Efficient Text Classification Using Term Projection

Yabin Zheng, Zhiyuan Liu, Shaohua Teng, Maosong Sun

State Key Laboratory on Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University,
Beijing 100084, China
yabin.zheng@gmail.com, liuliudong@gmail.com, tengshaohua@gmail.com
sms@mail.tsinghua.edu.cn

Abstract. In this paper, we propose an efficient text classification method using term projection. Firstly, we use a modified χ^2 statistic to project terms into predefined categories, which is more efficient compared to other clustering methods. Afterwards, we utilize the generated clusters as features to represent the documents. The classification is then performed in a rule-based manner or via SVM. Experiment results show that our modified χ^2 statistic feature selection method outperforms traditional χ^2 statistic especially at lower dimensionalities. And our method is also more efficient than Latent Semantic Analysis (LSA) on homogeneous dataset. Meanwhile, we can reduce the feature dimensionality by three orders of magnitude to save training and testing cost, and maintain comparable accuracy. Moreover, we could use a small training set to gain an approximately 4.3% improvement on heterogeneous dataset as compared to traditional method, which indicates that our method has better generalization capability.

Keywords: Text classification, χ^2 statistic, Term projection, Cluster-based classification

1 Introduction

Text classification [1, 2] is a fundamental task in text mining research area. The goal of text classification is to assign documents with pre-defined semantically meaningful labels. Traditional text classification methods always follow a supervised learning strategy: use some labeled training set and machine learning technologies, like Naïve Bayesian, KNN, and SVM [3, 4] to build a model, and then classify the documents in test set. In general, those algorithms have demonstrated reasonable performance.

Standard representation of text uses bag-of-word (BOW) model, with each term corresponds to a dimension. It is obvious that BOW representation will bring sparse and noisy problems, especially when the training set is relatively small. Moreover, this will also lead to curse of dimensionality issue, which is tough and common in text classification.

A sophisticated methodology to reduce feature dimensionality is feature selection [5], such as χ^2 statistic, mutual information and information gain. In [6], they show

that χ^2 statistic has better performance on Chinese text dataset when the dimensionality is relatively high. We did some modification on χ^2 statistic feature selection method. After feature selection, the next procedure is cluster-based text classification. The goal is to group the similar terms into clusters, in [7, 8, 9, 10], they do term clustering according to the distribution of terms with different clustering algorithm applied. Then they utilize the generated clusters as features to represent the documents for classification.

In this paper, we follow the similar procedure described above. First, we found that traditional χ^2 statistic method doesn't take term frequency into account. However, we argue that term frequency indeed shows relationship between terms and categories. The more a term emerges in a corresponding category, the stronger their relationship is. With this tiny modification, we get better performance on both English and Chinese dataset on varied dimensionality.

Second, we do term projection according to the modified χ^2 statistics, which can be considered as a rule-based clustering algorithm. The advantage of our projection method is that no additional computational cost is needed. In [7, 8, 9, 10], they have tried diverse term clustering algorithms, which bring different degrees of computational costs. Moreover, they have to determine the number of clustering result, which is difficult for different datasets.

After term projection step, we utilize the generated clusters as features to represent the documents. The benefit of using clusters as features include: (a) make most use of semantic meanings of terms. We can group similar terms into the same cluster, and condense the feature space to reduce the sparse problem to a certain extent, (b) efficient classification speed. We can reduce the feature dimensionality by three orders of magnitude, from 60,000 to 55 in Chinese dataset, while still and maintain comparable accuracy as compared to LSA. Besides, the classification speed can be greatly accelerated as a result of much smaller feature size, (c) small and better generalization classification model. With the feature dimensionality greatly reduced, we need less parameter to determine the model. Furthermore, experiment result shows that this model also has better generalization capability, (d) a complement to feature selection. Feature selection aims at removing noisy features, while term clustering is good at decreasing redundant features by putting them together. In practice, we generally do feature selection first to keep meaningful features, and then condense the feature space by clustering.

