..... & ..... \\ w_{j,1} & w_{j,2} & w_{j,3} & ..... & w_{j,k} \end{bmatrix} \tag{8}$$

There into,

$$w_{j,k} = tf(k) * \frac{N_D}{n_k} * \log(\frac{n_k}{N_D} + 0.01) \quad (9)$$

$tf(k)$ in Formula 9 represents the feature frequency of k_{th} feature word of j_{th} document in category i. $\frac{n_k}{N_D}$ represents the radio of documents number of feature k appearing in category i and total document numbers in category i.

Step 5 Centrifuge machine calculating

Centrifuge matrix CM adds respectively to the torque adjoint matrix $AM_{i,j}$ of each document in sample space. Two points should be emphasized here: first, the initial value of Centrifuge matrix CM is negative number or 0, and the weight of torque adjoint matrix in document is positive number or 0. This data type is the origin of data change (that is, the position move of small objects with quality in centrifuge model). Second, it is found that the calculation process of centrifuge machine contains all the addition operations. The calculation model embodies natural parallel computing characteristics and provides a good support for parallel computing.

Step 6 Do the nuclear process to the centrifuge matrix with the nuclear vector

Because some categories does not exist certain feature words in sample space and the calculation process of centrifuge machine in Step5 contains multiple matrix addition operations, it makes many feature words that does not emerge in assigned categories have weight. To keep the consistency of the whole calculating process, the logical mistake need to be avoided. If the feature word does not emerge in assigned category, the weight before and after calculation should not change, and are all 0. So certain calculation process should be done here. Meanwhile, zero value will emerges in the computing process. Two kinds of zero value have different semantics. Zero value emerging in the computing process represents the contribution of feature to the classification, and the other kind of zero value represents whether the feature emerges in the category. The operation to maintain the logical integrity is to use nuclear vector K_i of each category multiply the i_{th} row of centrifuge matrix CM. Because the nuclear vector is a bull vector, the weight of feature word that does not emerge in the corresponding category in centrifuge matrix is set to 0 after nuclear operation ($K\Theta CM$). The definition of nuclear operation is:

$$K\Theta CM = \begin{bmatrix} b_{11} & \ldots\ldots & b_{1k} \\ \ldots & \ldots\ldots & \ldots \\ b_{i1} & \ldots\ldots & b_{ik} \end{bmatrix} \otimes \begin{bmatrix} w_{11} & \ldots\ldots & w_{1k} \\ \ldots & \ldots\ldots & \ldots \\ w_{i1} & \ldots\ldots & w_{ik} \end{bmatrix} = \begin{bmatrix} b_{11}*w_{11} & \ldots\ldots & b_{1k}*w_{1k} \\ \ldots & \ldots\ldots & \ldots \\ b_{i1}*w_{i1} & \ldots\ldots & b_{ik}*w_{ik} \end{bmatrix} \quad (10)$$

Step7 Feature selection following the threshold rules

After steps 1-6, according to the centrifuge machine idea put forward by this paper, the dimension-reduced feature set can be gotten through certain rules. The extraction rules this paper use separately select the first m features with smallest weight in each category row vector of centrifuge matrix CM as the feature set after feature reduction, if these features repeat, assuming the repetition number is n, the n-1 unrepeated features with the biggest weight will be chosen in all the categories.

5 Experimental analysis

5.1 Experimental design

To effectively test the result of centrifuge model feature selection method, experimental procedure followed the following design ideas. Basing on the hypothesis that centrifuge model focus on feature category affinity, we judge whether this method will contribute to the classification. Basing on the analysis above, X^2test indeed exists the defect of low frequency feature words. Information gain is serious shortage at dealing with the problem of "whole feature selection" and sample set unevenly distributed. The centrifuge model feature selection method try to avoid the defect, what about the effect? The experimentation design for answering the questions above and adopt open standards Reuters-21578 data set [24]. For comparatively analysis, the experimentation respectively select X^2 test, information gain and centrifuge model methods as feature selection methods to do the experiment. Meanwhile, in order to observe the degree of adaptation of text classification to feature selection methods, and consider the bias of different indicators of feature selection methods, experiments use two classification methods: KNN and naive Bayes. The general idea of experimental design is as shown in Fig. 2.

