Expert Systems with Applications 36 (2009) 5432–5435

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Feature selection for text classification with Naïve Bayes

Jingnian Chen [a,b,*], Houkuan Huang [a], Shengfeng Tian [a], Youli Qu [a]

[a] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
[b] Department of Information and Computing Science, Shandong University of Finance, Jinan, Shandong, 250014, China

ARTICLE INFO

Keywords:
Text classification
Feature selection
Text preprocessing
Naïve Bayes

ABSTRACT

As an important preprocessing technology in text classification, feature selection can improve the scalability, efficiency and accuracy of a text classifier. In general, a good feature selection method should consider domain and algorithm characteristics. As the Naïve Bayesian classifier is very simple and efficient and highly sensitive to feature selection, so the research of feature selection specially for it is significant. This paper presents two feature evaluation metrics for the Naïve Bayesian classifier applied on multiclass text datasets: *Multi-class Odds Ratio* (MOR), and *Class Discriminating Measure* (CDM). Experiments of text classification with Naïve Bayesian classifiers were carried out on two multi-class texts collections. As the results indicate, CDM and MOR gain obviously better selecting effect than other feature selection approaches.

1. Introduction

Due to the proliferated availability of texts in digital form and the increasing need to access them in flexible ways, text classification becomes an elementary and crucial task. In the past several years, many methods based on machine learning and statistical theory have been applied to text classification.

Among this kinds of methods, decision trees (Lewis & Ringuette, 1994), *k*-nearest neighbors (*k*NN) (Cover & Hart, 1967; Tan, 2005; Yang, 1997; Yang & chute, 1994), neural networks (Wiener, Pedersen, & Weigend, 1995), Naïve Bayes (Lewis, 1998; McCallum & Nigam, 1998) and support vector machines (SVM) (Joachims, 1998) are all successful examples. As one of these successful methods, Naïve Bayes is popular in text classification due to its computational efficiency and relatively good predictive performance. In recent years, there are many literatures about the Naïve Bayes classifier applied in text classification (Frank & Bouchaert, 2006; Kim, Han, Rim, & Myaeng, 2006; Lewis, 1998; McCallum & Nigam, 1998; Mladenic & Grobelnik, 1999; Mladenic & Grobelnik, 2003).

For text classification a major problem is the high dimensionality of the feature space. It is very often that a text domain has several tens of thousands of features. Most of these features are not relevant and beneficial for text classification task. Even some noise features may sharply reduce the classification accuracy. Furthermore, a high number of features can slow down the classification process or even make some classifiers inapplicable. Hence feature selection is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

According to John, Kohavi, and Pfleger (1994) there are mainly two types of feature selection methods in machine learning: wrappers and filters. Wrappers use the classification accuracy of some learning algorithm as their evaluation function. Since wrappers have to train a classifier for each feature subset to be evaluated, they are usually much more time consuming especially when the number of features is high. So wrappers are generally not suitable for text classification.

As opposed to wrappers, filters perform feature selection independently of the learning algorithm that will use the selected features. In order to evaluate a feature, filters use an evaluation metric that measures the ability of the feature to differentiate each class. In general filters are much less time consuming than wrappers and have been widely used in text classification. Many feature evaluation metrics have been explored, notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index (Yang & Pedersen, 1997; Forman, 2003; Mladenic & Grobelnik, 1999; Shang, Huang, Zhu, Lin, et al., 2007) and so on.

As Mladenic pointed out (Mladenic & Grobelnik, 1999), a good feature selection metric should consider problem domain and algorithm characteristics. A given evaluation metric may get much different selecting effect for different domain or algorithms. For example, Yang and Pedersen (1997) reported that IG was one of the best metric in their experiments. However, Mladenic and Grobelnik (1999) found that IG got worst selecting effect on the domain they studied. Besides the difference of problem domain, one important reason that leads to the dis-

[*] Corresponding author. Address: School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China.
E-mail address: jnchen06@163.com (J. Chen).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.054

agreement in the performance of IG is the difference of algorithms used. In this paper, we will do some research on feature evaluation metrics specially for the Naïve Bayesian classifier applied on text data, which is very simple and efficient and highly sensitive to feature selection.

The rest of this paper is organized as follows: Section 2 gives two feature evaluation metrics developed for the Naïve Bayesian classifier. Section 3 describes Naïve Bayesian classifiers used in text classification. Section 4 presents the experimental results and analysis, and we conclude at last.

