International Journal of Hybrid Information Technology
Vol.8, No.9 (2015), pp.257-268
http://dx.doi.org/10.14257/ijhit.2015.8.9.24

Text Sentiment Classification Based on Mixed Cloud Vector Model Clustering and Kernel Fisher Discriminant

Yujuan Xing and Ping Tan

(School of digital media, Lanzhou University of Arts and Science, Lanzhou, 730000, China)
E-mail:xyj19811010@126.com

Abstract

In today's world, the web has dramatically changed the way that people express their opinions. People use the internet to express their opinion, attitude, feeling and emotion about films, goods, news etc. It is challenging to automatically classify mass subjectivity comments into different sentiment orientation categories (e.g. positive/negative). Furthermore, the ambiguity and randomness, which are existed in natural language, lead to lower classification accuracy in text sentiment classification. In this paper, we propose a novel chinese text sentiment classification algorithm based on mixed cloud vector model clustering and kernel fisher discriminant. In this algorithm, we firstly analysis the role of cloud model theory in conversion between qualitative concept and quantitative values, and explore a mixed feature cloud model (MFCM) based on cloud model to represent a single document. In MFCM, both effect of different part-of- speech features and ambiguity of sentiment tendency are considered. And then, documents are clustered according to their similarity between MFCM. Finally, kernel fisher discriminant (KFD) is adopted as the classifier to judge views. The experimental results demonstrate that our proposed method outperforms traditional approaches.

Keyword: *Text sentiment classification; cloud model; kernel fisher discriminant; support vector machine*

1. Introduction

In today's world, text is still the dominant means of communication on the internet despite the rising popularity of other media such as speech, video, and images. People continue to use text as a means to express their opinions, emotions and thoughts via blogs, emails, and comments to articles. With the rapid development of web technology, the comments in blog, microblogging, forums and shopping sites continue to grow increasingly. How to exact sentiment information from mass web data and determine sentiment views (e.g. positive/negative) become the research hotspots. Text sentiment classification is an important branch of text sentiment analysis. It is process of exploring user comments on the web to judge the overall opinion or feeling in reviews [1-3]. Firstly, sentiment keywords about related information are selected, and then views are determined according to these keywords. Therefore, the information selection and views decision is the key problems in sentiment classification.

Vector space model (VSM) was proposed by G Salton [4] to present a set of documents as feature vectors in a common vector space. In VSM, a document is conceptually represented by a vector of keywords extracted from the document. These keywords are related to weights which represent the importance of the keywords in the document. The weight of a keyword in a document vector can be determined in many ways, such as Boolean value, term frequency (TF) and inverse document frequency (IDF)[5,6]. The proposal of VSM makes various machine learning methods apply in sentiment classification easily [7,8,9]. Support vector machine (SVM) having excellent

ISSN: 1738-9968 IJHIT

classification performance is applied in sentiment classification widely. Jun LI [10] adopted SVM, Naïve Bayes (NB), Maximum Entropy (ME) and artificial neural networks (ANN) as sentiment classifiers in Chinese reviews from website. Experimental evaluation showed that machine learning method had better performance, especially SVM. Khin Phyu [11] combined ontology based on Formal Concept Analysis (FCA) design to SVM for classifying the software reviews are positive, negative or neutral. ZHU Jian [12] used individual model (i-model) based on artificial neural networks (ANN) to determine sentiment views. Experimental results showed that the accuracy of individual model is higher than that of support vector machines and hidden Markov model (HMM) classifiers on movie review corpus. Rudy Prabowo [13] combined rule-based classification, supervised learning and machine learning into a new combined method. The ideal experiment results were achieved on movie reviews, product reviews and MySpace comments. Rodrigo Moraes [14] carried out a detailed analysis and comparison of SVM and ANN in document-level text classification.

However, the original vectors have amazing dimensions in vector space model. This problem leads to a challenge in processing of text mining [15]. We look for useful subset of low-dimensional vectors to improve the accuracy and speed of machine learning significantly. At present, evaluation function is adopt as the common feature select method, such as information gain (IG), expected cross entropy, mutual information (MI) etc[16]. Suge Wang [17] proposed a feature selection method based on improved fisher's discriminant ratio for sentiment classification. The experimental results verified the effectiveness of the proposed method. Zhi-Hong Deng [18] proposed a supervised term weighting scheme that considered importance of a term in a document and importance of a term for expressing sentiment respectively, to improve the performance of sentiment analysis.