5.2 Experimental index

5.2.1 Data set

The experimental testing data set using in this paper is Reuters-21578 [25]. Reuters-21578 data set has totally 120 types according to subject, clearly marked types of 10789 texts. There into, training set has 7770 texts and 3019 testing set texts distributing in 90 non-empty types [26].

5.2.2 Evaluation index

The experiment adopts accuracy, recall rate and F1 test values as evaluation index of classification. The accuracy is defined as

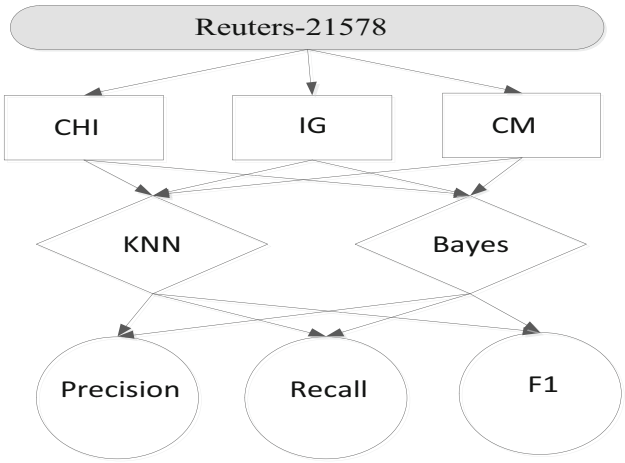

Fig. 2 Experimental design model

$$P(d_i) = \sum_{j=1}^{k} P(c_j)P(d_i|c_j), \tag{12}$$

Calculate probabilities for all documents classes in the given di case, the class whose probability is max value is the class that di in, which is

$$d_i \in c_j \quad if \quad P(c_j|d_i) = \max_{l=1}^{k}\{P(c_l|d_i)\}. \tag{13}$$

Native Bayesian classification algorithm assume the feature words of document are independent, denying the co-occurrence relationship of feature words, that means not approving the point of feature words class infinity in same document [31]. Meanwhile, the classification deviation of this method for small samples is probably bigger which is

$$P = \frac{Samples\ number\ that\ are\ put\ correctly\ in\ this\ category}{Samples\ number\ that\ are\ put\ correctly\ in\ category + samples\ number\ that\ are\ put\ incorrectly\ in\ other\ category},$$

which measures the precision of categories. The recall rate is defined as

$$R = \frac{Samples\ number\ that\ are\ put\ correctly\ in\ this\ category}{Samples\ number\ that\ are\ put\ correctly\ in\ this\ category + samples\ number\ that\ are\ put\ incorrectly\ in\ this\ category},$$

which depicts the recall of category [12]. F1 test value $F1 = \frac{2(P\times R)}{P+R}$ comprehensive measure the precision and recall of classification [27].

5.2.3 Classification method

For comparatively analysis, focusing on analyzing the effect of feature selection algorithm to classification at different index, and discovering the relationship between feature selection and classification, the experiment adopts native Bayesian classification algorithm and KNN [28]. Native Bayesian classification algorithm is basing on a hypothesis that in a given document category semantic environment, document properties are independent of each other, that means the feature words of a document are independent [29]. Suppose d_i is any a document that belonging to any certain categoryc_j of document category $C = \{c_1, c_2, \ldots, c_k\}$ [30]. Native Bayesian classification algorithm use the formula below:

$$P(c_j|d_i) = \frac{P(c_j)P(d_i|c_j)}{P(d_i)}. \tag{11}$$

suitable to comparatively analysis in the uneven environment [32].

KNN classification algorithm is a mature method in theoretically [33]. If multiple samples of k most similar (which is the closest in feature space) samples of a sample in feature space belong to certain category, the sample belong to the category too [34]. The description of KNN algorithm is:

$$\sum_{i=(K+1)/2}^{K} \binom{K}{i} P(\omega_i|X)^i [1 - P(\omega_i|X)]^{K-i} \tag{14}$$

It can be found that KNN is more concerned about the inherent polymerization of categories, so KNN classification method usually has a satisfying result [35].