2. Feature evaluation metrics for Naïve Bayes classifiers

As mentioned above, the Naïve Bayesian classifier is very simple and efficient. But it is highly sensitive to feature selection. So the study of feature evaluation metrics for it is very necessary. Mladenic and Grobelnik (1999) presented some good results in this area. They found that on all their binary-class domains Odds Ratio was among the best performing measures for the Naïve Bayes classifier. As numerous text datasets are multi-class, it is natural to adapt Odds Ratio for multi-class problems. As we have not expected that the directly extending of it performs badly. And so we present two effective metrics for the Naïve Bayes classifier applied on multi-class datasets: *Multi-class Odds Ratio* (MOR) and *Class Discriminating Measure* (CDM).

2.1. The MOR metric

The traditional Odds Ratio for binary-class domains is defined as

$$\text{Odds Ratio}(w) = \log\frac{P(w|\text{pos})(1-P(w|\text{neg}))}{P(w|\text{neg})(1-P(w|\text{pos}))}, \tag{1}$$

where $P(w|\text{pos})$ is the probability of the occurrence of word w in the positive class, and $P(w|\text{neg})$ is the probability that word w occurs in the negative class. The Odds Ratio extended directly for multi-class domains can be in the following two forms, named as *Extended Odds Ratio* (EOR) and *Weighted Odds Ratio* (WOR)

$$\text{EOR}(w) = \sum_j \log\frac{P(w|c_j)(1-P(w|\overline{c_j}))}{P(w|\overline{c_j})(1-P(w|c_j))}, \tag{2}$$

$$\text{WOR}(w) = \sum_j P(c_j)\cdot\log\frac{P(w|c_j)(1-P(w|\overline{c_j}))}{P(w|\overline{c_j})(1-P(w|c_j))}, \tag{3}$$

where $P(c_j)$ is the probability of the jth class value, $P(w|c_j)$ is the probability that word w occurs if the class value is j, and $P(w|\overline{c_j})$ is the probability that word w occurs when the class value is not j.

EOR and WOR perform badly. Why?

We can see that both EOR and WOR only prefer positive features. In domains where positive features dominate the classification results, just like the case described by Mladenic and Grobelnik (2003), the positive-feature-preferring metrics usually perform well. But for multi-class text data, negative features can usually contribute to the classification results. And positive-feature-preferring metrics may not perform well under such situation. Hence we present the MOR metric as

$$\text{MOR}(w) = \sum_j \left|\log\frac{P(w|c_j)(1-P(w|\overline{c_j}))}{P(w|\overline{c_j})(1-P(w|c_j))}\right|. \tag{4}$$

It can be seen from (4) that MOR metric prefers not only positive features but also negative features with high value of $P(w|\overline{c_j})$.

And so it can perform well, as shown in our experiments.

It should be pointed out that Zhou proposed MC-OR metric similar to MOR (Zhou, Zhao, & Hu, 2004). It is defined as

$$\text{MC}-\text{OR}(w) = \sum_j P(c_j)\left|\log\frac{P(w|c_j)(1-P(w|\overline{c_j}))}{P(w|\overline{c_j})(1-P(w|c_j))}\right|. \tag{5}$$

The difference between MC-OR and MOR is that MC-OR weighted each term in (4) with class distribution, and so give more emphasis to features in large classes. This will worsen the classification effect for small classes that are often in the majority of a multi-class text data. And so, as experimental results indicate in Section 4, MOR performs obviously better than MC-OR.

2.2. The CDM metric

Eq. (4) can be rewritten as

$$\text{MOR}(w) = \sum_j \left|\log\frac{P(w|c_j)}{P(w|\overline{c_j})} + \log\frac{1-P(w|\overline{c_j})}{1-P(w|c_j)}\right|. \tag{6}$$

The function of term $\log\frac{1-P(w|\overline{c_j})}{1-P(w|c_j)}$ in (6) is to show the contrast between $P(w|c_j)$ and $P(w|\overline{c_j})$, which is similar to that of $\log\frac{P(w|c_j)}{P(w|\overline{c_j})}$. So we can omit $\log\frac{1-P(w|\overline{c_j})}{1-P(w|c_j)}$ in (6) and define the CDM metric as

$$\text{CDM}(w) = \sum_j \left|\log\frac{P(w|c_j)}{P(w|\overline{c_j})}\right|. \tag{7}$$

From (5) and (7) we can see that CDM metric is much simpler than MC-OR. Moreover, as the experimental results in Section 4 indicate, its selecting effect is obviously better than that of MC-OR.

