A Probabilistic Approach to Feature Selection for Multi-class Text Categorization

Ke Wu[1], Bao-Liang Lu[1,⋆], Masao Uchiyama[2], and Hitoshi Isahara[2]

[1]Department of Computer Science and Engineering
Shanghai Jiao Tong University
800 Dong Chuan Rd., Shanghai 200240, China
{wuke,bllu}@sjtu.edu.cn
[2]Knowledge Creating Communication Research Center,
National Institute of Information and Communications Technology,
3-5 Hilaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan
{mutiyama,isahara}@nict.go.jp

Abstract. In this paper, we propose a probabilistic approach to feature selection for multi-class text categorization. Specifically, we regard document class and occurrence of each feature as events, calculate the probability of occurrence of each feature by the theorem on the total probability and utilize the values as a ranking criterion. Experiments on Reuters-2000 collection show that the proposed method can yield better performance than information gain and χ-square, which are two well-known feature selection methods.

1 Introduction

Text categorization is a process of assigning a text document into some predefined categories. Many information retrieval applications[1], such as filtering, routing or searching for relevant information can benefit from the text categorization research. However, a major characteristic, or difficulty of text categorization problem is the high dimensionality of the feature space. Dimension reduction techniques can be applied to handle the problem. They have attracted much recently since effective feature reduction can improve the prediction performance and learning efficiency, provide faster predictors possibly requesting less information on the original data, reduce complexity of the learned results, and save more storage space.

Dimension reduction techniques can typically be grouped into two categories, which are feature extraction(FE) and feature selection(FS). The traditional FE algorithms reduce the dimension of data by linear algebra transformations while FS algorithms reduce the dimension of data by directly selecting features from

[⋆] To whome correspondence should be addressed. This work was supported in part by the National Natural Science Foundation of China under the grants NSFC 60375022 and NSFC 60473040, and the Microsoft Laboratory for Intelligent Computing and Intelligent Systems of Shanghai Jiao Tong University.

D. Liu et al. (Eds.): ISNN 2007, Part I, LNCS 4491, pp. 1310–1317, 2007.

the original vectors. Although FE algorithms have been proved to be very effective for dimension reduction, the high dimension of data sets in text domain fails FE algorithms due to their expensive computational cost. Therefore FS algorithms are more popular in text domain.

In text categorization, FS algorithms are typically performed by assigning a score to each term and keeping some number of terms with the highest scores while discarding the rest. Numerous feature selection metrics have been proposed, e.g. information gain(IG), odds ratio, χ-square(CHI), document frequency (DF) , mutual information(MI) and SVM-based featurez selection [3], etc. These metrics have been extensively examined in binary classification and most have been extended to multi-class. However, SVM-based feature selection for multi-class text classification still has not been investigated although SVM-based feature selection yields better performance than some well-known feature selection metrics, at the same level of a feature set size. In this paper, we extend this metric to a multi-class classification case in the text domain and compared our proposed metric to two well-known feature selection measures, i.e., IG and CHI in the multi-class case.

The remainder of the paper is organized as follows. In section 2, we describe the multi-class learning algorithms and feature selection methods to be used in our experiments. In section 3, we introduce SVM-based feature selection method in the binary classification and extend it to multi-class classification. In section 4, the experimental results on Reuters-2000 collection are presented and analyzed. In Section 5, we conclude the paper.

2 Multi-class Classification and Feature Selection

2.1 Multi-class Classifiers

Naïve Bayes(NB). The multinomial model as described in [6] is used. The basic idea is based on the assumption that the probability of each word event in a document is independent of the word's context and position in the document and word event in a document conforms to a multinomial distribution. The predicted class for document d is the one that maximizes the posterior probability that $P(c|d) \propto P(c) \prod_t P(t|c)^{tf(t,d)}$, where $P(c)$ is the prior probability that a document belongs to class c, $P(t|c)$ is the probability that a word t occurs given class c and $tf(t,d)$ is the number of occurrences of word t in a document d.

k-Nearest-Neighbor(k-NN). Its basic idea is that it classifies a new sample into a predefined class based on a local vote by its k-nearest neighbors. In k-NN algorithm, classification is delayed until a new sample arrives. All samples in training set correspond to points in an n-dimensional Euclidean space and usually Euclidean distance is used to calculate the nearest neighbors of a new sample. If most of those k-nearest neighbors of a new sample are in class c, then assign the sample to c.

