An Algorithm of Feature Selection in Text Categorization Based on Gini-index

Wei-Dong Zhu[1,a], Bo Wang[2,b,*], Yong-Min Lin[3,c]

[1,2]School of Computer and Information Technology Beijing Jiaotong University, No.3 Shangyuancun, Haidian District Beijing. R. China 100044

[3]Hebei Polytechnic University Tangshan, China 063009

[a]wdzhu@bjtu.edu.cn, [b]14120474@bjtu.edu.cn, [c] linyongmin1208@126.com

*Corresponding author

Keywords: Text categorization, Feature selection, Gini-index, Feature selection function.

Abstract. TWith the rapid development of World Wide Web, text categorization has played an important role in organizing and processing large amount of text data. The first and major problem of text categorization is how to select the best subset from the original high feature space in order to reduce the high dimensionality of the original feature space and improve the classification performance. Gini-Index is the principle of multi-attribute selection very early used for attribute selection in Decision Tree, which performs near state-of-the-art level. However, relatively little work has been done on applying Gini-Index to text feature selection. We use improved Gini-index for text feature selection, constructing the measure function based on Gini-Index. We compare it to other four feature selection measures using two kinds of classifiers on two different document corpus. The result of experiments shows that its performance is comparable with other text feature selection approaches. However, it is perfect in the time complexity of algorithm.

Introduction

With the rapid development of network technology and digital libraries, online documentation is increasing quickly; automatic text categorization has become the key technology in organizing and processing large document data. The main difficulty of text categorization resides in the original features in a multi-dimensional space. Therefore, selecting the most significant characteristics from the original feature space, i.e. feature selection, will reduce the number of dimensions of the feature space, and improve the efficiency of the categorizer, boost the performance of classifier. Existing feature selection methods are based on statistical theory and machine learning methods, such as Information Gain, Expected Cross Entropy, the Weight of Evidence of Text, $\chi 2$ Statistic etc. [1, 2]. These methods have been proved good text feature selections by many researches via experiments.

Gini-Index is an impurity splitting method, which is proposed by Breiman in 1984 [3]. It has been widely used in many Decision Tree Algorithm, such as CART, SLIQ, SPRINT and Intelligent Miner, to select the splitting attribute, and achieved very good categorization accuracy. However, it is rare to use the Gini-Index in the text feature selection. Shankar has discussed [4] the application of the Gini-index principle in the text feature selection and weight adjustment issue, but the scope is limited to the centroid-based classification. However, the method described in this paper is totally different. Based on the analysis of Gini-index principle and text feature, we construct the evaluation function directly in the original feature space for feature selection, then to choose the most significant feature subset. It is not only good for centroid-based classification; it also suitable for other existing text classifier, such as kNN, SVM, LLSF, Bayes etc. According to the comparative experiments with other feature selection methods in both Chinese and English corpus, the performance is comparable to other feature selection methods. However, the time complexity of this algorithm is the most optimal.

Ext Feature Selection Method Based On Gini-Index

Gini-index principle

The specific algorithm: Suppose that S is a collection of data samples of the s, its class label attribute has m different values, which defines different classes of C_i,(i=1,…,m) . According to the class label attribute value, S can be divided into m subsets (S_i, i=1,…,m). If Si is the subset of samples belongs to class C_i, and si is the number of the samples in the subset Si, then the Gini-index of set S is

$$Gini(\boldsymbol{S}) = 1 - \sum_{i=1}^{m} P_i^2 \tag{1}$$

Where Pi is the probability of any sample of C_i, which estimated by s_i/s. When the minimum of *Gini(S)* is 0, i.e. all records belong to the same category at this collection, it indicates the maximum useful information can be obtained. When all the samples in this collection have uniform distribution for certain category, *Gini(S)* reaches maximum, it indicates the minimum useful information obtained.

If the collection of data samples *S* is divided into n subsets(S_j,j=1,…,n), based on certain attribute A. Then, the $Gini_{split}$ after splitting is

$$Gini_{split}(\boldsymbol{S}) = \sum_{j=1}^{n} \frac{S_j}{S} Gini(\boldsymbol{S}_j) \tag{2}$$

The property generating the minimum $Gini_{split}$ is chosen as splitting attribute.

