Improved Mutual Information Based on Relative Frequency Factor and Degree of Difference among Classes

Jianwen Gao [a*], Xi Yang[b], Wen Wen[c] and Jihua Yang[d]

Department of Information Engineering ,Engineering University of CAPF,

Xi'an 710086,China

[a]1838175910@qq.com, [b] 1274211197@qq.com, [c]1433128691@qq.com,

[d] 499295386@qq.com

Keywords: feature selection; relative frequency factor; MI; degree of difference among classes

Abstract. By introducing degree of difference among classes, an improved mutual information feature selection method is proposed to effectively improve the accuracy of feature selection, accuracy and efficiency of classification. At the same time, relative frequency factor is applied to solve the tendency of traditional methods to choose the shortage of low frequency words. The experimental results show that the improved method can reasonably improve the performance of mutual information feature selection.

1 Introduction

The rapid development of Internet technology enables online data to have explosive growth, and all of these data mainly exists in the form of text. In the text automatic classification, documents are often converted into models so as to better computer processing. But the high dimension of the feature space and data sparseness lead to increase of computing time and lowered efficiency in the process of text representation, which may have impacts on the accuracy of classification. Therefore, to reduce the dimension of original feature space and increase accuracy of classification become the difficulties of text automatic classification. At present, methods that are used in dimension reduction are feature extraction an feature selection[1].

Feature selection means to select out those features sets with strong communicative ability and greater contribution rate of classification in the integration of original feature items[2]. Currently, the common used selection methods are TF-IDF, Information Gain(IG)[3], Mutual Information(MI)[4], Chi-square Test(CHI)[5], Weight of Evidence for Text(WET), Expected Cross Entropy(ECE), etc. Literature[6] through experimental research shows that MI methods have relatively low effects in feature selection, because traditional MI methods do not take feature items of document frequency of different classifications into consideration, nor did it consider the word frequency of different documents in the same classification, and is prone to select low-frequency words in the selection process.

In this paper , through introducing degree of difference among classes, puts forward an improved feature selection method of MI, and introduces relative

© 2016. The authors - Published by Atlantis Press

frequency factor to deal with feature selection method which is prone to select the deficiency of low word frequency.

2 Studies on MI Method

2.1 Feature Selection Methods of Mutual Information

According to the appearance possibility of classification c_j and feature item t_i ,MI means the measurement of relevant degree between them[7]. In text classification, if the total number of document sets is N, the classification integration will be $\{c_1, c_2, ..., c_j, ..., c_m\}$, the feature items integration is $\{t_1,t_2,\ldots,t_i,\ldots,t_n\}$, and the computational formula of mutual information of classification c_j and feature item t_i is[8] :

$$MI=(t_i,c_j)=\log\frac{p(c_j,t_i)}{p(c_j)\times p(t_i)}=\log\frac{p(t_i/c_j)}{p(t_i)}, \quad (1)$$

in the formula: $p(c_j)$ means the appearance possibility of classification c_j ; $p(t_i)$ means appearance possibility of feature item t_i in the document integration; $p(c_j,t_i)$ means the simultaneous appearance possibility classification c_j and feature item t_i ; $p(c_j/t_i)$ means the appearance possibility of feature item t_i in the classification c_j .

From formula (1), the lower frequency of feature item t_i in the document sets and the higher frequency of classification c_j , the bigger mutual information value, which means stronger relevant degree between them. If feature item t_i did not appear in classification c_j , the mutual information value would be 0.

Considering that feature item may be distributed in other classification, in order to acquire feature item t_i in the average mutual information value of the whole text, the computational formula will be[9]:

$$\overline{MI(t_i,c_j)}=\sum_{j-1}^{m}p(c_j)\log\frac{p(t_i/c_j)}{p(t_i)}, \quad (2)$$

in the formula: m means the number of classification.

2.2 Analysis on The Mutual Information Methods

From analysis formula (1) and (2), elements that decide the size of the mutual information value only relies on the frequency number of feature item and the appearance frequency number of the whole text, so the results of the computation exist following deficiencies.