5.3 Important experimental result

5.3.1 Considering whether the feature word class infinity is benefit to classification

Table 1 shows the data comparative relationship of precise, recall and F1 of CHI, IG and centrifuge machine using the experimental data set of this paper, separately utilizing Native Bayes and KNN. Figure 3 described the classification effect

Table 1 Data comparative relationship of CHI, IG and centrifuge machine using different classification methods

Features selected method	Categories					
	Native Bayes			KNN		
	Precise	Recall	F1	Precise	Recall	F1
CHI	0.692	0.735	0.713	0.801	0.761	0.780
IG	0.712	0.782	0.745	0.839	0.816	0.827
CM	0.775	0.814	0.794	0.861	0.858	0.860

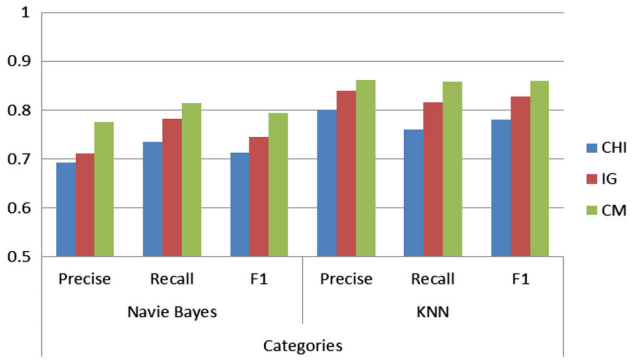

Fig. 3 F1 testing value of CHI, IG, CM in Navie Bayes/KNN (column chart)

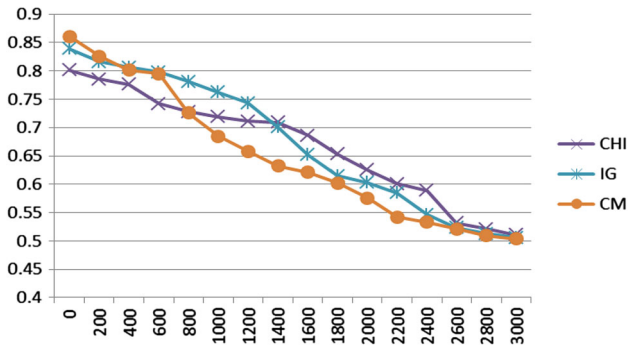

Fig. 4 The effect high category affinity feature words to classification effect

of three feature selection methods: CHI, IG and centrifuge machine using two classification methods. It can be seen in table1 that the classification effect of centrifuge machine method is obviously better than that of IG and the effect of CHI is worst. It can been seen that the classification effect of three feature selection methods in KNN is better than Native Bayes classification method.

For testing whether feature word category affinity influences classification effect, the experiment below is designed in this paper. Delete successively the features whose weight have biggest value in the feature set selected by centrifuge machine method and meanwhile delete the feature T_f in the feature set selected by CHI and IG (If it is included, delete; otherwise, do not delete other features). Figure 4 shows the classification effect of CHI, IG and centrifuge machine using KNN after successively deleting the biggest feature word of category affinity. From the figure, it can been seen that in this case the classification effects of three feature selection methods using KNN are all influenced.

5.3.2 What effect feature dimension to classification

Figure 5 shows the classification effect of feature dimension to CHI, IG and CM using two classification methods. From Fig. 5 we can see that when the feature dimension reduce, centrifuge model approach performance curve decline was significantly lower than CHI and IG.

5.3.3 The effect low-frequency feature word to classification

This experiment successively delete the low-frequency feature word set (maked as $T_{f)}$ from selected feature set by CHI, and meanwhile delete the part of features T_f (if included delete, otherwise not delete other features) from selected feature set by IG and CM. Figure 6 shows the classification effect of three feature selection methods using KNN after deleting different qunantity of low-frequency feature words. From Fig. 6 we can see it influence the classificaiton effect when deleting different qunantity of low-frequency feature words. There into, the influence of CHI is the most, IG follows and CM is the least.