The original form of the Gini-index is used to measure the "impurity" of attribute for categorization. The smaller its value, i.e. the lesser "impurity", the better attribute. On the other hand, measuring the "purity" of attribute for categorization, the bigger its value, the better "purity", the better attribute. The following formula shows this.

$$Gini(\boldsymbol{S}) = \sum_{i=1}^{m} P_i^2 \tag{3}$$

The text feature evaluation function based on Gini-index

Applying Gini-index principle with its "purity" into the text feature selection, the following formula can be built:

$$GiniIndTxt(W) = P(W)\sum_{i=1}^{m} P(C_i \mid W)^2 + P(\bar{W})\sum_{i=1}^{m} P(C_i \mid \bar{W})^2 \tag{4}$$

Based on our analysis of the pros and cons of the existing text feature selecting evaluation function, we have improved the above formula and created the following text feature selecting evaluation function:

$$GiniIndTxt(W) = \sum_{i=1}^{m} P(W \mid C_i)^{\alpha} P(C_i \mid W)^2 \tag{5}$$

To do such modification, it is based on the two reasons described below:

In the former research papers[1,2], high frequency words are emphasized when selecting text features, i.e. P(W) factor is included in the formula. Also the experiments show that certain words did not show up may contribute to determine the text category, but this contribution is often far less than the interference caused by the word does not show up, especially when the class distribution and Eigen value are highly uneven. Thus, when we create the Gini-index based text feature, the no show word condition is removed.

Since the distribution of documentation class is often uneven, it is particularly necessary to consider the

robustness of the feature evaluation function when processing unevenly distributed class. Consider the following case: $P(C_1) \neq P(C_2)$. If and only if W_1 appears in the document belong to the class C_1, and W1 appears in every document within the class C_1, i.e. $P(W_1)=P(C_1)$. If and only if W_2 appears in the document belong to the class C_2, and W_2 appears in every document within the class C_2, i.e. $P(W_2)=P(C_2)$. With the domain knowledge, we know W1 and W2 are equally important features. However, due to $P(C_1) \neq P(C_2)$, from the formula

$$GiniIndTxt(W) = P(W)\sum_{i=1}^{m} P(C_i \mid W)^2 \tag{6}$$

It can be calculated $GiniIndTxt(W_1) \neq GiniIndTxt(W_2)$, which is inconsistent with the domain knowledge. Therefore, we use the class conditional probability $P(W|C_i)$b to replace $P(W)$, and make the formula(5) to handle the class uneven distribution. Thus, when we create the Gini-index based text feature selecting evaluation function, the feature W class condition probability has been considered, using the combination of Posterior probability $P(C_i|W)$ and class condition probability $P(W|C_i)$ to evaluate the text feature, so that reduce the impact of the class uneven distribution on text feature selection.

Experimental setup

Classifier

In order to assess the effectiveness of the feature selection evaluation method, we use the two multi-valued classifiers which have better classification performance: *fk*NN text classifier and SVM classifier. There is a great difference in the statistical theory between these two classifiers. *fk*NN is a classifier with nonlinear parameters, the classification process traverses all the training data points. While SVM retain only the data points in the decision-making plane (called supporting vector), removing other data points will not affect the result of the algorithm.

The differentiating function of *fk*NN uses the FSWF rule mentioned in the reference:

$$\mu_j(\boldsymbol{x}) = \left\{ \sum_{i=1}^{k} \frac{\mu_j(\boldsymbol{X}_i) sim(\boldsymbol{X}, \boldsymbol{X}_i)}{(1-sim(\boldsymbol{X}, \boldsymbol{X}_i))^{2/(b-1)}} \right\} / \left\{ \sum_{i=1}^{k} \frac{1}{(1-sim(\boldsymbol{X}, \boldsymbol{X}_i))^{2/(b-1)}} \right\} \tag{7}$$

Where j=1, 2,…, c, $\mu_j(\boldsymbol{x}_i) sim(\boldsymbol{x}, \boldsymbol{x}_i)$ is the membership values of the j class of the known sample X. If the sample X belongs to j class, then $\mu_j(\boldsymbol{x}_i)$ is 1, otherwise, 0. Obviously, the membership value defined above is actually weighted the role of each closely classified samples according to the effect of each closely classified samples. The role of the parameter b is to determine the degree of the weight. Thus, the fuzzy k decision-making role of close neighbor is: If $\mu_j(\boldsymbol{x}) = \max_i \mu_i(\boldsymbol{x})$, then decision $\boldsymbol{x} \in \omega_j$.

SVM is proposed by V. Vapnik in 1995 [13] to solve the binary classification pattern recognition problem. We use the SVM linear model in the experiment.