Simultaneously, the vagueness and randomness of text will affect the judgment of text view [19]. For example, the human description of color, sound, smell, humidity, as well as the words "occasionally", "probably" etc. have vagueness and randomness. How to convert qualitative concept and quantitative values is the challenge in natural language processing. Li Deyi[20-23]proposed a novel cloud model for this conversion. Cloud model has been widely used in many fields to solve the vagueness and randomness in data. Qian Fu[24] used cloud model for reliability evaluation. The experimental results show that the reliability evaluation result is closer to the reality when using the cloud model that considering the uncertainty of the most typical sample. Ren-Long Zhang[25] proposed a novel fuzzy hybrid quantum artificial immune clustering algorithm based on cloud model(C-FHQAI) to solve the stochastic problem. Xiaojun Yang[26] proposed a new representation model based on cloud model for a word from interval-valued data to solve the interpersonal and intrapersonal uncertainties existed in the process of constructing the model of a word by collecting interval-valued data from a group of individuals.

Scholars have done a lot of research work in feature selection and classification algorithms. In this study, we propose a cloud vector model clustering (CVMC) algorithm based on mixed part-of-speech features, and adopt kernel fisher discriminant (KFD) as classifier to judge the text views. Our major work is as follows.

(1) Information Gain (IG) and Mutual Information (MI) are combined to features of different part-of-speech respectively to select low-dimensional and discriminant mixed features.

(2) For the sake of resolving the problems of uncertainty, fuzziness and randomness existed in natural language, we establish cloud vector models based on mixed features to represent documents.

(3) We design the clustering algorithm according to the similarity between cloud vector models. By doing so, the decrease of model numbers helps us to reduce the computational complexity of classification algorithm.

(4) We adopt KFD as classifier to decide the sentiment orientation categories (e.g. positive/negative).

The remainder of the paper organized as follows: Section 2 provides a review of cloud model theory. Section 3 provides our proposed method detailed. Section 4 includes an experimental evaluation of the proposed clustering and classification method in comparison with existing common techniques. Finally, Section 5 outlines conclusions and future directions.

2. Cloud Model Theory

Cloud model (CM) is a conversion between qualitative concept and quantitative values, proposed by De-yi. Li [21]. Suppose U as a quantity domain expressed with accurate numbers, and C is qualitative representation in U. If quantitative value $x \in U$ and x is a random realization of the quality concept C, random value $\mu(x) \in [0,1]$ which has stable tendency is the certainty degree of x to C : $\mu : U \to [0,1], \quad \forall x \in U, x \to \mu(x)$. Then the distribution of x in U is called cloud and each x is a cloud droplet. If U is n-dimensional space, it can be extended to n-dimensional cloud.

The integrity of cloud concept can be represented by three numerical characteristics, namely Ex (expectation), En (entropy) and He (hyper entropy), denoted cloud vector $\vec{C}(Ex, En, He)$. Ex indicates the expected distribution of cloud droplet in U. En is used to depict the uncertainty of samples in the concept. He reflects the thickness and dispersion of cloud, used to measure the uncertainty of En. Cloud models make it possible to get the distributing range of a qualitative concept. For example, how to judge the age of "young" is fuzzy. In [27], Dai Jin proposed a cloud model referred to "young" as showed in figure1.

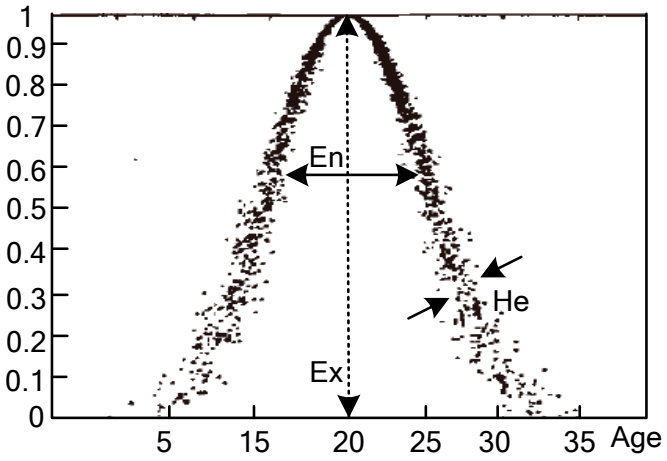


Figure 1: The Cloud Model of "Young"

Meanwhile, two cloud transformations, namely forward cloud transformation (FCT) and backward cloud transformation (BCT), are used to realize the bidirectional cognitive transformation between quantitative date and qualitative concept. FCT is used to implement the transformation from qualitative concept to quantitative date, and BCT realizes the transformation from quantitative date to qualitative concept.