5.3.4 What effect sample uneven to classification

To test what effect sample uneven to classification, the sample distribution that data centralize is adjusted in experiment. The experiment took the ratio of small sample category (the quantity of samples in category ≤ 10) represent the sample distribution. Figure 7 shows the effect CHI, IG and CM using two-classification algorithms to classification in the different sample distribution. From Fig. 7 we can see the influence of sample uneven to IG is the most, CHI follows and CM is the least.

5.4 The experimental result analysis of CM, CHI and IG

Basing on the experimental analysis, it is easy to find that it is worth to consider feature category affinity to classifi-

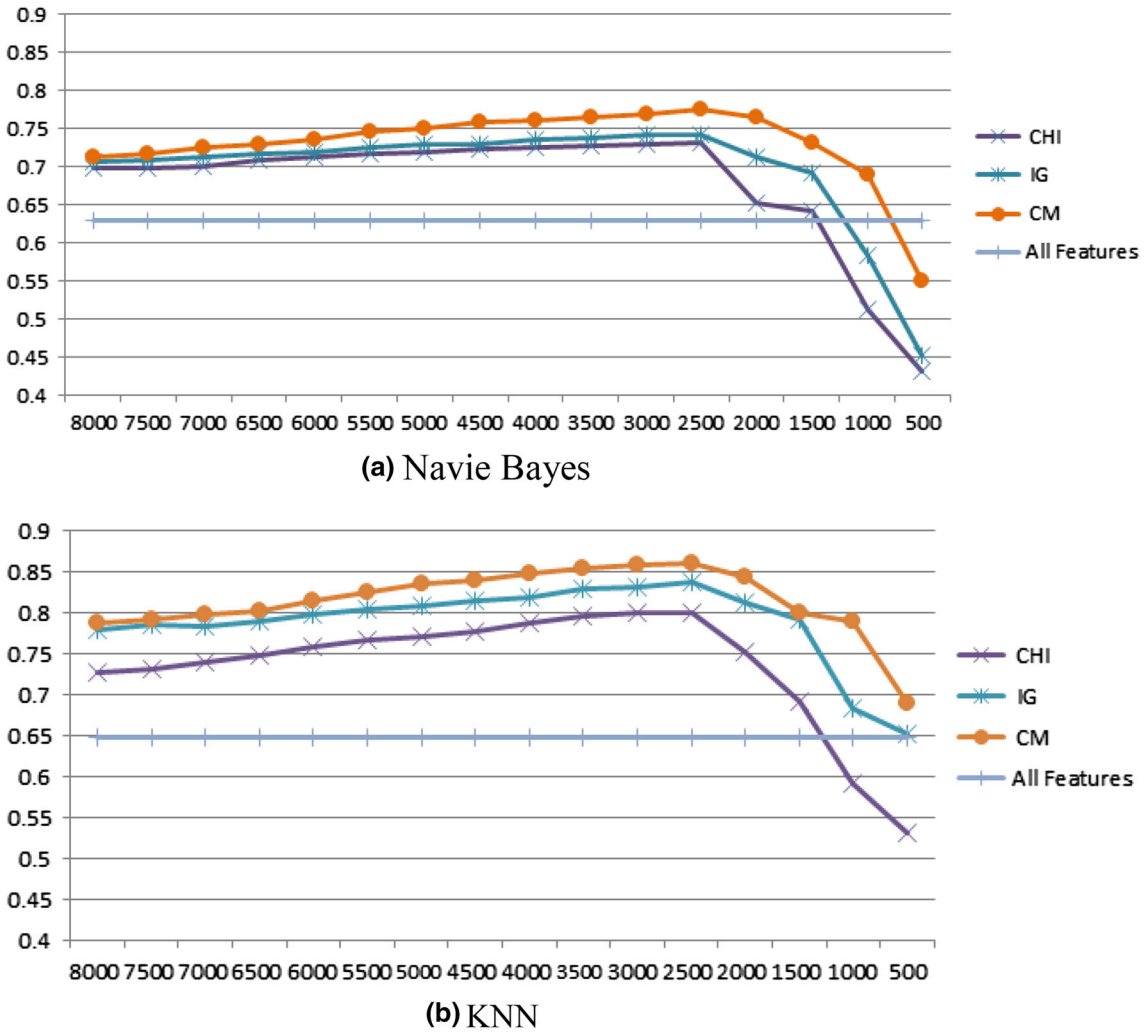

Fig. 5 The effect feature dimension to classification result

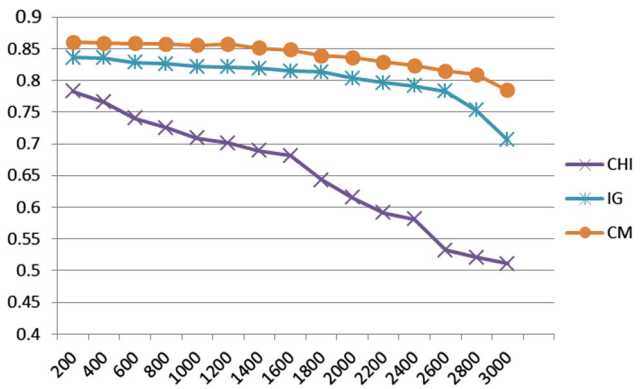

Fig. 6 The effect ab low-frequency feature word to classification

cation. CM farthest matains category shallow semantic in the words semantic. The experiment reault also illustrate the classification effect of CM is better than that of CHI and IG. When the feature dimension reduce, the influence centrifuge model approach is significantly lower than CHI and IG. Meanwhile, from the experimental result, we can see the classification result of CM using KNN is much better than using Native Bayes. This