In order to further investigate the effect of the algorithm, we use VC++6.0 to implement the algorithm, and partial of the source code is from the text classifier source code provided by Li Ronglu of Department of Computer and Information Technology, Fu Dan University.

Data Sets

In our experiment, two corpuses have been used. One of them is the recognized English Standard classification corpus Reuters-21578. We use the most common 10 classes, training set of 7053 documents, testing set of 2726 documents. After word root recovery and removing the un-used words, there are 23225 words. Within the experiment set, the class distribution is uneven. There are 2875 documents belong to the largest class, which takes 40.764% of the total training documents. While there are 170 documents belong to the smallest class, only 2.41% of the total training documents.

The second corpus in our experiment is Chinese Corpus from the International Database Center of

Department of Computer Information and Technology, Fu Dan University. Totally it includes 19637 documents, which divided into 20 classes. We used 9 classes from them. There are 1798 document in the training set and class distribution is not even. Among them, 619 documents are political class, which is 34.43% of the training document set A. Meanwhile the document of class energy has only 59 of them, just 3.28% of the total document set. After word root recovery and removing the un-used words, there are 78494 words.

Preprocessing

For every classifier in the preprocessing, we have done the feature selection by using information gain, expectation cross-entropy, text weight of Evidence, $\chi 2$ statistics and the revised Gini-index based text feature selecting evaluation function. The related formulas are listed below:

Information gain of the text feature W

Expectation cross-entropy of the text feature W

$$InfGainTxt(W) = P(W)\sum_{i}^{m} P(C_i \mid W)\log_2 \frac{P(C_i|W)}{P(C_i)} + P(\bar{W})\sum_{i}^{m} P(C_i \mid \bar{W})\log_2 \frac{P(C_i|\bar{W})}{P(C_i)} \tag{8}$$

$$CrossEntryTxt = P(W)\sum_{i}^{m} P(C_i \mid W)\log_2 \frac{P(C_i|W)}{P(C_i)} \tag{9}$$

$\chi 2$ statistics of the text feature W

$$\chi^2(W) = \sum_{i}^{m} P(C_i) * \frac{N(A_1A_4 - A_2A_3)^2}{(A_1 + A_3)(A_2 + A_4)(A_1 + A_2)(A_3 + A_4)} \tag{10}$$

Text weight of evidence of the text feature W

$$WeightEviTxt(W) = P(W)\sum_{i=1}^{m} P(C_i) \mid \log \frac{P(C_i \mid W)(1 - P(C_i))}{P(C_i)(1 - P(C_i \mid W))} \mid \tag{11}$$

A feature subset has been selected after applying above feature evaluation functions. Now, we can weight the selected feature via TF-IDF method as the formula below:

$$w_{ik} = \frac{tf_{ik} \times \log(N/n_i)}{\sqrt{\sum_{j=1}^{M}\left[tf_{ik} \times \log(N/n_i)\right]^2}} \tag{12}$$

Where b is equal to 2 in *fk*NN, k is determined by parameter training optimization result, when Reuters-21578, k is 40; when document set A, k is 10; when document set B, k is 35. In the formula (4), α is 2, which is also determined by parameter training optimization result.

Experiment results and analysis

To evaluate the feature selection method, we study the performance of the feature selection method from four aspects below:

Classification accuracy: Adopted performance evaluation index is, commonly used internationally, Micro-average accuracy rate [1, 5, 6] (p) and Micro F_1 of breakeven point of accuracy and recall rate.

Reduction of dimensionality: Under the premise of keeping the accuracy of the classification, it is better to have fewer numbers of selected features. During the experiment, among the feature selection methods for the training set, we recorded the number of the features that made the best classification accuracy, and

the percentage of the original total features. This helped us to evaluate the dimension reduction capability of each selection methods.

The capability to process class uneven distribution: In reality, class distribution is extremely uneven. Thus, it is necessary to compare the performance of processing unevenly distributed data set among the different feature selection methods.

Algorithm computational complexity: The logarithm calculation is used in feature selection functions such as Information Gain, Expected Cross Entropy and the Weight of Evidence of Text. However, only the multiplication is used in the Gini-index based feature selection function. Undoubtedly, from the computational complexity point of view, the latter is superior to the first three.

Table 1 and Table 2 summarized the performances of the classification of five feature selection methods, based on the most common 10 classes in the Reuters-21578 corpus, and data set A with uneven class distribution, and data set B with even class distribution. The classifier *k*NN and SVM were used.