Our contributions in this paper include: (a) we modify the traditional χ^2 statistic, to the best of our knowledge, no one has take term frequency into consideration when using χ^2 statistic to do feature selection. This modification improves the performance, especially at lower dimensionalities, (b) rule-based term projection algorithm, we project the terms according to the modified χ^2 statistics, which is quite efficient and practical, (c) we reduce the feature dimensionality by three orders of magnitude, and still maintain comparable accuracy in comparison with LSA on homogeneous dataset. This indicates that our method require less computational cost to achieve the same performance, (d) we have observed some improvement both on classification accuracy and speed on heterogeneous dataset using a small training corpus.

The rest of the paper is organized as follows. In Section 2, we will review some related works that using term clustering technology for text classification as well as transfer learning that used to solve the heterogeneous dataset problem. Then we

introduce our term projection method in section 3. After that, projection based classification algorithms are discussed in section 4, using a rule-based manner or via SVM. Experiment results and discussions are shows in section 5. Section 6 concludes the whole paper and gives some future works.

2 Related Work

Pereira *et al.* [11] firstly proposed the distributional clustering scheme of English words in 1993, followed by a group of researchers in text classification area [7, 8, 9, 10], to establish a more sophisticated text representation than bag-of-words model via term clusters.

Baker and McCallum [7] apply term clustering according to the distributions of class labels associated with them. Then use these learned clusters to represent the documents in a new reduced feature space. They get only a slight decrease in accuracy; however, the cluster-based representation is significantly more efficient than BOW model.

Bekkerman *et al.* [8, 9] follow the similar idea. They introduced a new information bottleneck method to generate cluster-based representations of documents. What's more, combined with SVM, their experiment result outperforms other methods in both accuracy and efficiency. The shortcoming of all mentioned work is that they spend extra computational cost on term clustering, and some parameters like number of clusters should be determined. Unfortunately, this is always tough for different applications. In this paper, we use modified χ^2 statistic to do term projection, which can be obtained straightforwardly from the feature selection step, with no additional computational cost introduced.

On the other hand, traditional text classification strategies always make a basic assumption: the training and test set are sampling from the same distribution. However, this assumption may be violated in reality. For example, it is not reasonable to assume that web-pages on the internet are homogeneous because they change frequently. New terms emerge; old terms disappear; identical terms have different meanings. Recently, transfer learning [12, 13] is designed to solve this problem. Transfer learning is the application of skills and knowledge learned in one context being applied in another context. In this paper, we make use of cluster-based representations of documents to alleviate this heterogeneous problem. We gain improvement on both classification accuracy and speed, which give evidence of better generalization ability of our method.

3 Term Projection

In this section, we will introduce our term projection algorithm, which is significantly more efficient than other clustering algorithms. First, we present the modified χ^2 statistic formula, with term frequency taken into consideration. We also give some explanation and benefit of doing this. We also use the modified χ^2 statistic to do

feature selection in the following experiment. Second, we straightforwardly utilize modified χ^2 statistic to do term projection, which is quite efficient.

3.1 Modified χ^2 Statistic

Yang *et al.* [5] has investigated several feature selections for text classification. They found that information gain and χ^2 statistic is most effective on English text dataset among five feature selection methods. In fact, χ^2 statistic measures the lack of independence between term *t* and class label *c*. We first review the traditional χ^2 statistic formula as follows.

Using a two-way contingency table of term *t* and class label *c*, we can found four elements in the table, where *A* is the number of times that both *t* and *c* occur, *B* is the number of times that only *t* occurs, *C* is the number of times that only *c* occurs, *D* is the number of times that neither *c* nor *t* occurs. *N* is the number of documents in the training set. Statistics are performed at document level. The formula is defined to be:

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{1}$$

Ideally, *t* and *c* always occur or disappear together, which means that *t* and *c* have a strong relationship. Then once a document contains term *t*, maybe we are confident enough to classify it to class *c*. On the other hand, if *t* and *c* are completely independent, then we get a value of zero in formula (1). We computed the χ^2 statistic between a particular term *t* and all the class labels in the training set. We use the maximum value to represent the final score of term *t*, which is known as $\chi^2_{\max}$. The formula is defined as: (m is the number of classes in the training set)

$$\chi^2_{\max}(t) = \max_{1 \le i \le m} \chi^2(t, c_i) \tag{2}$$

We also record the χ^2_{total} value for future use, which is defined as:

$$\chi^2_{total}(t) = \sum_{i=1}^{m} \chi^2(t, c_i) \tag{3}$$

We argue that traditional χ^2 statistic ignores the term frequency information. For example, a document *d* which belongs to class *c* contains two terms, *t1* and *t2*. Suppose that *t1* occurs only one time in *d*, while *t2* occurs 1000 times in *d*, which is much more frequent than *t1*. It is reasonable to consider that *t2* has a stronger relationship with c than *t1* has. But this term frequency information is not revealed in formula (1). Traditional χ^2 statistic makes statistics at the document level, which assumes *t1* and *t2* have the same relation with *c*.