3. Sentiment Classification Based On CVMC and KFD

In this paper, we design a novel document digital model and clustering algorithm by means of conversion role of BCT, and adopt KFD as classifier to make decision.

3.1. Mixed-Feature Selection

The part-of-speech feature of text gets wide application in sentiment analysis and opinion mining because of its excellent disambiguation in polysemous word. The basic part-of-speech includes Noun (N), Verb (V), adjective (A), adverb (D), pronouns (R), prepositions (P), idioms (I), idioms (L) and conjunctions (C) etc[28]. In this paper, Noun(N), Verb (V), adjective (A) and adverb (D) having stronger sentiment characteristics are selected due to considering the consuming computation of cloud vector model(CVM). We define the feature combination "N+D+A", "D+A", "A" as $f(1)$, $f(2)$ and $f(3)$ respectively. Since these three features appeared in document is more, we utilize Information Gain (IG) and Mutual Information (MI) to select low-dimensional feature.

3.1.1. Mixed-Feature Based On IG

IG is regularly employed as measure of term entropy according to the presence or absence of a term in a document. Suppose a term as t, $IG(t)$ can be define as follows.

$$IG(t) = H(D) - H(D \mid t) = \sum_{d \in D} (P(d,t)\log(\frac{P(d,t)}{P(d)P(t)}) + P(d,\bar{t})\log(\frac{P(d,\bar{t})}{P(d)P(\bar{t})})) \quad (1)$$

where d is the category of document, D represents the set of documents, $H(D)$ is system entropy. $H(D \mid t)$ indicates system entropy under the conditions of choosing t. Clearly seen by the formula (1), $IG(t)$ indicates that the contribution of t for classification. In our research, we compute the information gain for Noun (N), Verb (V), adjective (A) and adverb (D) respectively, and select foremost q terms having biggest information gain value as final features. So, the three mixed feature can be denoted as $f_{IG}(1)$, $f_{IG}(2)$ and $f_{IG}(3)$.

3.1.2. Mixed-Feature Based On MI

Mutual information is a measure method of feature information on the basis of correlation between random variables. Suppose document category as d, the mutual information between term t and document category d can be define as formula (2).

$$MI(d,t) = \log(\frac{P(d,t)}{P(d)P(t)}) \quad (2)$$

In formula (2), if term t is unrelated to current document c, $MI(d,t) = 0$. We adopt the average mutual information $MI_{avg}(t) = \sum_{d \in D} P(d)MI(d,t)$ as MI measure threshold. Formula (2) also shows that MI is proportional to correlation of term and category. And the same as IG, we only compute the mutual information for Noun (N), Verb (V), adjective (A) and adverb (D) respectively, and choose the former features having largest MI values as final features. Three mixed feature can be denoted as $f_{MI}(1)$, $f_{MI}(2)$ and $f_{MI}(3)$.

3.2. Cloud Vector Model Clustering (CVMC)

3.2.1. Generation Of Cloud Vector Model

Suppose m documents as $D = \{D_1, D_2, \cdots, D_m\}$, where $D_i (i = 1, 2, \cdots, m)$ indicates ith document in D and $d_i = \{w_{i1}, w_{i2}, \cdots, w_{in}\}$ represents feature set of document D_i, w_{ij} is

jth feature. If w_{ij} is regarded as cloud droplet, we can compute Ex, En and He of D_i by BCT algorithm as follows.

$$Ed_i = \frac{1}{n}\sum_{j=1}^{n} w_{ij} \quad (3)$$

$$En_i = \sqrt{\frac{\pi}{2}} \times \frac{1}{n}\sum_{i=1}^{n} \left| w_{ij} - Ed_i \right| \quad (4)$$

$$S_i^2 = \frac{1}{n-1}\sum_{i=1}^{n} (w_{ij} - Ed_i)^2 \quad (5)$$

$$He_i = \sqrt{S_i^2 - En_i^2} \quad (6)$$

Obviously, the ith document D_i can be denoted as cloud vector $\vec{C}_i = (Ed_i, En_i, He_i)$. And then the document set is digitized into cloud vector model (CVM), denoted as $D = \{C_1, C_2, \cdots, C_m\}$.

3.2.2. Clustering

Inspired by the diversity measure in vector space model, we define the similarity between D_i and D_j as formula (7).

$$\cos\theta = \frac{\vec{C}_i \bullet \vec{C}_j}{\left\|\vec{C}_i\right\|\left\|\vec{C}_j\right\|} \quad (7)$$

Where $\vec{C}_i$ and $\vec{C}_j$ indicate the cloud vectors of D_i and D_j respectively. However, document set is larger and has similar documents, these lead to consuming computation of classifier. For the sake of solving these problems, we proposed a novel clustering algorithm as follows based on similarity of document CVMs.