Tending to select low frequency words. In classification c_j , when feature item $p(t_1)>p(t_2)$ and $p(t_1/c_j) <p(t_2/c_j)$, we could acquire $MI(t_1,c_j) <MI(t_2,c_j)$,from formula (1). According to the ranking of mutual information value, t_1 was listed in the last. In the final threshold value selection, t_1 was eliminated. But for classification c_j , high-frequency word t_1 carries more information, and is more good at expressing text content.

2) There exists some identical mutual information value of some feature items, but feature items listed behind are easily eliminated, which may cause the loss of some valuable information.

3) In different classification, feature item t_1 appears in one or several classifications, and feature item t_2 is uniformly distributed according to different classification. When $MI(t_1, c_j)<MI(t_2, c_j)$ appears in computational mutual information value, feature item t_1 has more representation ability in classification.

4) In the same classification, feature item t_1 appears mostly in rare documents,

and feature item t_2 is uniformly distributed in contained documents. But for this classification, t_2 has more representation ability and higher contribution rate. In computing mutual information value, the value of t_1 may be bigger than t_2, which makes t_2 leave behind, and t_2 may be eliminated in the final threshold value.

This thesis puts forward a feature selection method based on degree of difference among classes , so as to enhance precision rate of feature selection method , while taking relative frequency factor into consideration.

3 Improved Feature Selection Method

The thesis puts forward improved feature selection method based on the two elements of degree of difference among classes and relative frequency factor .

3.1 Degree of Difference among Classes

The ideology of degree of difference among classes is feature items that have strong representation ability which should focus on one or several classifications, and the contained documents of all these classifications are uniformly distributed[10]. Degree of difference among classes method integrates between-class scatter *AC* with coupling with the classification *DC*, namely considering the frequency number of feature item in different classification and distributed difference of feature items in different documents of same classification.

(1)Introducing between-class scatter *AC* to describe distribution condition of feature items. One feature item with strong classification ability should focus on one or several classifications rather than uniformly distributed. The computational formula of between-class scatter is:

$$AC = \sqrt{\frac{1}{m-1}(\sum_{j-1}^{m}(df_j(t_i) - \overline{df(t_i)})^2)}, \quad (3)$$

in the formula: $df_j(t_i)$ represents classification c_j which contains documents number of feature items; $\overline{df(t_i)}$ means average documents number of every classification contained feature item t_i, and m is classification number. The greater the between-class scatter, the greater classification ability of feature item.

(2)Introducing classification *DC* to describe distribution conditions of classification text. A strong feature item with representation ability should be uniformly distributed in the documents rather than focusing on several documents. The computational formula of classification is:

$$DC = \sqrt{\frac{1}{n_j}\sum_{k}^{n_j}(f_{kj}(t_i) - \overline{f_j(t_j)})}, \quad (4)$$

in the formula: n_j represents the total documents number of c_j; $f_{kj}(t_i)$ represents the appearance number of the k documents of feature item t_i; $\overline{f_j(t_j)}$ represents the average number of classification c_j of feature item t_i. The bigger the coupling with the classification, the stronger ability of feature item.

At last degree of difference among classes β is introduced:

$$\beta = \frac{AC}{DC}. \quad (5)$$

After introducing degree of difference among classes into formula (1), we

acquire:

$$MI = \beta * \log \frac{p(t_i / c_j)}{p(t_i)}. \quad (6)$$

3.2 Relative Word Frequency Factor relative frequency factor

Introducing relative frequency factor is mainly to solve deficiencies that are prone to select low-frequency words of feature selection in mutual information method. Word frequency is based on the frequency number of feature item in classification. If $f_j(t_i)$ represents frequency number of feature item t_i in classification c_j, the relative word frequency degree of feature item t_i is λ and λ represents:

$$\lambda = \frac{f_j(t_i)}{\overline{f(t_i)}}, \quad (7)$$

in the formula: $\overline{f_j(t_j)}$ represents the average value of appearance frequency of feature item t_i in all classifications.