We perform statistic at the term level to consider the missing term frequency information discussed above. Use the same annotations before; now *A* is the total frequency of *t* occurs in class *c*, *B* is the total frequency of *t* occurs in other classes except *c*, *C* is the total frequencies of other terms occur in *c*, *D* is the total frequencies of all terms (term *t* not included) outside class *c*, *N* is the total frequency of all terms

in training set. Experiment results on both English and Chinese datasets show that our modified χ^2 statistic has better performance, especially at lower dimensionalities. Actually, we obtain 18.4% improvement on Chinese dataset at 200 dimensions.

3.2 Projection by χ^2 Statistic

As stated above, we use modified χ^2 statistic and χ^2_{max} to do feature selection. In traditional methods, they sort the terms according to their χ^2_{max} values, and choose top T candidates. However, χ^2_{max} values have natural semantic meanings. For example, in our experiment dataset, the term "导演(director)" gains χ^2_{max} in class "电影(movie)", which gives evidence that "导演" has the strongest semantic relationship with "电影" among all the classes. So, we have every reason to use χ^2_{max} information to do semantic term projection.

Furthermore, we record the χ^2_{max} values of each term and the corresponding class that it gains χ^2_{max} value. Then we make semantic matching between terms and classes, which can be considered as term projection (clustering) procedure straightforwardly after feature selection. Similar terms are projected to the same cluster.

The benefit of our proposed projection method is: (a) make most use of semantic meanings of the dataset. Dataset is usually labeled by human, which is in high quality. We use exactly the same taxonomy of training set to do projection, (b) no clustering parameters introduced. Other clustering algorithms always require extra parameters, such as clustering numbers, iteration convergence control parameters. It is always difficult to set those extra parameters for various dataset, (c) our method is significantly more efficient. Unlike other clustering algorithms, all those projections are generated straightforwardly after previous feature selection step without any extra computational cost brought about.

Meanwhile, we do some post-processing jobs to reduce noise. Certain terms may appear uniformly in classes, such as some stop words. It can be projected to every class on different datasets. Consider classifying documents into classes by individual sport (like basketball, football, volleyball). It is suitable to project term "coacher" to either class. To solve this problem, we only keep the terms whose χ^2_{max} value makes up at least λ% of their χ^2_{total} value. λ is set as 50 in this paper.

4 Cluster-based Classification

In this section, we will show how to utilize the generated clusters as features to represent the documents. We reduce the feature dimensionality by three orders of magnitude using this more sophisticated text representation. The direct benefit of this approach is that much more efficient classification speed. For practical applications, we always desire splendid processing speed as the documents on the internet accumulate exponentially. Besides, we also gain better generalization performance using this representation on heterogeneous dataset.

In subsection 4.2, we proposed two strategies to do classification task using cluster-based representation. The former performs in a rule-based manner, which

classifies the document based on the values on individual cluster features; the latter takes advantage of classification power of SVM. In our experiment, the latter method achieves better performance with a little more computational cost. While in practice, we are free to choose either method under different situation.

4.1 Cluster-based Representation

First, we will introduce how to use clusters to represent documents in detail. As discussed before, this representation is more sophisticated than traditional BOW model with semantic meanings of terms taken into consideration. Similar terms are projected to the identical cluster. We project the documents from previous feature space to the newly created feature space, in which each dimension corresponds to a cluster.