is because KNN classification also considers the category feature aggregated information. KNN algorithm believes that training samples contain the accurate information of its category belongings. That is there is great association relation between feature of the sample and classification category of the sample. While Native Bayes algorithm consider the features of the samples are independent and the relationship of feature and category is also independent. Basing on the designing idea of CM, It fully considers the category affinity of features. The belonging categories of k training samples nearest the testing samples are usually same in KNN algorithm. So the classification effect of CM in KNN classification method is obviously better than that of Native Bayes. The tests low frequency feature words and sample distribution uneven to classification shows that CM is much adaptable to different sample environments.

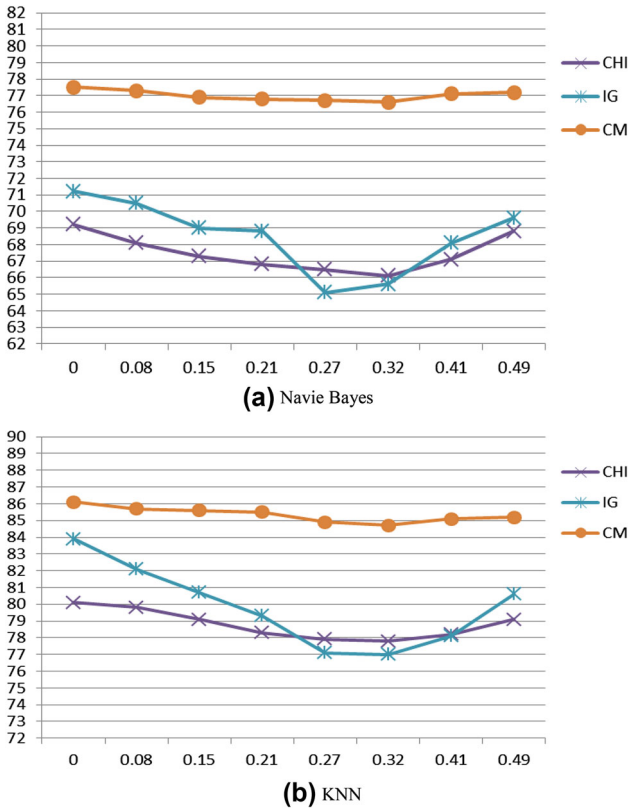

Fig. 7 The effect ab sample uneven to classification

There is serious low frequency feature words defect in CHI. IG performance badly in dealing with "whole feature selection" and sample distribution uneven and CM improves well.

6 Evaluation of the centrifuge model algorithm

Basing on the experimental analysis, a comprehensive analysis and evaluation to centrifuge model is put forward in this paper.