Rocchio [4]. It constructs a prototype vector for each category using both the centroid of positive training samples and the centroid of negative training samples. The prototype vector is calculated as follows:

$$\boldsymbol{c_j} = \alpha\frac{1}{|C_j|}\sum_{\boldsymbol{d}\in C_j}\frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} - \beta\frac{1}{|D - C_j|}\sum_{\boldsymbol{d}\in D-C_j}\frac{\boldsymbol{d}}{\|\boldsymbol{d}\|}, \tag{1}$$

where α and β are parameters that adjust the relative impact of positive and negative training samples, respectively. When classifying a new document, Rocchio classifier computes either the dot product or cosine value between the new document and the prototype vector of each class, and then assigns the new document into the class with the highest dot product or cosine value.

2.2 Feature Selection Methods

Information Gain [2]. This feature ranking criterion is based on information theory. It measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. Let $\{c_i\}_{i=1}^{m}$ denote the set of categories in the target space. The information gain of term t is defined as follows:

$$\begin{aligned} IG(t) = &-\sum_{i=1}^{m} P(c_i)\log P(c_i) \\ &+ P(t)\sum_{i=1}^{m} P(c_i|t)\log P(c_i|t) + P(\bar{t})\sum_{i=1}^{m} P(c_i|\bar{t})\log P(c_i|\bar{t}). \end{aligned} \tag{2}$$

Chi-square(CHI). CHI [2,7] measures the lack of independence between t and c_i and can be compared to the chi-square distribution with one degree of freedom to judge extremeness. It is defined as follows:

$$\chi^2(t, c_i) = \frac{N[P(t, c_i)P(\bar{t}, \bar{c_i}) - P(t, \bar{c_i})P(\bar{t}, c_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c_i})}, \tag{3}$$

where N is the total number of documents. For each category, we caculate the χ^2 statistic between each unique term in training data set and the corresponding category, and then we combine the category-specific scores of each term into two scores as follows:

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P(c_i)\chi^2(t, c_i), \tag{4}$$

$$\chi^2_{max}(t) = \max_{i=1}^{m}\{\chi^2(t, c_i)\}. \tag{5}$$

In this paper, we use $\chi^2_{avg}(t)$ for multi-class text classification.

3 A Probabilistic Feature Selection Approach

In the linear case of binary classification, the output of a trained SVM can be expressed as:

$$F(x) = sign(\boldsymbol{w}^T \cdot \boldsymbol{x} + b) = sign(\textstyle\sum_i w_i x_i + b). \tag{6}$$

From (6), we can see that a feature i with the weight w_i close to 0 has a smaller effect on the prediction than features with large absolute values of w_i. If the classifier performs well, the input feature subset with the largest weights should correspond to the most informative features [9]. As a result, $|w_i|$ is an evidence for feature ranking. This method was introduced by Brank *et al.* [3] for binary text categorization in 2002.

On the other hand, how to extend svm-based binary feature selection into multi-class feature selection remains unresolved in text categorization. In this Section, we apply a probabilistic approach to implement the extension of svm-based binary feature selection.

Intuitively, we try to extract features that all classes regard as important from candidate features and rank them higher. To the end, we apply a similar method to one-versus-all multi-class SVMs [12] to decompose a multi-class problem into a series of two-class sub-problems and combine these results in a probabilistic way.

More specifically, we first construct k two-class classifiers, one for each class, where k is the number of classes. The ith SVM is trained with all the samples from the ith class against all the samples from the rest classes and thus k decision functions are generated. Consequently, the ith decision function $F_i(x)$ is used as a binary classification sub-model criterion for discriminating the ith class from the all other classes. On the other hand, we regard document class and occurrence of each feature as events, calculate the probability of occurrence of each feature by the theorem on the total probability and utilize the values as a ranking criterion.

Assume that there are sure event E and impossible event $\varnothing$. Let E_i indicates the event that the ith class is true. According to probability theory, events E_1, E_2, ..., E_k constitute a sample space S. S corresponds to sure event E, where $E = E_1 \cup E_2 \cup \ldots \cup E_k$ and $E_i \cap E_j = \varnothing, i \neq j$. $P(E_i)$ is the prior probability that the ith class is true. Define a random event F as a event that a feature is selected as a discriminative feature. Let f_m denote the mth feature and let $P(F = f_j|E_i)$ denote the conditional probability of the event that f_j is selected as a discriminative feature given that E_i has occurred. When event E_i occurs, the ith binary classification sub-model is effective for determining the final classification result.Under the feature ranking criterion $R^{(i)}$of ith sub-model, we can derive $P(F = f_j|E_i)$ by the following equation.

$$P(F = f_j|E_i) = \frac{r_j^{(i)}}{\sum_{j=1}^{n} r_j^{(i)}}. \tag{7}$$

According to the theorem on the total probability, $P(F = f_j)$ can be derived from $P(F = f_j|E_i)$ and $P(E_i)$.

$$P(F = f_j) = \sum_{i=1}^{k} P(F = f_j|E_i)P(E_i). \tag{8}$$

The above probability can be exploited as a feature selection metric for multi-class classification.