Then, we elaborate the concept of discriminability and a straight metric of this concept [14]. Discriminability measures how unbalanced is the distribution of term among the classes. A term is said to be have high discriminability if it appears significantly more frequent in one class *c* than others. Once this term appears for quite a lot of times in one document, it is reasonable to infer that this document belongs to class *c*. Forman [15] proposed a straight metric named *probability ratio* (for brief, we use PR instead hereafter) to measure the discriminability of a term, where *df* means number of documents:

$$PR(t,c) = \frac{P(t \mid c_+)}{P(t \mid c_-)} = \frac{df(t,c_+)/df(c_+)}{df(t,c_-)/df(c_-)} \tag{4}$$

Like χ^2 statistic discussed before, we use the maximum value to represent the final discriminability of term *t*, which is denoted as PR_{max}.

$$PR_{\max}(t) = \max_{1 \le i \le m} PR(t,c_i) \tag{5}$$

Suppose document *d* contains three terms, t_1, t_2 and t_3, with occurrences of n_1, n_2 and n_3, respectively. For simplicity, we only label *d* as positive or negative, which is a binary classification problem. Assume that t_1 and t_2 are projected to positive class c_1 and t_3 is projected to negative class c_2 in term projection step. Moreover, suppose t_1 and t_2 have better discriminability ($PR_{max}(t_1) > PR_{max}(t_3)$, $PR_{max}(t_2) > PR_{max}(t_3)$) as well as more occurrences in *d* than t_3 ($n_1 > n_3$, $n_2 > n_3$). Then, we represent *d* in the new 2-dimensional feature space, the first dimension corresponds to c_1 and the second dimension corresponds to c_2. Corresponding feature weighting is performed as:

$$\begin{aligned} weight_1 &= \log(n_1+1) \times \log(PR_{\max}(t_1)) + \log(n_2+1) \times \log(PR_{\max}(t_2)) \\ weight_2 &= \log(n_3+1) \times \log(PR_{\max}(t_3)) \end{aligned} \tag{6}$$

The weighting schema is similar to traditional TF*IDF, with discriminability taken into consideration. As a result, $weight_i$ indicates the possibility that *d* belongs to class c_i. Therefore, we tend to label *d* as positive according to above information.

As we can see, feature dimensionality is greatly reduced using this representation, which is only related to the number of classes in the training set. Straightforwardly,

our method has significantly more efficient training and classification speed, especially when the dataset contains hundreds of thousands of class labels.

4.2 Classification

In the last subsection, we have demonstrated have to use the clusters to represent the documents. The dimensionality of the new feature space is exactly the number of classes in training set, with each dimension corresponds to a class.

In fact, the value of $weight_i$ in the new representation of document d implies the probability that d belongs to $class_i$. The naive and intuitive idea is classifying d to the class in which d obtains maximum weight. Based on this idea, the classification can be performed in a rule-based manner, which is quite efficient. At the meantime, experiment results prove this method maintain comparable accuracy. In addition, we can exploit the classification power of SVM, which is considered as a powerful tool for machine learning task especially text classification. We apply the same cluster-based representations both to training and test sets, then use training set to build a SVM model, which is used to classify documents in the test set.

5 Experiment

5.1 Experimental Setting

We carry out experiments both on Chinese and English datasets. 20Newsgroups [16] is a widely used English document collection. We choose this collection as a secondary validation case for modified χ^2 statistic.

For Chinese document collection, we involve two datasets. One is the electronic version of Chinese Encyclopedia (**CE**). This collection contains 55 categories and 71669 single-labeled documents (9:1 split to training and test set). This collection is homogeneous. The other is Chinese Web Documents (**CWD**) collection. It has the same taxonomy as CE, including 24016 single-labeled documents. The distributions of two Chinese text collections are diverse though under the same taxonomy, which reflects the heterogeneous problem.

Libsvm [17] with linear kernel is used as our SVM classifier. Previous work [6] shows that Chinese character bigram has better performance than Chinese word unit at higher dimensionality. Besides, we don't have to consider Chinese word segmentation problem. We use bigram as our term unit. Finally, Micro-average F1-Measure is adopted as performance evaluation metric.

In addition, "traditional method" used hereafter follows a straightforward strategy: use traditional χ^2 statistic with various dimension cutoff values to do feature selection, and then use Libsvm to train a SVM model, finally, classify the documents in test set.

5.2 Modified χ^2 Statistic

In this subsection, we will first illustrate that modified χ^2 statistic that takes term frequency into account indeed improves accuracy especially at lower dimensionalities. We use SVM as classifier in this experiment, with different χ^2 statistic methods applied.

