An Improved Feature Selection Algorithm Utilizing the within Category Variance

P.J. Zhang
School of Software Engineering,
South China University of Technology,
Guangzhou, China

S.C. Gan
School of Software Engineering
South China University of Technology
Guangzhou, China

***Abstract*-Theχ2 statistics is a commonly used and effective method of feature selection for corpus. However, it suffers several deficiencies. First, it only counts the document frequency for each feature. Secondly, this method does not distinguish among features that have different frequency distributions within a category. To overcome these shortcomings, two indexes, naming, the within category frequency and the within category variance, are introduced. Experiments are carried out to compare the traditionalχ2 statistics, some existing improvement, and the improvedχ2 statistics proposed in this paper using either naive Bayesian or SVM on the corpus collected by Fudan University and Sogou. Experimental results reveal that the improvement of this paper is effective and robust with respect to various classifiers and corpus.**

Keywords- *text classification; feature selection; χ2 statistics*

I. INTRODUCTION

Typical issues with text classification are the high dimensionality of feature space and the high sparse of the document-feature matrix. Thus, reduce the dimensionality of feature space is of extreme importance and attracts much research effort. Generally, methods of dimension reduction fall into two categories: feature selection, which choose a portion of features from the whole set by some criteria, and feature extraction, which maps the original feature space to a low dimensional space. Among feature extraction approaches, there are classical Document Frequency, Information Gain, Mutual Information,χ2statistics and ODDS Ratio, etc., [1]. Some novel methods are proposed recently, say, Orthogonal Centroid Feature Selection [2], SVM based feature selection [3] and Bayes network based feature selection [4].

Though χ2statistics is effective and has found wide applications [5], [6], it has two drawbacks. First, it only counts the document frequency for each feature, without consideration of frequency within documents, thus, it favours low-frequency features. Secondly, this method does not distinguish among features that have different frequency distributions within a category. Research work shows that features with balanced distribution within a category contribute more than features with unbalanced distribution. There are already some improvements on theχ2statistics in the literature. The first kind of improvements combine theχ2statistics with other methods, e.g., with Information Gain [7], or with Mutual Information[8]. The second kind of improvements aims to introduce some extra indexes to compensate the deficiency of the χ2statistics. In [9], frequency, degree of concentration, and degree of decentralization are introduced. Based on the indexes proposed in [9], reference [10] makes modifications so as to filter out the negative correlations between the χ2statistics and categories.

This work follows the second direction, sets forth the index of within category frequency and the index of within category variance, and defines a new χ2statistics based on them. Experimental results demonstrate that, compared to the originalχ2statistics and those modifications in [9, 10], the new χ2statistics enhances the micro F1 index.

II. IMPROVEMENT ON THE X2STATISTICS

Intuitively, given a corpus, features with high frequency are more qualified to represent categories. However, the χ2 statistics utilizes document frequency onlyand does not give enough considerations to features of high frequency. For a feature *t* and a category c_j, the within category frequency of *t* w.r.t. c_j is defined as

$$FI(t,c_j)=\frac{ctf_j(\mathrm{t})}{n_j} \qquad (1)$$

Where $ctf_j(\mathrm{t})$ is the count of t in c_j , n_j is the number of documents that c_j contains. In fact, the within category frequency is defined to be a relative measure so that it can handle unbalanced corpus.

Another observation is that, if a feature distributes evenly in some category, then, it represents this category better than features with unbalanced distributions. To characterize the balanced or unbalanced distributions, some index that measures variance of distribution is needed. Towards this, the first step is to define the distribution function. For a feature *t* and a document *d*, the distribution of *t* in *d* is defined as

$$F(t,d)=\frac{TF_d(t)}{|\mathrm{d}|+|\mathrm{V}_d|} \qquad (2)$$

Where $TF_{\mathrm{d}}(\mathrm{t})$ is the count of t in d , $|d|$ is the count of all features in d , $|V_d|$ is the count of distinct features in d .This formula has taken into account the document length, and the vocabulary issues. Now it is ready to define the within category variance as

$$V(t,c_j)=\frac{1}{n_i}\sum_{t\in c_j}(F(t,d)-F(t,c_j))^2 \qquad (3)$$

Where $F(t,c_j)=\frac{1}{n_j}\sum_{d\in c_j}F(t,d)$, n_j is the number of documents that c_j contains. Based on this two indexes, a novel χ2statistics for feature *t* against category c_j , called *NewCHI*, can be defined as

$$NewCHI(t,c_j)=\frac{\log(1+\mathrm{FI}(t,c_j))\times\chi^2(t,c_j)}{\mathrm{V}(t,c_j)+\alpha} \qquad (4)$$