Introducing relative frequency factor α:

$$\alpha = \frac{\lambda}{\sqrt{\lambda^2}}, \quad (8)$$

from above definitions, we could know that for one certain feature item, the bigger the relative word frequency, the bigger the classification difference, and the contribution rate to text classification will be higher. Therefore, introducing relative word frequency factor and classification difference can acquire the computational formula of new feature selection method:

$$MI = \alpha * \beta * \log \frac{p(t_i / c_j)}{p(t_i)}. \quad (9)$$

4 Experimental Results and Analysis

4.1 Experimental Preparation

The experiment of the thesis selects the Chinese corpus data of Fudan University, and carries out the experiments of five classification : environment, politics, economy, military, and sports. Various types of documents are selected as shown in Table 1.

Table1. Data Sets

Classification	Training Sets	Test Sets
Environment	167	33
Politics	421	84
Economy	271	54
Military	208	41
Sports	375	75

The selected proportion of training sets and test sets is 5:1. The experiment will be carried out according to the pre-process of text data, feature selection, classifier training, data test, and results analysis. The hardware environment is CPU i5, 2.6G Hz, RAM 4 GB; applied software environment is Chinese word segmentation system of Chinese Academy of Sciences(ICTCLAS), Eclipse; the classifier selects Support Vector Machine(SVM); programming language based on JAVA. The experiment is divided into two parts, one makes a comparison of improved MI method and traditional MI method under the same dimension; another makes a comparison of

improved MI method and traditional MI method under different dimension.

4.2 Evaluation Indicator

Evaluation indicator of classification effects adopts *R*(Recall) and *P*(Precision) that are universally recognized, and the computational formula of the two indicators is :

$$R = \frac{Text\ Number\ of\ Correct\ Classification}{Total\ Number\ of\ Testing\ Docunents}, \quad (10)$$

$$P = \frac{Text\ Number\ of\ Correct\ Classification}{Text\ Number\ of\ Actual\ Classification}, \quad (11)$$

R represents the ability to measure the text number of correct arithmetic classification; *P* represents the ability to measure the text number of declining arithmetic errors. Both of them reflects text quality from different aspects, and they are indispensable. In actual conditions, what is generally adopted is the harmonic mean of the two indicators, namely, comprehensive evaluation indicator F_1, and the computational formula is:

$$F_1 = \frac{2 \times P \times R}{P + R} \quad . \quad (12)$$

4.3 Experiment Results and Analysis

(1) Experiment 1

To verify the effects of the method, the thesis makes a comparison of traditional mutual information method and improved method. The feature dimension selection is 300 with SVM classifier test, and its statistical classification effects are as Table 2.

Table2. Results from Experiment 1

Classification	Traditional MI Method		Improved MI Method	
	R/%	*P*/%	*R*/%	*P*/%
Environment	35.72	66.75	48.07	76.74
Politics	70.58	54.63	71.14	66.67
Economy	69.26	63.33	75.01	68.13
Military	65.60	70.06	80.39	83.37
Sports	64.54	66.72	87.08	66.76

Table 2 shows the comparison results of various recall ratio and precision ratio adopted the two methods. After classification, the recall ratio and precision ratio of all classifications are: environment classification increased 12.67% and 9.99%; political classification increased 0.56% and 12.04%; economic classification increased 5.75% and 4.8%; military classification increased 14.79% and 13.31%; sports classification increased 22.54% and 0.04%. The biggest rise of recall ratio is 8.04%, and the highest precision ratio is military. The average recall ratio increased 11.26%, and the average precision ratio 8.04%. The data results show that the recall ratio and precision ratio results that adopted the method have a certain degree of rise compared with traditional information methods. According to the above results, F_1 value of comprehensive appraisal indicator is as Table 3.

Table3. Result 2 from Experiment 1

Classification	F_1		Increased ratio/(%)
	Traditional MI	Improved MI	
Environment	46.53	59.79	37.11
Politics	61.59	68.83	11.76
Economy	66.16	71.40	7.92
Military	67.76	81.85	20.79
Sports	65.61	75.58	15.19

From Table 3, every classification of F_1 value adopted the method in the thesis has a certain degree of rise, with environment classification the highest and economic classification the lowest. The average F_1 increases 18.55%, which means that the method put forward in the thesis is better than the traditional mutual information methods in feature selection.

(2) Experiment 2

In order to further verify the enhanced effects of feature selection in the method, the thesis makes a comparison of the method put forward in the thesis and traditional mutual information methods in different dimensions. The experiment adopts the SVM, and the results are as the Fig.1.