3. Naïve Bayesian classifiers used on text data

In the area of text classification there are two different models of Naïve Bayes classifiers in common use: the *Multi-Variate Bernoulli Event Model* and *the Multinomial Event Model* (McCallum & Nigam, 1998). Both of these two models use the Bayes rule to classify a document. Given a document d_i, the probability of each class c_j is calculated as

$$P(c_j|d_i) = \frac{P(d_i|c_j)\cdot P(c_j)}{P(d_i)}. \tag{8}$$

As $P(d_i)$ is the same for all class, then label(d_i), the class label ofd_i, can be determined by

$$\text{label}(d_i) = \arg\text{Max}_{c_j}\{P(c_j|d_i)\} = \arg\text{Max}_{c_j}\{P(d_i|c_j)\cdot P(c_j)\}. \tag{9}$$

The calculation of probability $P(d_i|c_j)$ in (9) is different in these two models.

In the multi-variate Bernoulli event model, a vocabulary V is given. A document is represented with a vector of $|V|$ dimensions. The kth dimension of the vector corresponds to word w_k from V and is either 1 or 0, indicating whether word w_k occurs in the document. To simplify the calculation of probability $P(d_i|c_j)$ in (Eq. (9)), the Naïve Bayes assumption is made in this model: that in a document the probability of the occurrence of each word is independent of the occurrence of other words. Suppose document d_i is represented with the vector $(t_1, t_2, \ldots, t_{|V|})$, then $P(d_i|c_j)$ can be calculated under the Naïve Bayes assumption as

$$P(d_i|c_j) = \prod_{k=1}^{|V|} P(w_k|c_j)^{t_k}(1-P(w_k|c_j))^{1-t_k}. \tag{10}$$

In the multinomial event model, a document is regarded as “a bag of words”. No order of the words is considered, but the frequence of each word in the document is captured. In this model, a similar Naïve Bayes assumption is made: that the probability of the occurrence of each word in a document is independent of the word’s position and the occurrence of other words in the document. Denote the number of times word w_k occurs in document d_i as n_{ik}. Then the probability $P(d_i|c_j)$ from Eq. (9) can be computed by

$$P(d_i|c_j) = P(|d_i|)|d_i|! \prod_{k=1}^{|V|} \frac{P(w_k|c_j)^{n_{ik}}}{n_{ik}!}, \tag{11}$$

where $|d_i|$ is the number of words in document d_i.

Given a training set D, the probability $P(c_j)$ from (9) is estimated as

$$P(c_j) = \frac{1 + n_j}{l + n_{\text{all}}}, \tag{12}$$

where n_j is the number of documents in class c_j, l is the number of classes, and n_{all} is the number of all documents in the training set D. There are two ways to calculate probability $P(w_k|c_j)$ in (10) and (11). By the first means $P(w_k|c_j)$ is computed as

$$P(w_k|c_j) = \frac{1 + n_{c_jk}}{n_{\text{all}} + n_j}, \tag{13}$$

where n_j and n_{all} is the same as that in Eq. (12), and n_{c_jk} is the number of documents in class c_j that contain word w_k. In this research, we will compute $P(w_k|c_j)$ in this way.

By the second means $P(w_k|c_j)$ is estimated as

$$P(w_k|c_j) = \frac{1 + N_{c_jk}}{N_{\text{all}} + N_j}, \tag{14}$$

where N_j is the number of words in class c_j, N_{c_jk} is the number of word w_k in class c_j, and N_{all} is the number of all words in the training set D.

As McCallum and Nigam (1998) pointed out, the multinomial model performs usually better than the multi-variate Bernoulli model. So we use the former in this research.

4. Experiments

4.1. Data collections and performance setting

We ran experiments with two corpora in this study: Reuters-21578 and the “Chinese text classification corpus” given by Ronglu Li on the website: http://www.nlp.org.cn.