Experiment result on CE is shown in Fig.1. X-axis represents the dimension cutoff value, and Y-axis means the corresponding F1 value. It is remarkable that we gain 17% improvement at dimensionality of 100, from 35.1% to 52.1% and 18.4% improvement at 200, from 44.4% to 62.8%. Furthermore, we can promote the performance on various dimensionalities to a certain extent.

To verify the effectiveness of our method further, we also carry out the same experiment on 20NG dataset. Result is shown in Fig.2, which is similar to CE. This shows that our proposed modified χ^2 statistic approach has good generalization performance. On the other hand, we can infer that term frequency is helpful information for feature selection. In this context, we adopt the modified χ^2 statistic as our feature selection method.

In addition, we also obtain greater promotion on Chinese dataset compared to English dataset, which indicates that our method is more appropriate for Chinese text classification. Therefore, we use CE (for homogenous case) and CWD (for heterogeneous case) collections in the following experiments.

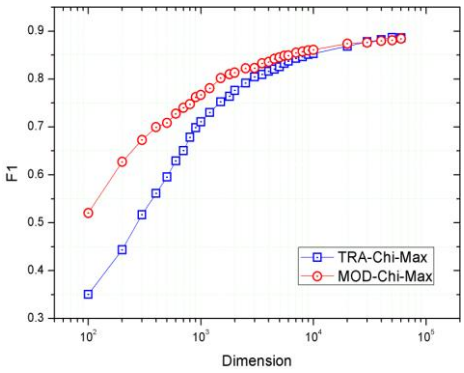

Fig. 1. Modified χ^2 statistic on CE

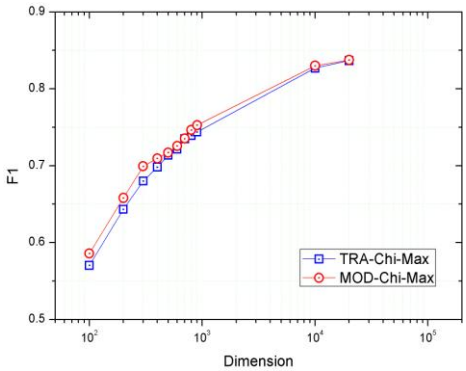

Fig. 2. Modified χ^2 statistic on 20NG

5.3 Homogeneous Dataset

We first focus on homogeneous case, which means that training and test set are sampling from the same distribution. We split CE dataset to training and test set according to the proportion of 9:1. With 64529 documents used as training set and 7140 used as test set.

Follow the instructions described in section 3 and 4. We first use the modified χ^2 statistic to do feature selection and term projection. Two parameters introduced in this step, one is the dimension cutoff value *T*, which determines the selected feature size, and the other is λ, which remove the noisy term whose χ^2_{max} value is relatively small

compared to its χ^2_{total} value. We set λ to 50 in the following experiments. In other words, we only keep the terms whose χ^2_{max} value makes up at least 50% of their χ^2_{total} value. We have done different trials with various values of *T*. In fact, we are able to utilize all the terms in the training set, as we only need to record the projection result of them. While in traditional method, each term corresponds to a dimension, it's impossible to handle the dimension at that level.

Then we use the term projection information to represent the documents in training and test set, following the procedure in subsection 4.1. Using this cluster-based representation, we can extremely reduce the feature dimensionality to the number of classes in training set. In other words, we use vectors in a 55-dimensional feature space to represent the documents in CE dataset. We can apply rule-based or SVM-based algorithms to do classification discussed in subsection 4.1. The results of both methods are shown in Fig.3 along with different values of *T*.

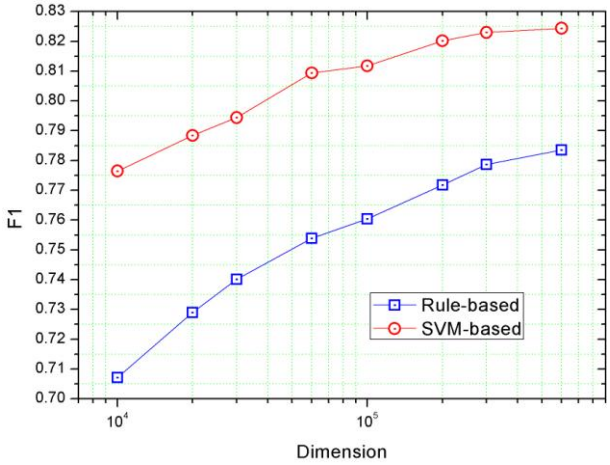

Fig. 3. Rule-based and SVM-based results with different *T* on homogeneous datasets