Table 1. Accuracy rates of k-NN classifiers for various feature set sizes

Dimensionality	IG	CHI	PSVM
50	41.00±0.41	44.75±0.39	**50.30**±0.38
100	38.89±0.40	41.00±0.36	**56.63**±0.34
200	43.02±0.50	42.80±0.52	**61.92**±0.41
300	43.62±0.48	44.80±0.52	**63.44**±0.54
400	43.91±0.49	44.66±0.50	**63.33**±0.51
500	44.20±0.49	44.91±0.50	**63.25**±0.54
800	45.28±0.53	45.65±0.55	**64.62**±0.51
1000	45.96±0.47	46.44±0.54	**65.06**±0.51
2000	47.08±0.47	47.30±0.49	**64.94**±0.42
3000	47.57±0.47	47.72±0.47	**53.10**±0.46
4000	47.69±0.46	47.84±0.47	**53.31**±0.46
5000	47.89±0.43	**47.93**±0.44	47.64±0.46
8000	**48.26**±0.42	48.22±0.44	48.11±0.40
10000	48.37±0.43	**48.38**±0.42	48.20±0.40
50000	**48.63**±0.41	48.62±0.42	48.55±0.41
100000	**48.63**±0.41	**48.63**±0.41	48.58±0.42
159300	48.63±0.41	48.63±0.41	48.63±0.41

4 Experiments

In the paper, the Reuters-2000 collection[1] is used to conduct all experiments. It includes a total of 806,791 documents, with news stories covering the period from 20 Aug 1996 through 19 Aug 1997. We divided this time interval into a training period, which includes all the 504,468 documents dated 14 April 1997 or earlier, and test period, consisting of the remaining 302,323 documents. We used the same 16 categories that were selected in [3]. These statistics for the selected subset of 16 categories approximately follows the distribution for all 103 categories. The selected set of categories includes: godd, c313, gpol, ghea, c15, e121, gobit, m14, m143, e13, e21, gspo, e132, c183, e142, and c13. A document may belong to one or more categories, but we simply think that a document belongs to one category.

In our experiments, data are documents from the above 16 categories. More specifically, the training data set contains 282,010 document and the test data set consists of 175,807 documents was divided into 10 parts with the approximately equal size. Two state-of-the-art feature selection methods for text categorization, i.e. IG and CHI as our baseline, are investigated on the data set. We ignored the case of the word surface form and removed words according to a standard stop list containing 523 stop-of-words and words that occur less than 4 times from the corpus. Consequently, we used the bag-of-words model to represent documents containing a vocabulary of 159,300. Additionally, the normalized TF-IDF score was used to weight features. In addition, we respectively applied the above three

[1] http://about.reuters.com/researchandstandards/corpus/

Table 2. Accuracy rates of the NB classifiers for various feature set sizes

Dimensionality	IG	CHI	PSVM
50	**56.04**±0.56	53.80±0.48	53.58±0.42
100	62.62±0.44	**65.25**±0.43	60.71±0.37
200	69.75±0.31	69.88±0.40	**70.41**±0.20
300	71.27±0.27	71.62±0.43	**73.20**±0.15
400	72.61±0.36	72.99±0.28	**75.70**±0.17
500	73.32±0.35	73.34±0.35	**76.37**±0.24
800	74.95±0.28	74.75±0.32	**77.71**±0.19
1000	75.59±0.33	75.74±0.34	**78.08**±0.24
2000	77.23±0.28	77.00±0.27	**78.94**±0.25
3000	78.38±0.31	78.34±0.27	**79.38**±0.25
4000	78.95±0.29	78.96±0.29	**79.49**±0.28
5000	79.22±0.29	79.27±0.28	**79.32**±0.26
8000	79.59±0.25	79.75±0.25	**79.79**±0.22
10000	79.84±0.25	**80.05**±0.21	79.97±0.20
50000	81.50±0.24	**81.63**±0.24	81.26±0.19
100000	82.33±0.27	**82.37**±0.28	82.04±0.24
159300	82.29±0.22	82.29±0.22	82.29±0.22