The top ten classes of Reuters-21578 were adopted and divided into 7053 training documents and 2726 test documents. The distribution of the class in the training set is skewed. The maximum class has 2875 documents and the minimum class has only 170 documents, occupying respectively 40.762% and 2.41% of the training set. Stop-word removal and stemming were applied.

The second data set contains 2816 documents belonging to 10 classes. All these documents are divided into 1882 training samples and 934 test samples. In this training set the class distribution is relatively uniform. The maximum class has 338 documents, occupying 17.96% of training documents. The minimum class has 134 documents, occupying 7.12% of the training set. Word segmentation and stop-word removal were applied.

To assess the effectiveness of MOR and CDM we compare them with, EOR, WOR and MC-OR, the three variations of Odds Ratio for multi-class datasets. We also compare MOR and CDM with IG, which is usually among the best performing metrics for many text datasets. EOR, WOR and MC-OR are defined in Eq. (2), (3), and (5). IG is described as follows:

$$\text{IG}(w) = \sum_j P(w, c_j) \log \frac{P(w|c_j)}{P(c_j)} + \sum_j P(\overline{w}, c_j) \log \frac{P(\overline{w}|c_j)}{P(c_j)}. \tag{15}$$

To evaluate the performance, we calculated the accuracy, micro-F1 measure, and macro-F1 measure (Forman, 2003) for each feature evaluation metric when the number of features takes each of the four values: 500, 1000, 2000 and 5000.

4.2. Experimental results and analyses

Table 1 presents the experimental result for each of the 6 metrics on the top 10 classes of Reuters-21578.

It can be seen from Table 1 that the highest accuracy 85.62% is acquired by CDM when the number of selected features is 5000. MOR follows CDM and its highest accuracy is 84.92%. MC-OR and IG also perform well but are obviously inferior to CDM and MOR. EOR and WOR perform badly and get very low accuracy.

CDM also acquires the highest Micro-F1 measure 0.8257 when the number of selected features is 5000, and the highest Macro-F1 measure 0.5601 when the number of selected features is 500. MOR gets second highest Micro-F1 measure 0.8154 and MC-OR gets second highest Macro-F1 measure 0.5262.

Table 1
The performance of 6 metrics on the top 10 classes of Reuters-21578

Metrics	Accuracy				Micro-F1				Macro-F1			
	500	1000	2000	5000	500	1000	2000	5000	500	1000	2000	5000
CDM	0.8353	0.8474	0.8532	**0.8562**	0.8121	0.8208	0.8239	**0.8257**	**0.5601**	0.5482	0.5428	0.5416
MOR	0.837	0.8431	0.8492	0.8379	0.8052	0.8107	0.8154	0.8013	0.5195	0.5201	0.5205	0.5065
EOR	0.4611	0.4896	0.5059	0.5988	0.3300	0.4000	0.4243	0.5503	0.1561	0.2192	0.2447	0.3501
WOR	0.7957	0.7306	0.6711	0.6958	0.7381	0.6570	0.5912	0.6664	0.1824	0.1629	0.1715	0.4255
MC-OR	0.8404	0.8455	0.8415	0.8147	0.8106	0.8152	0.8079	0.7722	0.5183	0.5262	0.5157	0.4768
IG	0.8272	0.8276	0.8247	0.8221	0.7879	0.7859	0.7833	0.7793	0.4918	0.4877	0.4856	0.4777

Table 2
The performance of 6 metrics on the Chinese text classification corpus

Metrics	Accuracy				Micro-F1				Macro-F1			
	500	1000	2000	5000	500	1000	2000	5000	500	1000	2000	5000
CDM	0.7077	0.6542	0.6210	0.6574	0.6929	0.6317	0.6015	0.6496	0.6698	0.6031	0.5777	0.6313
MOR	**0.7259**	0.6520	0.6253	0.6338	**0.7080**	0.6291	0.5999	0.6159	**0.6826**	0.5977	0.5682	0.5960
EOR	0.1640	0.1712	0.2360	0.3994	0.1241	0.1327	0.2059	0.3378	0.1440	0.1620	0.2552	0.4183
WOR	0.4960	0.4503	0.4652	0.6510	0.3702	0.3314	0.3780	0.6223	0.2461	0.2428	0.3078	0.5534
MC-OR	0.6991	0.6488	0.6424	0.6638	0.6767	0.6170	0.6159	0.6343	0.6385	0.5703	0.5745	0.5813
IG	0.6552	0.6574	0.6510	0.6777	0.6362	0.6409	0.6355	0.6646	0.6269	0.6226	0.6115	0.6328

As a whole, on the top 10 classes of Reuters-21578 CDM performs best, MOR follows CDM, MC-OR and IG are inferior to CDM and MOR, and EOR and WOR perform worst.