As illustrated in Fig.3, we get better performance when *T* increases, which shows that our noise reducing method is effective. Generally speaking, using rule-based manner, we can achieve an acceptable and comparable performance with extremely efficient classification speed. Furthermore, we can an approximately 5% improvement by SVM compared to rule-based method with a little more computational cost. We gain the best performance when *T* reaches 600,000, which indicates that we almost use all the useful terms in the training set, this is impossible for traditional method because of dimensionality curse problem.

Rule-based method gets F1-value of 78.3%, a comparable result with traditional method with dimension cutoff value of 2,000. While SVM-based method gains F1-value of 82.4%. Using the traditional method, we gains the same performance with dimension cutoff value of 4,000. In other words, we can get an acceptable and comparable result with less training efforts and greater classification efficiency. The training and test time are shown in Table 1, in a PC with Intel Core2 2.10GHz CPU and 3G memory.

As we can see clearly from the table, using SVM-based method, we can save training time by 75 percent, as well as test time by 50 percent without any lost in F1. Furthermore, rule-based method does not require model training process, and the classification step is really efficient. In fact, we spend only about 2 percent test time of traditional method, but gain a small improvement.

Table 1. Training and test time on homogeneous datasets

Algorithms	Dimensionality	Training Time	Test Time	F1
Traditional Method	4,000	430.676s	139.873s	82.4%
SVM-based	**55**	**101.95s**	**66.795s**	82.4%
Traditional Method	2,000	247.805s	91.957s	77.9%
Rule-based	**55**	**0s**	**2.238s**	**78.3%**

We also do comparisons between Latent Semantic Analysis (LSA) and our method. As we all known that LSA is time consuming and computational intractable, which is not suitable for practical application. In a PC with Intel Core2 2.10GHz CPU and 3G memory, it takes about an hour to do singular value decomposition (SVD) when the dimensionality is reduced to 200. We also perform similar experiments on dimensionalities of 55 and 100. Results are shown in Table 2. On the contrary, we obtain some improvement both on classification accuracy and speed compared with LSA. Rule-based and SVM-based methods get F1-value of 78.3% and 82.4% with dimensionality reduced to 55, which gains improvements of 2.4% and 6.5% compared to LSA with the same dimensionality.

Table 2. Comparisons with LSA

Algorithms	Dimensionality	F1
LSA	**55**	75.9%
LSA	100	78.5%
LSA	200	80.9%
Rule-based	**55**	78.3%
SVM-based	**55**	**82.4%**

5.4 Heterogeneous Dataset

To verify the generalization performance of our proposed methods, we carry out similar experiments on heterogeneous datasets. As shown before, distributions of CE and CWD datasets are distinct. CE stands for a more constant distribution, while CWD reflects the characteristic of web documents which change from time to time. We use the smaller portion of CE as our training set in the following experiment, which contains 7140 documents. CWD with 24016 documents is used as test set.

We follow the same steps in previous experiment. Results of rule-based and SVM-based methods on heterogeneous datasets are shown in Fig. 4. F1-value of both methods increases along with larger value of *T*. Overall, performance of SVM-based method exceeds rule-based method by 3%. Both methods reach peak performance when *T* is 200,000, with F1-values of 64.4% and 68.3%.

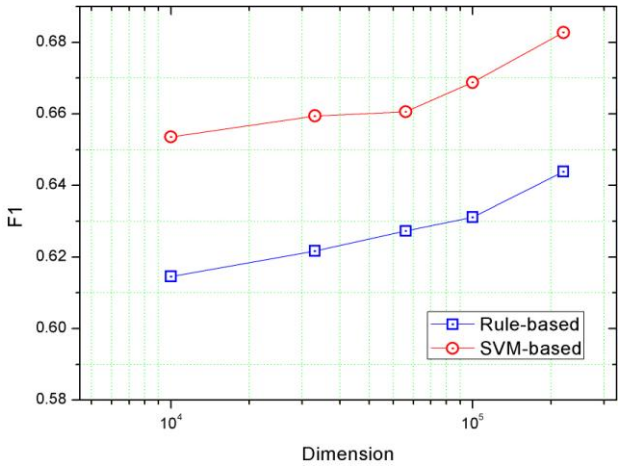

Fig. 4. Rule-based and SVM-based results with different T on heterogeneous datasets