Table 1: CVMC Algorithm

Step1: Set the number of categories as K, assign K cloud vectors to initialize the cluster center randomly. $R_s = 0$ ($s = 1, \cdots, K$) is used to record the number of clustered documents in current category s;

Step2: Compute similarity between cloud vectors $\vec{C}_i$ ($i = 1, \cdots, N$) and cluster center cloud vector C_s ($s = 1, \cdots, K$) by formula (7);

Step3: Select cloud vector $\vec{C}_i = (Ed_i, En_i, He_i)$ having the smallest similarity, and merge it into current category s whose cluster center cloud vector is $\vec{C}_s = (Ed_s, En_s, He_s)$, $R_s = R_s + 1$, compute the new cluster center $\vec{C}_{sn} = (Ed_{sn}, En_{sn}, He_{sn})$ as following formulas:

$$Ed_{sn} = \frac{Ed_i En_i + Ed_s En_s}{En_i + En_s}$$

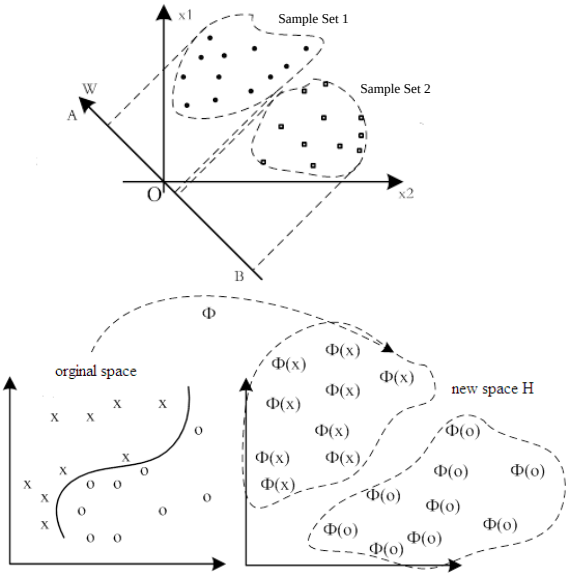
$$En_{sn} = En_i + En_s$$

$$He_{sn} = \frac{He_i En_i + He_s En_s}{En_i + En_s}$$

Step4: Repeat step2 and step3 until the cluster center cloud vector changes no longer.

3.3. Classifier Based On KFD

Kernel fisher discriminant (KFD) [29] is an extension of Fisher linear discriminant (FLD). Its basic idea can be described that input space is mapped into some feature space (usually nonlinear space) H using a nonlinear function column vector $\Phi(g)$, consequently FLD seeks a direction, which can maximize the ratio of the inter-class scatter matrix of the projected samples to the intra-class scatter matrix of the projected samples so as to separate the objects from different classes as much as possible in the feature space H. Figure 2 shows the schematic diagram of FLD and KFD.



a) **FLD** **b)KFD**

Figure 2: The Schematic Diagram Of FLD And KFD

KFD use all the training samples rather than some special samples such as support vector, so KFD is superior to SVM in some aspects. However, with the increase of number of samples, the computational complexity of KFD will be become larger. In this paper, the number and dimension of features are reduced entirely by the proposed mixed feature select method and cloud vector model clustering algorithm. The problem of high computational existed in KFD is solved precisely.

We suppose that there are K cluster center cloud vectors $\vec{C}_s = (Ed_s, En_s, He_s), s = 1, 2, \cdots, K$ which are got from CVMC. $C = \{C_1, C_2, \cdots, C_s\}$, $R_1 = \{K_1$ cloud vectors of positive view $\}$ and $R_2 = \{K_2$ vectors of negative view$\}$ $K_1 + K_2 = K$.

In the space H, the corresponding object function is defined as:

$$J(w) = \frac{w^T S_b^{\Phi} w}{w^T S_w^{\Phi} w} \tag{11}$$

where

$$S_b^{\Phi} = (\mu_1^{\Phi} - \mu_2^{\Phi})(\mu_1^{\Phi} - \mu_2^{\Phi})^T \tag{12}$$

$$S_w^{\Phi} = \sum_{i=1}^{2} \sum_{\vec{C} \in C} (\Phi(\vec{C}) - \mu_i^{\Phi})(\Phi(\vec{C}) - \mu_i^{\Phi})^T \tag{13}$$

are the inter-class and intra-class scatter matrices respectively. $\mu_i^{\Phi} = \frac{1}{K_i}\sum_{j=1}^{K_i}\Phi(\vec{C}_j)$ $i = 1, 2$ is mean vector of input cluster center vector $\vec{C}$ and w is projection direction.