Tab.1 performances of five feature selection methods on Reuters-21578

methods	SVM		*fk*NN	
	p(%)	Number	*p*(%)	Number (%)
GiniIndTxt	88.59	500(2.15)	86.54	1000(4.31)
InfGainTxt	88.45	500(2.15)	86.13	1000(4.31)
CroEntTxt	88.45	500(2.15)	86.21	1000(4.31)
$\chi 2$	88.23	500(2.15)	86.06	3000(12.93)
WeiEviTxt	88.48	1000(4.31)	86.28	1000(4.31)

From Table 1, we noticed, with SVM and *fk*NN classifier, Gini-index selection function not only achieved the best classification performance, but also the best on dimension reduction. Meanwhile, we also noticed these five feature selection methods all performed well. The difference of Micro-average accuracy rate is just 0.36% between the best and worst with SVM classifier. With fkNN classifier, the difference is 0.48%.

Tab.2 performances of five feature selection methods on training set A

methods	SVM		*fk*NN	
	p (%)	Number (%)	*p* (%)	Number (%)
GiniIndTxt	90.94	3000(3.822)	83.86	3000(3.82)
InfGainTxt	90.71	3000(3.822)	82.81	1000(1.27)
CroEntTxt	90.71	4000(5.096)	82.58	3000(3.82)
$\chi 2$	91.06	6000(7.644)	84.01	2000(2.55)
WeiEviTxt	90.83	4000(5.096)	85.02	3000(3.82)

From Table 2, we can tell, when processing unevenly distributed class data set, with SVM classifier, the Micro-average accuracy rate of Gini-index is only inferior to the Weight of Evidence of Text by 0.12%, but better than the other four methods. Again, it is the best on dimension reduction. With fkNN classifier, the Gini-index method is inferior to the Weight of Evidence of Text and $\chi 2$ Statistic on classification accuracy, but better than Information Gain and Expected Cross Entropy.

Via the test results from above two tables, we found that there is no big difference among the five feature selection methods on both classification accuracy and dimension reduction. They all perform quite well. However, which feature selection evaluation function is the best for specific corpus and specific classifier; it is very hard to tell. On the other hand, according to these experiments, we can declare the Gini-index based text feature selection evaluation function has equivalent good performance to the other four evaluation functions. In some cases, it offer better classification accuracy and dimension reduction.

Figure 1 and Figure 2 show the performance curves of Gini-index, information gain, expectation cross-entropy, text weight of Evidence, $\chi 2$ statistics on classifier fkNN and SVM for the most common 10 classes in the Reuters corpus and data set A in the Chinese corpus.

In Figure 1, according to the classification result of SVM and fkNN on the most common 10 classes in the Reuters corpus, besides the $\chi 2$ statistics on fkNN has slightly better accuracy between dimension 1200

and 4000, all five feature selection methods showed surprisingly similar classification performance. For SVM classifier, under the condition of not hurting the classification accuracy, about 98% or more features can be removed. However, for *fk*NN classifier, only about 96% features can be removed. Because during our experiments, we found that, when the dimension is less than 1000, all five feature selection methods have trouble to correctly classify some document. Thus, we only show the *fk*NN performance with dimension bigger than 1000. (Observation: if removing the document that cannot be classified, the average accuracy rate can be improved.)