Table 3. Accuracy rates of the Rocchio classifiers for various feature set sizes

Dimensionality	IG	CHI	PSVM
50	40.86±0.32	43.20±0.37	**52.28**±0.41
100	51.01±0.38	53.47±0.43	**56.69**±0.39
200	58.53±0.40	59.96±0.32	**62.52**±0.31
300	62.98±0.47	61.87±0.39	**63.51**±0.25
400	64.40±0.39	63.87±0.40	**66.56**±0.21
500	65.40±0.34	63.60±0.43	**67.74**±0.24
800	68.07±0.37	67.06±0.34	**69.26**±0.24
1000	69.29±0.35	69.44±0.29	**70.11**±0.28
2000	71.52±0.35	70.87±0.35	**72.17**±0.41
3000	72.79±0.40	72.40±0.36	**73.27**±0.37
4000	73.45±0.37	73.06±0.38	**73.69**±0.38
5000	73.76±0.35	73.57±0.39	**73.96**±0.36
8000	74.38±0.36	74.13±0.35	**74.60**±0.35
10000	74.64±0.35	74.35±0.39	**74.70**±0.36
50000	**75.17**±0.35	75.08±0.34	75.13±0.34
100000	**75.20**±0.34	75.19±0.35	75.19±0.34
159300	75.21±0.35	75.21±0.35	75.21±0.35

feature selection metrics to three well-known multi-class classifiers, that is , kNN, NB and Rocchio.

We use LibSVM[13] for obtaining the weight of each feature. Three typical classifiers, k-NN, NB and Rocchio, are from Rainbow toolkit and default parameters were used in the experiments. The experimental results are shown in Tables 1 through 3, where IG denotes information gain metric, CHI denotes Chi-square metric, and PSVM denotes SVM-based probabilistic criterion metric. The first column indicates the feature set size in the experiments and the last three columns indicate the correct rates and their scale in percentage.

From Tables 1 through 3, we can observe that our metric can perform better than both IG and CHI. On the one hand, for three feature selection metrics, there is a comparable performance when the size of feature set becomes large. On the other hand, the proposed metric performs better than two existing metrics when the size of feature selection remain small. In Table 1, there is amazingly better performance for the proposed metric than IG and CHI until the size of feature set reaches 5000. In Tables 2 and 3, PSVM has better performance than that of IG and CHI in almost all cases. The results in Table 1 indicate that our proposed metric can make more relevant features have high feature rank, since k-NN classifier is sensitive to irrelevant features. In addition, the results in Table 2 indicate that our metric can obtain more informative features for all classes, since NB classifier has a preference to positive features for each class. Also, our metric can suppress some features with high rank, which is obtained by a statistical bias, since tf-idf scheme in Rocchio classifier has a strong relation with data.

5 Conclusion

We have presented a novel feature selection metric for multi-class classification. The empirical results of our study indicate that the proposed method has better performance than IG and CHI. Although IG and CHI are two state-of-the-art feature selection metrics, they simplely consider the relation between the single feature and some classes, and they easily suffer from the data bias. The proposed probabilistic approach can effectively avoid the case. However, it could be affected by the noise samples on support boundary. Therefore, in future work, we will investigate effect of the noise samples on support boundary on feature selection.

Acknowledgment

The authors thank Zhi-Gang Fan for his valuable advice.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrival. Addison-Wesley (1999)
2. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. Proc. Of the 14th Int. Conf. on Machine Learning (1997) 412-420
3. Brank, J., Grobelnik, M., Milic-Frayling, N., Mladenic, D.: Feature Selection Using Support Vector Machines. In Proc. 3rd Int. Conf. on Data Mining Methods and Databases for Engineering, Finance, and Other Fields (2002)
4. Ittner, D.J., Lewis, D.D., Ahn, D.D.: Text Categorization of Low Quality Images. In Symposium on Document Analysis and Information Retrieval,Las Vegas (1995) 301-315
5. Fan, Z.G., Lu, B.L.: Fast Recognition of Multi-View Faces with Feature Selection. 10th IEEE International Conference on Computer Vision (2005)76-81
6. McCallum, A., Nigam, K.: A Comparision of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization (1998)
7. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM computing Surveys **34** (1) (2002) 1-47
8. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (2000)
9. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning **46** (2002) 389-422
10. Fan, Z.G., Wang, K.A., Lu, B.L.: Feature Selection for Fast Image Classification with Support Vector Machines. Proc. ICONIP 2004, LNCS, **3316** (2004) 711-720
11. Heisele, B., Serre, T., Prentice, S., Poggio, T.: Hierarchical Classification Andfeature Reduction for Fast Face Detection with Support Vector Machines. Pattern Recognition **36**(2003) 2007-2017
12. Rifkin, R., Klautau, A.: In Defense of One-Vs-All Classification. Journal of Machine Learning Research **5** (2004) 101-141
13. Chang, C.C., Lin, C.J.: LIBSVM: A Library for Support Vector Machines [EB/OL]. http://www.csie.ntu.edu.tw/ cjlin/libsvm (2001)