According to kernel trick [30] and generalized rayleigh entropy, we could maximize equation (11) to find the optimal projection direction w of a testing sample $\Phi(\vec{C})$ in feature space H. It can be computed by $w\Phi(\vec{C}) = \sum_{i=1}^{K} a_i K(\vec{C}_i, \vec{C})$. By using a linear support vector machine (SVM) to determine an optimal threshold, the discriminant decision function is obtained as following:

$$f(x) = \mathrm{sgn}[w^T\Phi(\vec{C}) + b] = \mathrm{sgn}[\sum_{i=1}^{R} a_i K(\vec{C}_i, \vec{C}) + b] \quad (14)$$

When KFD classifier is used for final decision, each cloud vector possesses an optimal projection direction. In the optimal projection direction, positive cloud vector will be distinguished from negative vector completely.

4. Experiments

In order to examine the effect of our proposed mixed feature select method, CVMC and KFD classifier, we use chinese reviews collected by Dr. Tan [31]. In this corpus, there are three kinds of reviews about hotel, computer and book. The positive/ negative of each kind have 2000 views. In our experiment we select 1200 hotel views both positive and negative randomly to train the cloud vector model, and the rest is used to test. We utilize ICTCLAS chinese analysis system to preprocess the views, and don't consider the affect of punctuation and auxiliary word. The sample information is showed as Table 2.

Table 2: Experimental Example Information

corpus		Training	Testing
Hotel views	positive	1200	800
	negative	1200	800

Experiment 1：Performance comparison of different Part-of-speech combination

In this experiment, we test performance of classifier such as KFD, SVM and Naïve Bayes (NB) based on features of different part-of-speech combination. The results are showed as table3. The evaluation index is correct classification number (CCN), classification accuracy (CA) and Time consuming (TC). The number of testing views is 1600.

Table 3: Performance Comparison Of Different Part-Of-Speech Combination

Part-of-speech combination	Dimensions	KFD			SVM			NB		
		CCN	CA (%)	TC	CCN	CA (%)	TC	CCN	CA (%)	TC
$f^{(1)}$ (N+D+A)	8137	1372	85.75	6.17s	1279	79.94	5.33s	1202	75.13	5.82s
$f^{(2)}$ (D+A)	1329	1429	89.31	4.21s	1324	82.75	3.92s	1292	80.75	4.37s
$f^{(3)}$ (A)	824	1245	77.81	3.29s	1126	70.38	3.03s	1097	68.56	3.49s

Table 3 shows that:

(1) “D+A” shows best performance in classification algorithms. Though its dimension is far less than “N+D+A”, it is superior to “N+D+A” in classification. The classification accuracy of “D+A” is 89.31% in KFD classifier. However, fewer features are existed in adjectives, and different combination of adjective and other part-of- speech has different sentiment tendency. So, “A” has lower classification accuracy.

(2) KFD has most excellent performance both in three different part-of- speech combination. In “N+D+A”, its CA is higher than SVM the nearly 6 percentage points, and higher than NB 10.62%. In “D+A”, its classification accuracy is 89.31%, higher than SVM 6.56% and higher than NB 8.56%. In “D+A”, it also has optimal classification accuracy. However, the time consuming of KFD is largest. In “N+D+A”, its time consuming is even to 6.17s.

Experiment 2：Performance comparison of mixed feature

In experiment 1 we test four classifiers, KFD has optimal performance. However, KFD has disadvantage of high computational complexity. In this paper, we adopt IG and MI as feature selection method to select more discriminant and low-dimensional mixed features. We adopt three kinds of classical evaluation measures generally used in text classification, Precision , Recall and F_value to evaluate the effectiveness of mixed feature combinations. By PP (PN), RP (RN) and FP (FN) we denote Precision, Recall and F_value of positive (negative) subjectivity views respectively. The experiment results are shown as table 4 and Figs. 3-5.