Where $\chi^2(t,c_j)$ is the classicalχ2statistics. This formula incorporates the effect of both within category frequency and within category variance. α is a small positive number to force the denominator to be positive. In this paper, α is taken to be 0.01.

The last step is to define the χ2statistics of a feature for the whole corpus. This can be taken to be the sum of χ2statistics for all categories.

III. Experiments

The Fudan corpus contains 20 categories, with 19637 documents. It is a typical unbalanced corpus: some categories contain more than 1000 documents, while some others contain less than 50 documents, c.f. Table 1. On the other hand, the Sogou Compact corpus is a balanced one, each of the 9 categories contains 995 documents.

TABLEI. Document Frequency of the Fudan Corpus.

Category	Training set	Test set	Category	Training set	Test set
Economy	1369	1127	Education	58	58
Sports	1024	980	Transport	57	58
Computer	1019	591	Law	51	52
Politics	1010	989	Medical	51	52
Agriculture	847	635	Philosophy	40	33
Environment	805	371	Literature	33	32
Art	510	286	Mine	33	29
Space	506	248	Energy	30	31
History	466	468	Electronics	26	26
Military	74	75	Communication	25	22

To verify the effectiveness of the improved χ2 statistics, some groups of experiment are designed. In each of the groups, the χ2 statistics (CHI), NewCHI proposed in this work, and other improvements on CHI are compared against a specific corpus, using a specific classifier. Results are summarized below.

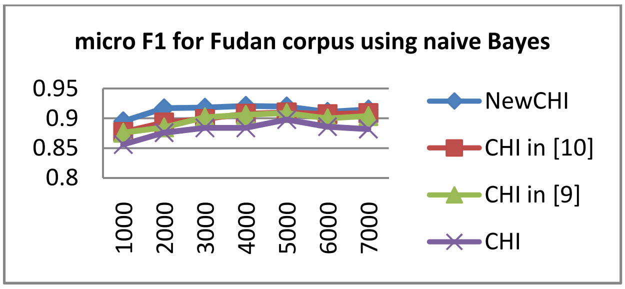

FIGURE I. MICRO F1 FOR FUDAN CORPUS USING NAIVE BAYES UNDER VARIOUS DIMENSIONS.

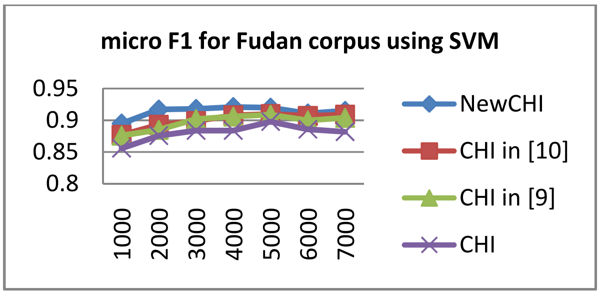

FIGURE II. MICRO F1 FOR FUDAN CORPUS USING SVM UNDER VARIOUS DIMENSIONS.

It is observed from Fig.1 and Fig.2 that, New CHI results higher F1 scores than others. Moreover, New CHI achieves maximum F1 scores by selecting fewer features. The only difference is that SVM performs much better than naive Bayes. Fig. 3 and Fig. 4 show the same trend for New CHI on the Sogou compact corpus using either naive Bayes or SVM. Since this is a balanced corpus, the F1 score increases monotonically with the dimension of feature space.