We also compare our methods with traditional one. Traditional method gains best F1-value of 64% at dimensionality of 60,000. Both of our methods get better performance with less training and test time. SVM-based method uses about 2 percent training time and 7 percent test time of traditional method, but gains improvement of 4.3%. Besides, Rule-based method ignores training process and uses only about 1 percent test time of traditional method, while still maintains comparable performance.

Table 3. Training and test time on heterogeneous datasets

Algorithms	Dimensionality	Training Time	Test Time	F1
Traditional Method	60,000	179.51s	462.875s	64%
SVM-based	**55**	4.065s	31.678s	**68.3%**
Rule-based	**55**	**0s**	**5.938s**	64.4%

6 Conclusion

In this paper, we proposed an efficient text classification method based on term projection. First, we show that our modified χ^2 statistic promotes the performance especially at lower dimensionalities. Then, we project the terms to appropriate classes using the modified χ^2 statistic, this can make most use of semantic meanings of terms. We also use a more sophisticated cluster-based text representation to reduce the feature dimensionality by three orders of magnitude. Finally, Rule-based and SVM-based methods are adopted to do classification. Experiment results on both homogeneous and heterogeneous datasets show that our method can greatly reduce the training and test time and cost, while still maintains comparable or even better performance than traditional method. As a result, our method is practical in the large-scale text classification tasks which require efficient classification speed. Whether our method is effective on other heterogeneous datasets is left as future work.

Acknowledgement

This work is supported by the National 863 Project under Grant No. 2007AA01Z148 and the National Science Foundation of China under Grant No. 60873174.

References

1. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys. 34(1), 1-47 (2002)
2. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning, pp. 137-142 (1998)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification (2nd Edition). Wiley-Interscience, New York (2000)
4. Yang, Y.M., Liu, X., A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval, pp.42–49 (1999)
5. Yang, Y.M., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of 14th International Conference on Machine Learning, pp. 412–420 (1997)
6. Li, J.Y., Sun, M.S., Zhang, X.: A comparison and semi-quantitative analysis of words and character-bigrams as features in Chinese text categorization. In: Proceedings of COLING-ACL '06, pp. 545–552 (2006)
7. Baker, L.D., McCallum, A.K.: Distributional clustering of words for text classification. In: Proceedings of 21st ACM International Conference on Research and Development in Information Retrieval, pp. 96-103 (1998)
8. Bekkerman, R., El-Yaniv, R., Tishby, N., Winter, Y.: On feature distributional clustering for text categorization. In: Proceedings of 24th ACM International Conference on Research and Development in Information Retrieval, pp. 146-153 (2001)
9. Bekkerman, R., El-Yaniv, R., Tishby, N., Winter, Y.: Distributional word clusters vs. words for text categorization. Journal of Machine Learning Research, 3, 1183-1208 (2003)
10. Chen, W.L., Chang X.Z., Wang H.Z., Zhu J.B., Yao T.S.: Automatic word clustering for text categorization using global information. In: First Asia Information Retrieval Symposium, pp.1-6 (2004)
11. Pereira.F., Tishby,N., Lee, L.: Distributional clustering of English words. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 183-90 (1993)
12. Ling, X., Dai, W.Y., Jiang, Y., Xue, G.R., Yang Q., and Yu Y.: Can Chinese Web Pages be Classified with English Data Source?. In: Proceedings of the 17th international conference on World Wide Web (2008)
13. Dai, W.Y., Xue, G.R., Yang Q., and Yu Y.: Transferring Naive Bayes Classifiers for Text Classification. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence (2007)
14. Li, J.Y., Sun, M.S.: Scalable term selection for text categorization, In: Proceedings of EMNLP' 07, pp. 774-782 (2007)
15. Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3, 1289-1305 (2003)
16. Rennie,J.: 20Newsgroups dataset, http://people.csail.mit.edu/jrennie/20Newsgroups/
17. Chang, Chih-Chung, Lin, Chih-Jen: LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001)