Table 4: Performance Comparison Of Mixed Feature

Dimensionality	Mixed feature	PP (%)	RP (%)	FP (%)	PN (%)	RN (%)	FN (%)	Accuracy(%)
4193	$f_{IG}(1)$	80.20	77.55	78.85	76.54	79.28	77.89	78.38
	$f_{MI}(1)$	77.44	81.79	79.56	79.91	75.24	77.51	78.58
1006	$f_{IG}(2)$	89.56	87.20	88.37	87.17	89.53	88.34	88.35
	$f_{MI}(2)$	86.82	87.27	87.05	86.43	85.95	86.19	86.63
611	$f_{IG}(3)$	75.45	69.65	72.44	68.40	74.35	71.25	71.86
	$f_{MI}(3)$	66.25	64.01	65.11	60.36	62.69	61.50	63.39

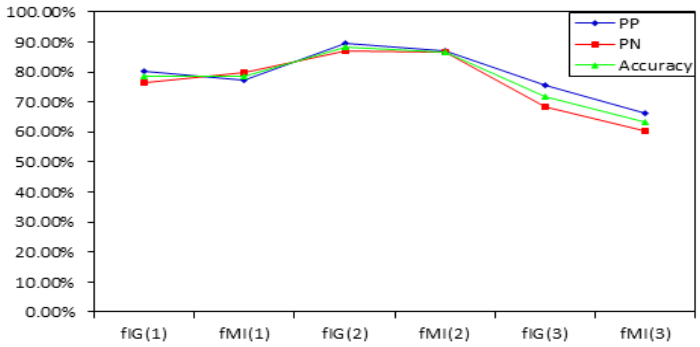


Figure 3.The Curve Of Precision And Accuracy Of Mixed Feature Selection Methods

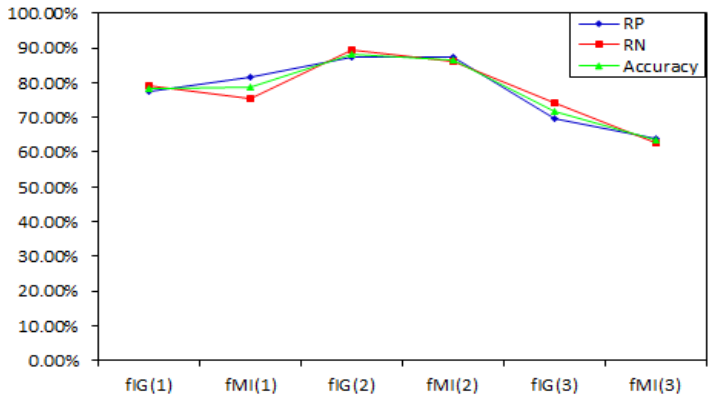


Figure 4: The Curve Of Recall And Accuracy Of Mixed Feature Selection Methods

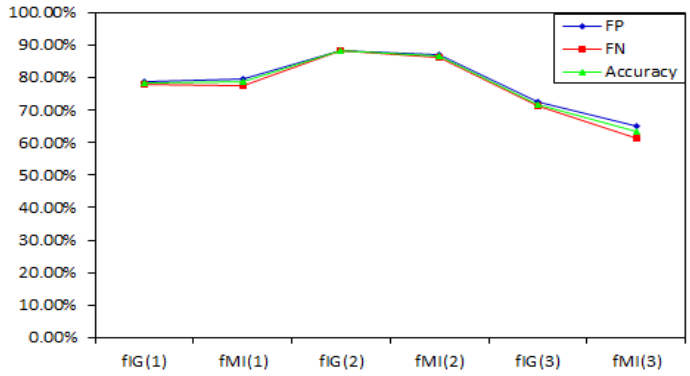


Figure 5: The Curve Of F_Value And Accuracy Of Mixed Feature Selection Methods

From Table 3, Figures 2–4, we can see that IG is superior to MI in three kinds of speech features. This is mainly due to the lack of consideration of terms in different documents in MI. And results also verify that the part-of-speech combination "A+D" is efficient compared to "A" and "A+N+D". In summary, feature selection method and speech feature have a crucial role in text sentiment classification.

Experiment 3: Analysis and comparison on CVMC

In this experiment, we mainly focus on testing the performance of CVMC using features based on $f_{IG}(2)$ and adopting KFD as classifier. The experiment results are shown as table 5 and Fig 4.

Table 5: Performance Comparison of CVMC

Number of clusters（K）	Correctly classified documents	Accuracy（%）
2000	1413	88.31
1800	1432	89.50
1500	1476	92.25
1200	1322	82.63
1000	1209	75.56
800	1008	63.00
500	873	54.56

From table 5 and Figure 6, we can easily see that the accuracy is highest when the number of clusters is 1500. With the decrease of cluster numbers, the training features are reduced. By doing this, the classification accuracy of KFD is affected. The number of training views is reduced to 1500 by CVMC, which is 2400 originally. The reduction ratio is 37.5%. So, CVMC can saves storage space effectively, reduces the computational complexity of KFD and improves classification accuracy.