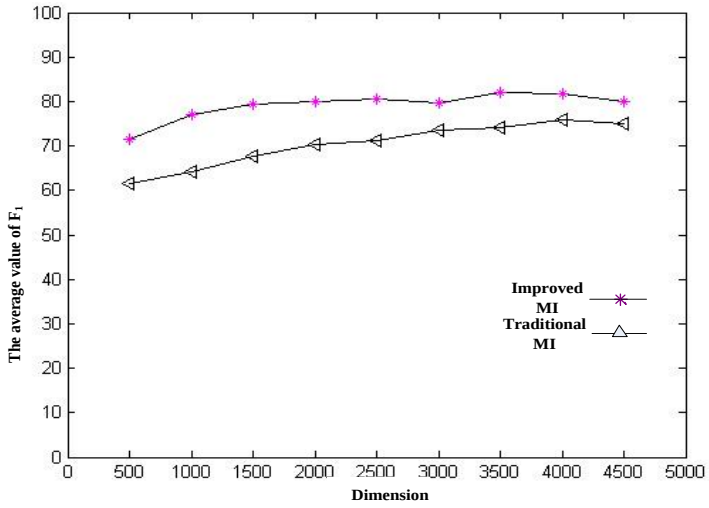

Fig.1. Effects Comparison of Text Classification of Different Dimension

From Fig.1, we could see that F_1 value of the two methods would increase with the rise of dimension number. When the dimension number reaches 4,000, the curve tends to be horizontal. The F_1 value of results acquired from traditional mutual information methods in low dimensionality is relatively low, and with the increase of dimensionality number, the growth rate may fluctuate. The thesis adopts the improved MI method. When the F_1 value reaches the number of 3,000, there is a turning point; when reaching 4,000 and 5,000, there is a decline. Generally speaking, all the F_1 values that are put forward in the thesis are higher than traditional mutual information methods in different dimension numbers.

4 Conclusion

The good or the bad of the feature selection methods has direct impact on the accuracy of the results of text classification. Targeted at the existing deficiencies of the traditional mutual information methods, the thesis puts forward an improved mutual information method based on degree of difference among classes, and introduces relative frequency factor which are prone to select high-frequency words. The experiment on text classification shows that the methods put forward in the thesis are helpful to enhance the recall rate, precision rate and F_1 value compared with the traditional mutual information methods.

Feature dimensionality reduction is the key and the difficulty of the text classification research. Feature selection and feature extraction are two methods of feature dimensionality reduction which have their own advantages and disadvantages. The stress of the next-step research should be attracting the advantages of the two methods, putting forward a comprehensive method to feature selection so as to satisfy the demand of text classification.

References

[1] Yan.J, Liu.N, Yan.S.C, et al. Trace-oriented feature analysis for large-scale text data dimension reduction. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7): 1103-1117.

[2] Zhao.Z, Wang.L, Liu.H, et al. On similarity preserving feature selection. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(3): 619-632.

[3] Harun.U . A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 2011, 24(7): 1024-1032.

[4] Amiri.F, Rezaei.M. Mutual information-based feature selection for intrusion detection systems. Journal of Network and Computer Applications, 2011, 34: 1184-1199.

[5] Forman.G. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 2003, 3(1):1289-1305.

[6] Yang.Y.H, J.O.Pedersen. A comparative study on feature selection in text categorization // Proceedings of the 14th International Conference on Machine Learning. Nashville: Morgan Kaufmann, 1997: 412-420.

[7] Yang.J.M, Wang.J, Qu.Z.Y. Feature selection method based on the relative contribution. Journal of Northeast Dianli University, 2014, 34(4): 62-68.

[8] Bakus.J, Kamel.M.S. Higher order feature selection for text classification. Knowledge and Information Systems, 2006, 9(4): 468-491.

[9] Cheng.W.Q, Tang.X. A text feature selection method using the improved mutual information and information entropy. Journal of Nanjing University of Posts and Telecommunications, 2013, 33(5): 63-68.

[10] Zhou.Q.N, Zhang.Z.H, Xu.D.C. Feature selection method for Chinese text categorisation based on class discriminating words. Computer Applications and Software, 2013, 33(7): 193-195.