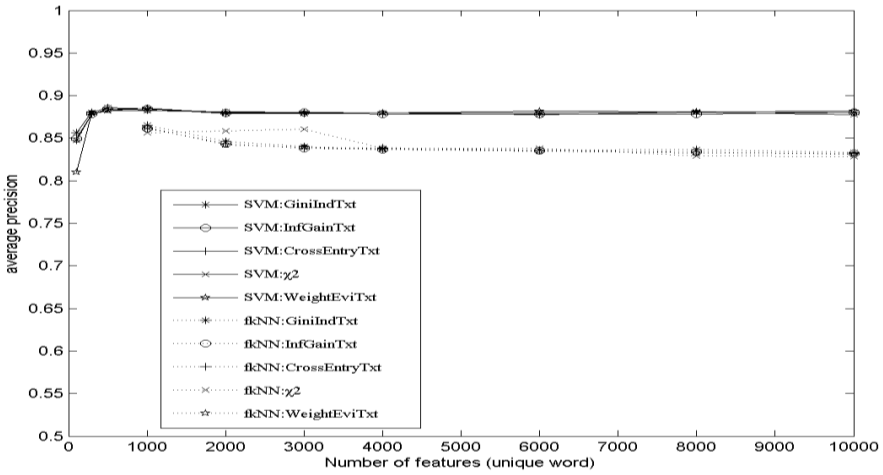

Fig.1. Curves of five feature selection methods on Reuters-21578

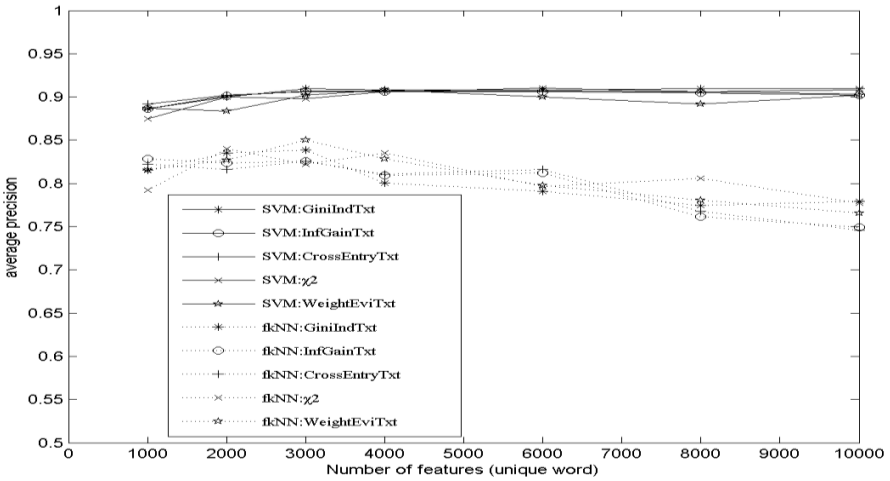

Fig.2 Curves of five feature selection methods on Training A

From the Figure 2, we observed, when processing the Chinese corpus, all five feature selection methods have a flat change with the feature dimension numbers on the SVM classifier; only the text weight of Evidence method shows little worse performance. However, with the fkNN classifier, the classification accuracy is dropping quickly with the change of feature dimensions. Each feature selection methods reach its best classification accuracy at different dimension numbers.

Sum up the information from above two diagrams, we can find out: (1) All five feature selection methods have superior classification performance on SVM than fkNN classifier; and the average accuracy changes smoothly with the feature dimension numbers on SVM. In contrast, the accuracy changes dramatically on the fkNN classifier. This could be due to the different methodology used by these two classification algorithms. (2) With two type corpus and using two different classification algorithms, we can tell these five feature selection methods have similar classification performance. Because of the close relation among the corpus, evaluation function and classifier, based on the experiment of single data set, it is impossible to conclude that certain feature selection method is better than the others. Yang has reported in his paper[1], Information Gain and $\chi 2$ Statistic have similar feature selection performance, when he compare the Information Gain, $\chi 2$ Statistic, document frequency, the words right and mutual

information.

Conclusion

In this paper, we study the text feature selection based on Gini-index. We compare and analysis the experiments in four aspects: Classification accuracy, dimension reduction, processing the class unevenly distributed document set and complexity of the calculation. Five selection methods have been used in our experiments: Gini-index based text feature selection, Information Gain, Expected Cross Entropy, the Weight of Evidence of Text and $\chi 2$ Statistic. The experiment result showed that Gini-index based text feature selection function is a very promising feature selection method. However, from the large amount of experiments, we noticed that, for certain randomly chosen training data set, five selection methods received different results. The statistical testing standard of micro-average accuracy rate hides some specific difference, which can cause less attention on some techniques deserve in deep study, even though its overall performance is not great. For specific data set, how to choose the feature selection method need more research work.

References

[1] Uguz Harun, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm", Knowledge-Based Systems, v 24, n 7, p 1024-1032, October 2011.

[2] Lei, Shang ,"A feature selection method based on information gain and genetic algorithm", 2012 International Conference on Computer Science and Electronics Engineering, ICCSEE 2012, v 2, p 355-358, 2012.

[3] Zge Uncu. "A Novel Feature Se-lection Approach: Combining Feature Wrappers and Filters", Information Sciences,Volume 177, Issue 2, 15 January 2007, Pages 449–466.

[4] Breiman L., Friedman J. H., et al, "Classification and Regression Trees. Monterey", CA: Wadsworth International Group, 1984.

[5] Yang Y., Liu X., "A re-examination of text categorization methods".,The 22nd Annual Int'l ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM PressL, 1999.

[6] Yang Y, "An evaluation of statistical approaches to text categorization", Information Retrieval, 1(1), 76~88.1999.