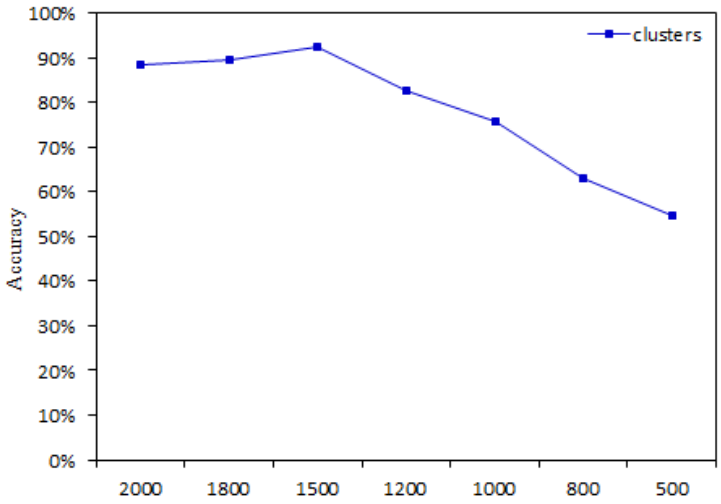


Figure 6: Curve of CVMC

5. Conclusions

Feature selection and classifier design are the key issues in text sentiment classification. However, the mass features, fuzziness and randomness in text lead to inefficient classification accuracy and high computational complexity. In this paper, we propose a novel cloud vector model clustering method based on mixed part-of-speech feature combination, and adopt kernel fisher discriminant as classifier. We use IG and MI to generate mixed part-of-speech feature combination, at the same time the dimensions of features are reduced. And then we construct cloud vector models of documents based on reduced mixed part-of-speech feature vectors. Using the similarity between cloud vector models, we cluster the models. Through the above process, the numbers of documents and dimensions of features are reduced entirely. It solves the problems of slow training in large-scale data existed in KFD. In order to validate the validity of the proposed method, we design three experiments to test performance comprehensively. The results of experiment 1 show that "D+A" has more sentiment tendency, and KFD has excellent classification performance. In experiment 2, we can conclude that IG is superior to MI in mixed feature selection. And in clustering method experiment, the results show that the accuracy achieves 92.25% when the number of clusters is 1500.

References

[1] Deyu Li, Suge Wang, Lidong Zhao and Jiahao Zhang. Sample cutting method for imbalanced text sentiment classification based on BRC. Knowledge-based systems, Vol. 37, pp.451-461**(2013)**
[2] Aimin Yang, Yongmei Zhou and Jianghao Lin. A Method of Chinese Texts Sentiment Classification Based on Bayesian Algorithm. International Conference on Information Technology and Management Innovation, Guangzhou, China, pp.2157-2162 **(2012)**
[3] Xiaoni Wang, Zhenjiang Zhang and Wei Cao. An Improved KNN Algorithm in Text Classification. International Conference on Information Science and Computer Applications, Changsha, China, pp.268-273**(2013)**