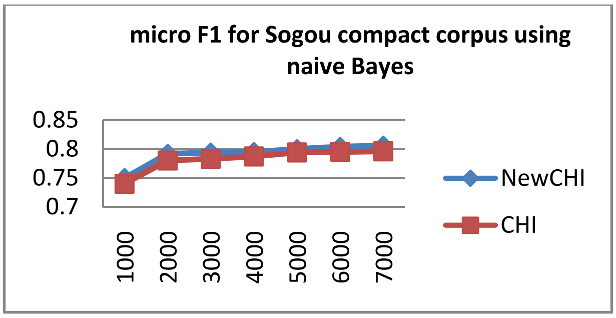

FIGURE III. MICRO F1 FOR SOGOU COMPACT CORPUS USING NAIVE BAYES UNDER VARIOUS DIMENSIONS.

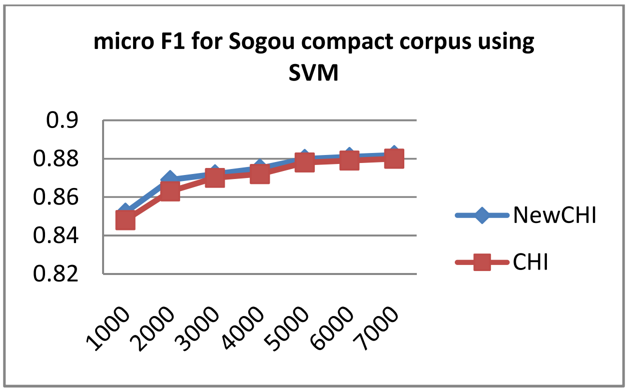

FIGURE IV. MICRO F1 FOR SOGOU COMPACT CORPUS USING SVM UNDER VARIOUS DIMENSIONS.

IV. CONCLUSIONS

This work reviews usual feature selection method in text classification, especially the $\chi2$ statistics and its improvements. Two new indexes are then introduced and used to construct the so called New CHI statistics. Experimental results show that the New CHI statistics is effective and handles unbalanced corpus well.

ACKNOWLEDGEMENTS

This work is supported by National Supercomputer Center in Guangzhou (No. 2013Y2-00036).

REFERENCES

[1] Mladenic D, GrobelnikM. Feacture selection for unbalanced class distribution and naïve Bayes, Proceedings of the 16th International Conference on Machine Learning(ICML), 1999:258-267.

[2] Yan J, Liu N, Zhang B,et al. OCFS: optional orthogonal centroid feature selection for text categorization, Proceeding of 28th ACM Internation Conference on Research and Development in Information Retrieval (SIGIR-05), Salvador, Brazil, 2005:122-129.

[3] Hardin D, Tsamardinos I, Aliferis C F. A theoretical characterization of linear SVM-based feature selection, Proceedings of the Twenty-first International Conference on Machine Learning(ICML 2004), Banff, Alberta, Canada, 2004:48.

[4] Al-Rousan N, Haeri S, Trajkovic L. Feature selection for classification of BGP anomalies using Bayesian models, Proceedings of 2012 International Conference on Machine Learning and Cybernetics.

[5] Patil P, Attar V. Intelligent detection of major network attacks using feature selection methods, Advances in Intelligent and Soft Computing, v 131 AISC, n VOL. 2, p 671-679, 2012.

[6] Feki A, Ishak A B, Feki S.Feature selection using Bayesian and multiclass Support Vector Machines approaches: Application to bank risk prediction, Expert Systems with Applications, v 39, n 3, p 3087-3099.

[7] Uysal A K, Gunal S.A novel probabilistic feature selection method for text classification, Knowledge-Based Systems, v 36, p 226-235.

[8] Adeli A, Sinaee M, Zomorodian M J, Hamzeh A.Harmony-based feature weighting to improve the nearest neighbor classification, Advances in Intelligent Systems and Computing, v 177 AISC, n VOL. 2, p 27-36, 2013.

[9] Xiong Z,Zhang P,Zhang Y.Improved approach to CHI in feature extraction, Journal of Computer Applications, Vol 28, No. 2, 513-515, 2008.

[10] Pei Y,Liu X.Study on improved CHI for feature selection in Chinese text categorization, Computer Engineering and Applications, 2011, 47(4):128-130.