The experimental results on the second data set are described in Table 2.

From Table 2, we can see that the highest accuracy 72.59%, the highest Micro-F1 measure 0.7080 and the highest Macro-F1 measure 0.6826 are all acquired by MOR when the number of selected features is 500. CDM follows MOR and gets the second highest accuracy 70.77%, the second highest Micro-F1 measure 0.6929 and the second highest Macro-F1 measure 0.6698. MC-OR and IG also perform well but are inferior to CDM and MOR. EOR and WOR perform still badly on the second data set.

In summary, experimental results on these two data sets show that CDM and MOR are among the best performing metrics for the Naïve Bayes classifier applied on multi-class datasets.

5. Conclusion

This paper presents two feature evaluation metrics (CDM and MOR) for the Naïve Bayes classifier applied on multi-class text collections. We compared CDM and MOR with EOR, WOR and MC-OR, three variations of Odds Ratio for multi-class datasets. We also compare them with IG, which is usually among the best performing metrics for many text datasets. Experimental results on two data sets show that CDM and MOR are among the best performing metrics for the Naïve Bayes classifier applied on multi-class text datasets. Moreover, the computation of CDM metric is simpler than other feature evaluation metrics.

For future work, we will experiment on more multi-class datasets. Furthermore, we will explore feature selection metrics for high skewed datasets and text collections containing unlabeled documents.

Acknowledgement

This research is supported by National Natural Science Foundation of China under Grant Nos. 60503017 and 60673089.

References

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transaction on Information Theory IT, 13*(1), 21–27.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289–1305.

Frank, E., & Bouchaert, R. R. (2006). Naive Bayes for text classification with unbalanced classes. In *Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases* (pp. 503–510). Berlin: Springer.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European conference on machine learning* (pp. 137–142). New York: Springer.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. In *Proceedings of the 11th International Conference on machine learning* (pp. 121–129). San Francisco: Morgan Kaufmann.

Kim, S., Han, K., Rim, H., & Myaeng, S. (2006). Some effective techniques for Naive Bayes text classification. *IEEE Transactions on Knowledge and Data Engineering, 18*(11), 1457–1466.

Lewis, D. D. (1998). Naive Bayes at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European conference on machine learning* (pp. 4–15). New York: Springer.

Lewis, D.D., & Ringuette, M. (1994). Comparison of two learning algorithms for text categorization. In *Proceedings of the third annual symposium on document analysis and information retrieval* (pp. 81-93). Las Vegas, NV.

McCallum, A. & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*.

Mladenic, D., & Grobelnik, M., 1999. Feature selection for unbalanced class distribution and Naive Bayes. In *Proceedings of 16th international conference on machine learning* (pp. 258–267). San Francisco.

Mladenic, D., & Grobelnik, M. (2003). Feature selection on hierarchy of web documents. *Decision Support Systems, 35*(1), 45–87.

Shang, W., Huang, H., Zhu, H., Lin, Y., et al. (2007). A novel feature selection algorithm for text categorization. *Expert System with Applications, 33*(1), 1–5.

Tan, S. (2005). Neighbor-weighted K-nearest neighbor for unbalanced text corpus. *Expert System with Applications, 28*(4), 667–671.

Wiener, E. D., Pedersen, J. O., & Weigend, A. S. (1995). A neural network approach to topic spotting. In *Proceedings of SDAIR-95, 4th annual symposium on document analysis and information retrieval* (pp. 317–332).

Yang, Y. (1997). An evaluation of statistical approaches to text categorization. *Information Retrieval, 1*(1), 76–88.

Yang, Y., & Chute, C. G. (1994). An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information System, 12*(3), 252–277.

Yang, Y., & Pedersen, J.O. (1997). A Comaprative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th international conference on machine learning* (pp. 412-420). Nashville, USA.

Zhou, Q., Zhao, M., & Hu, M. (2004). Study on feature selection in Chinese text categorization. *Journal of Chinese Information Processing, 18*(3), 17–23.