[4] Salton G and Wang A. Yang C S. A vector space model for automatic indexing. Communication of the ACM. Vol.18, pp.613-620 **(1975)**
[5] Tian Xia and Yi Du. Improve VSM text classification by title vector based document representation method. International Conference on Computer Science & Education (ICCSE), pp. 210 – 213 **(2011)**
[6] Man Lan , Tan, C.L. , Jian Su and Yue Lu. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, no. 4, pp. 721 – 735 **(2009)**
[7] Farhoodi and M.,Yari, A. Applying machine learning algorithms for automatic Persian text classification. 2010 6th International Conference on Advanced Information Management and Service (IMS), pp. 318 – 323 **(2010)**
[8] Frunza, O., Inkpen, D and Tran, T.A. Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts. IEEE Transactions on Knowledge and Data Engineering, Vol. 23, no. 6, pp. 801-814 **(2011)**
[9] Sang-Bum Kim , Kyoung-Soo Han , Hae-Chang Rim and Sung Hyon Myaeng. Some Effective Techniques for Naive Bayes Text Classification. IEEE Transactions on Knowledge and Data Engineering, Vol. 18 , no.11, pp. 1457 – 1466 **(2006)**
[10] Jun LI and Maosong SUN: Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques. Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering. Beijing, pp.393 -400 **(2007)**
[11] Khin Phyu Phyu Shein and Thi Thi Soe Nyunt: Sentiment Classification based on Ontology and SVM Classifier. 2010 Second International Conference on Communication Software and Networks. pp.169-172, Singapore **(2010)**
[12] ZHU Jian, XU Chen and WANG Han-shi: Sentiment classification using the theory of ANNs. The Journal of China Universities of Posts and Telecommunications. Vol.17, pp.58-62 **(2010)**
[13] Rudy Prabowo1 and Mike Thelwall: Sentiment analysis: A combined approach. Journal of Informetrics. Vol. 3,pp.143-157**(2009)**
[14] Rodrigo Moraes, João Francisco Valiati, Wilson P and Gavião Neto. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, Vol. 40,pp.621-633 **(2013)**
[15] Rogati M and Yang Y. High-Performing Feature Selection for Text Classification. In Proceedings of the 11th ACM International Conference on Information and Knowledge Management (CIKM-02). McLean: ACM Press,pp.659-661 **(2002)**
[16] Yang Y and Pedersen JO. A Comparative Study on Feature Selection in Text Categorization. Proceedings of the 14th International Conference on Machine Learning. Nashville: Morgan Kaufmann Press, pp.412-420 **(1997)**
[17] Suge Wang, Deyu Li, Xiaolei Song, Yingjie Wei, and Hongxia Li: A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Systems with Application. Vol. 38, pp.8696-8702 **(2011)**
[18] Zhi-Hong Deng, Kun-Hu Luo and Hong-Liang Yu: A study of supervised term weighting scheme for sentiment analysis. Expert Systems with Applications, Vol. 41, pp.3506–3513 **(2014)**
[19] De-yi. Li. Knowledge representation and discovery based on linguistic atoms. In: Proceedings of the 1st Pacific2Asia Conference. pp.3- 20, Singapore**(1997)**
[20] LI DY. Uncertainty in knowledge representation. Engineering Sciences, Vol. 10, pp.73-79 **(2000)**
[21] Li D Y, Di K C and Li D. Mining association rules with linguistic cloud models. In: PAKDD'98 Proceedings of the Second Pacific-Asia Confon Knowledge Discovery and Data Mining. Melbourne, pp.392–394 **(1998)**
[22] Li Deyi and Liu Changyu, Study on the universality of the normal cloud model. Engineering Science, Vol. 6, no. 8, pp.29–33 **(2004)**
[23] Li Deyi, Meng Haijun and Shi Xuemei. Membership cloud and membership cloud generator. Computer Research and Development, Vol. 32, no. 6, pp.15–20 **(1995)**
[24] Qian Fu, Zhi-hua Cai and Yi-qi Wu. A New Method for Reliability Evaluation Based on Cloud Model. 2010 Second International Conference on Information Technology and Computer Science (ITCS), pp. 118-121 **(2010)**
[25] Ren-Long Zhang, Mi-Yuan Shan, Xiao-Hong Liu and Li-Hong Zhang. A novel fuzzy hybrid quantum artificial immune clustering algorithm based on cloud model. Engineering Applications of Artificial Intelligence, Vol. 35, pp.1-13 **(2014)**
[26] Xiaojun Yang, Liaoliao Yan, Hui Peng and Xiangdong Gao. Encoding words into Cloud models from interval-valued data via fuzzy statistics and membership function fitting. Knowledge-Based Systems, Vol. 55, pp.114-124 **(2014)**
[27] Dai Jin. Research on Key Problems in Text Mining Based on Cloud Method. Chongqing: Chongqing University **(2011)**
[28] Turney Peter D and Littman Michael L.. Measuring praise and criticism: inference of semantic orientation from association. ACM Transactions on Information Systems, Vol. 21,no. 4,pp.315-346 **(2003)**

[29] Xiang, C., Fan, X.A., Lee and T.H: Face recognition using recursive Fisher linear discriminant. IEEE Transactions on Image Processing, Vol. 15, pp.2097 – 2105**(2006)**
[30] Shu Yang, Shuicheng Yan and Chao Zhang. Bilinear Analysis for Kernel Selection and Nonlinear Feature Extraction. IEEE Transactions on Neural Networks, Vol. 18, pp.1442-1452**(2007)**
[31] Tan, S. B.and Zhang, J.. An Empirical study of sentiment analysis for chinese documents. Expert Systems with Application. Vol. 34, 2622–2629**(2008)**

Authors

Yujuan Xing

Sex: Female
Date of Birth: Sep. 13th, 1981
Native Place: Tianshui Gansu, P.R.China
Educational Attainment: master degree
Professional Title: associate professor
Tel: +8613893466082
E-mail: xyj19811010@126.com
Research Interests: natural language processing

Ping Tan

Sex: Female
Date of Birth: Jul. 14th, 1981
Native Place: Lanzhou Gansu, P.R.China
Educational Attainment: master degree
Professional Title: associate professor
Tel: +8613519609280
E-mail: 429687745@qq.com
Research Interests: natural language processing

