Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 2160 – 2164

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineeringand Information Science

An Improvement to Naive Bayes for Text Classification

Wei Zhang[a], Feng Gao[a,]a*

[a]*MOE KLINNS Lab, Xi'an Jiaotong University, Xi'an, Shaanxi Province, 710049 ,China*

Abstract

Naïve Bayes classifiers which are widely used for text classification in machine learning are based on the conditional probability of features belonging to a class, which the features are selected by feature selection methods. In this paper, an auxiliary feature method is proposed. It determines features by an existing feature selection method, and selects an auxiliary feature which can reclassify the text space aimed at the chosen features. Then the corresponding conditional probability is adjusted in order to improve classification accuracy. Illustrative examples show that the proposed meth-od indeed improves the performance of naïve Bayes classifier.

Selection and/or peer-review under responsibility of [CEIS 2011]

Keywords：Text classification; Feature selection; Machine learning; Naïve Bayes.

1. Introduction

Naive Bayes is based on Bayes' theorem and an attribute independence assumption [1],[2]. Its competitive performance in classification is surprising, because the conditional independence assumption on which it is based, is rarely true in real world applications. Naïve Bayes have been studied extensively by some researchers in text classification task [3],[4].

The existing literatures about text classification with naïve Bayes have focused on three aspects. One is to construct and improve naïve Bayes model[5],[6],[7], another is to discuss the "naïve hypothesis" enable or not, and analyze the effect on classify performance, then present the corresponding improvement[8]. The other is to improve feature selection because naïve Bayes is highly sensitive to feature selection

* Corresponding author. Tel.: 86-15339027665.
E-mail address: wzhang@sei.xjtu.edu.cn.

doi:10.1016/j.proeng.2011.08.404

[9],[10],[11],[12]. In theses literatures, the naïve Bayes classifiers are based on the conditional probability of the features belonging to one class after features selection using the existing methods. An auxiliary feature method is proposed in this paper. It determines features by an existing feature selection method, and selects an auxiliary feature which can reclassify the class space aimed at the chosen features. Then the corresponding conditional probability is adjusted in order to improve classification accuracy. The experiments with data sets obtained from CCERT shows that the auxiliary feature method improves the classification accuracy of naïve Bayes because of reclassification of the class space. In this paper, we propose the auxiliary feature method. The feature with auxiliary feature was found and the probability of the feature with auxiliary feature was adjusted after feature selection.

This paper is structured as follows. Section 2 introduces Naïve Bayes model. In Section 3 we present our approach for enhancing naive Bayes by using auxiliary feature adjust probability. Section 4 contains experimental results demonstrating that the predictive accuracy of naive Bayes can be improved by auxiliary feature method. Section 5 discusses related work and future work.

2. Naïve Bayes Model in Text Classification

Denote a vector of variables $D = \langle d_i \rangle$, $i = 1,2,...,n$, represent document, where d_i is corresponding to a letter, a word, or other attributes about some text in reality, and a set of $C = \{c_1, c_2, ..., c_k\}$ is predefined classes. Text classification is to assign a class label c_j, $j = 1,2,...,k$ from C to a document.

Bayes classifier is a hybrid parameter probability model in essence:

$$P(c_j \mid D) = \frac{P(c_j)P(D \mid c_j)}{P(D)} \tag{1}$$

Where $P(c_j)$ is prior information of the appearing probability of class c_j, $P(D)$ is the information from observations, which is the knowledge from the text itself to be classified, and $P(D \mid c_j)$ is the distribution probability of document D in classes space. Bayes classifier is to integrate these information and compute separately the posteriori of document D falling into each class c_j, and assign the document to the class with the highest probability, that is

$$c^*(D) = \arg\max_j P(c_j \mid D) \tag{2}$$

Assume the components d_i of D are independent with each other since conditional probability $P(D \mid c_j)$ cannot be computed directly in practice. Thus

$$P(D \mid c_j) = \prod_i P(d_i \mid c_j) \tag{3}$$

The model with the above assumption is called native Bayes model, and equation (1) becomes