- CM farthest considers the feature difference among categories. So the feature set after feature selection can represent the category difference farthest, which has positive value for classification.
- CM avoids the low frequency feature words defect of CHI, which effectively solving the overall feature selection problem and improving the adaptability of algorithm using uneven sample set.
- CM algorithm can be adaptive to parallel computing, which farthest improve the efficiency of feature selection. Basing on the description of operating mechanism of CM algorithm, the computing process of centrifuge model is mostly adding operations of data. This computing model represents a natural parallel computing character, which can ease the burden of large calculation well and improve calculation efficiency effectively.
- CM can calculate feature weight only in categories, which can decompress document feature preprocess well.
- CM great respect category latent semantic environment, for the core feature set of category contains unquantified latent semantic.
- CM takes category as the unit of considering feature word class discrimination ability and rise the feature selection granularity to category dimension.
- CM algorithm is more complicated, whose whole efficiency of feature selection raise is much limited.

Above all, CM maintained the original semantic of "rough" classification as much as possible and take full account of characteristic differences between categories. The experimental result shows that CM effectively avoid the design defect of traditional feature selection methods basing on statistics and furthermore improve the classification effect. The focus of next work will be how to quantify the extent of semantic loss of the different feature selection methods to categories and systematic study the semantic classification effect of kinds of feature selection methods including CM.

Acknowledgements This work was financially supported by the National Natural Science Foundation of China (61373067,61672301, 61662057), the Science and Technology Innovation Guide Project of Inner Mongolia Autonomous Region of china (2016), the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education Open Foundation (93K172016K05), the Research Program of science and technology at Universities of Inner Mongolia Autonomous Region of china (NJZY16177), the Philosophy and Social Science Planning Project of Inner Mongolia Autonomous Region of china (2015D033), the Natural Science Foundation of Inner Mongolia Autonomous Region of china (2016MS0624), the Program of Science and Technology Development Plan of Jilin Province (20140101195JC).

References

1. Garcia-Torres, M., Gomez-Vela, F., Melian, B., Moreno-Vega, J.M.: High-dimensional feature selection via feature grouping: a variable neighborhood Searc approach. Inf. Sci. **326**, 102–118 (2016)
2. Saeed, F., Salim, N., Abdo, A.: Voting-based consensus clustering for combining multiple clusterings of chemical structures. J. Cheminformatics **4**(1), 1–8 (2012)
3. Wang, Y., Mei, Y.: A multistage procedure for decentralized sequential multi-hypothesis testing problems. Seq. Anal. **31**(4), 505–527 (2012)
4. García, S., Fernández, A., Luengo, J., Herrera, F.: A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Comput. **13**(10), 959–977 (2009)
5. Aliferis, C.: Local causal and markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. J. Mach. Learn. Res. **11**, 171–234 (2010)
6. Gheyas, I.A., Smith, L.S.: Feature subset selection in large dimensionality domains. Pattern Recognit. **43**(1), 5–13 (2009)

7. Berrya, M.W., et al.: Algorithms and applications for approximate on negative matrix factorization. Comput. Stat. Data Anal. **52**, 155–173 (2007)
8. Hanchuan, P., Fuhui, L., Ding, C.: Feature selection based on mutual information criteria of max-dependency max-relevance and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
9. Apte, C., Damerau, F., Weiss, S.: Towards language independent automated learning of text categorization models. In: Proceedings of the 17th Annual ACM/SIGIR Conference, 1994
10. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
11. Salton, G., Wong, A., Yang, C.S.: On the specification of term values in automatic Indexing. J. Doc. **29**(4), 351–372 (1973)
12. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semamtic analysis. J. Am. Soc. Inf. Sci. **1**(6), 391–407 (1990)
13. Frakes, W.B.: Stemming algorithms. In: Information Retrieval: Data Structure & Algorithms, pp. 131–160. TPR Prentice Hall (1992)
14. Hyunki, K., Sushing, C.: Associative naïve Bayes classifier: automated linking of gene ontology to medline documents. Pattern Recognit. **42**(9), 1777–1785 (2009)
15. Joachims, T.: Aprobabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In: Proceedings of the 14th International Conference on Machine Learning, pp. 143–151. Morgan Kaufmann, San Francisco (1997)
16. Dunning, T.E.: Accurate methods for the statistics of surprise and coincidence. Comput. Linguist. **19**(1), 61–74 (1993)
17. Lewis, D.D.: Feature selection and feature extraction for text categorization. In: Proceedings of the Workshop on Speech and Natural Language, pp. 23–26 (1992)
18. John, G.H., Khavi, R., Pfleger, K.: Irrelevant feature and the subset selection problem. In: Proceedings of the 11th International Conference on Machine Learning, New Jersey, pp. 121–129 (1994)
19. Yang Y., Pederson J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning, pp. 412-420. Morgan Kaufmann, Nashville (1997)
20. Mitchell, T.: Machine Learning. McCraw Hill, New York (1996)
21. Koller, D., Sahami, M.: Toeard optimal feature selection. In: Proceedings of the Thirteenth International Conference on Machine Learning (1996)
22. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retieval of Information by Computer. Addison-Wesley, Reading (1989)
23. Ying, C., Jiu-Lin, S.: Research on the automatic classification: present situation and prospects. J. China Soc. Sci. Tech. Inf. **1**, 20–27 (1999)
24. Li, Y.H., Jain, A.K.: Classification of text documents. Comput. J. **41**(8), 537–546 (1998)
25. Lam, W., Ho, C.Y.: Using a generalized instance set for automatic text categorization. In: Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, Melbourne, AU, pp. 81–89 (1998)
26. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines. In: Advances in Kernel Methods-Support Vector learning, pp. 185–208. MIT Press, Cambridge, MA (1999)
27. Apte, C., Damerau, F.J., Weiss, S.M.: Automated learning of decision rules for text categorization. ACM Trans. Inf. Syst. **12**(3), 233–251 (1994)
28. Schapire, R.E., Singer, Y., Singhal, A.: Boosting and Rocchio applied to text filtering. In: Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, Melbourne, AU, pp. 215–223 (1998)
29. Mladenic, D., Brank, J., Grobelnik, M., Milic-Frayling, N.: Feature selection using linear classifier weights: interaction with classification models. In: Jarvelin, K., Allan, J., Bruza, P., Sanderson, M. (eds.) Proceedings of the 27th ACM International Conference on Research and Development in Information Retrieval (SIGIR- 04), pp. 234-24. ACM Press, Sheffield (2004)
30. Aizerman, M., Brave, M.A.N.E., Rozonoer, L.: Theoretical foundations of the Potential function method in pattern recognition learning. Autom. Remote Control **25**, 821–837 (1964)
31. Gil-Garcia, R., Pons-Porrata, A.: Dynamic hierarchical algorithms for document clustering. Pattern Recognit. Lett. (2009)
32. Ng, H.T., Goh, W.B., Low, K.L.: Feature selection, perceptron learning, and a usability case study for text categorization. In: Proceedings of the 20th ACM International Conference on Research Development in Information Retrieval, SIGIR-97, pp. 67–73 (1997)
33. Anaya-Sanchez, H., Pons-Porrata, A., Berlanga-Liavori, R.: A document clustering algorithm for discovering and describing topics. Pattern Recognit. Lett. (2009)
34. Drewes, B.: Some Industrial applications of text mining. Knowl. Min. **185**, 233–238 (2005)
35. Chu, H.-C., Chen, M.-Y., Chen, Y.-M.: A semantic-based approach to content abstraction and annotation for content management. Expert Syst. Appl. **36**(2), 2360–2376 (2009)

Qinghu Wang born in 1983. His main research interests include data mining and multi-relation data mining, cloud computing, cluster computing, big data, software engineering, data structure and computer algorithms.

Lisha Liu born in 1981. Her main research interests include data mining, big data, informatization and transportation informatization.

Jingqing Jiang born in 1968. Her research interests include machine learning and computational intelligence.

Mingyang Jiang born in 1983. His main research interests include data mining, machine learning and deep learning.

Yinan Lu born in 1969. Her main research interests include cover image processing, data mining, computational intelligence and bioinformatics.

Zhili Pei born in 1968. His main research interests include data mining and multi-relation data mining, bioinformatics, pattern recognition theory and application.