$$P(c_j \mid D) = \frac{P(c_j)\prod_i P(d_i \mid c_j)}{P(D)} \tag{4}$$

Because the sample information $P(D)$ is identical to each class c_j, $j = 1,2,...,k$, equation (2) becomes

$$c^*(D) = \arg\max_j P(c_j)\prod_i P(d_i \mid c_j) \tag{5}$$

This paper will deal with two kinds of classification problems for simplicity, and denote class set $C=\{+,-\}$. We choose multi-variable Bernoulli model as our Bayes model[13], that is, if d_i corresponding to feature presence in the document, assign its value 1, otherwise, assign the value 0.

3. Auxiliary Feature Method

There are lots of feature selection methods in text classification, such as DF (Document Frequency)、IG (Information Gain)、MI (Mutual Information)、OR(Odds Ratio)、χ^2 (CHI Squared Statistic)、ECE (Expected Cross Entropy)、WET (the Weight of Evidence for Text)[14],[15],[16],[17], and plentiful works have been researched on these methods. This paper won't evaluate the pros and cons of them here, and we just adjust the probability of feature with the auxiliary property to improve the performance of original naïve Bayes classifier, after feature selection using existing algorithm.

Denote the feature corresponding to the component d_i as w_i, and denote $p(w_i|+)$ and $p(w_i|-)$ as p_{i+} and p_{i-} separately for simplicity. For every independent feature, if $p_{i+}>p_{i-}$, Bayes classifier will assign the document to class +. The idea here is to find a auxiliary feature for feature w_i， which satisfies $p_{i-}^{'}-p_{i+}^{'}>p_{i+}-p_{i-}$， and $p_{i+}^{'}$ and $p_{i-}^{'}$ is the probability of document belonging to class + and - respectively when feature w_i and auxiliary feature $w_i^{'}$ presence simultaneously in the document. The geometric illustration for the method is as follows:

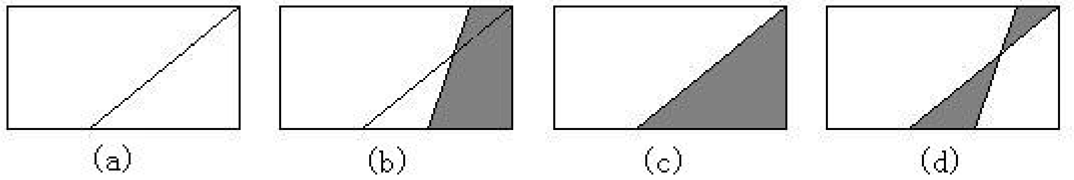

Fig.1. The geometric illustration of auxiliary feature method. (a) The text space division by the feature w_i. (b) The text space division by the feature w_i and the auxiliary feature $w_i^{'}$. (c) The misclassification induced by the feature w_i. (d) The misclassification induced by the feature w_i and the auxiliary feature $w_i^{'}$.

In figure 1, (a) denotes all the document including the features w_i and is a subspace of the whole text space, and the diagonal separate different classes; The shadow part of (b) denotes the text subspace simultaneously including auxiliary feature $w_i^{'}$; The shadow part of (c) denotes the classification fault of original naïve Bayes method. The shadow part of (d) denotes classification fault of auxiliary feature method. The constraint condition $p_{i-}^{'}-p_{i+}^{'}>p_{i+}-p_{i-}$ when finding auxiliary feature is to guarantee the top part of the shadow is smaller than the blank portion below in (d).

Algorithm of auxiliary feature:

Step1：make feature selection to determine with text vector of the dimension n.

Step2：find the set Θ composed of all document in which d_i presence.

Step3：find the auxiliary feature for d_i in Θ.

Step4：repeat step2 and step3 for $i=1,2,...,n$.

To classify a new document with a feature vector $d_1, d_2, ..., d_n$, hypothesis the auxiliary feature of the feature corresponding to the component d_i exist. While the auxiliary feature and original feature all presence in the document make $p_{i+} = p_{i+}^{'}, p_{i-} = p_{i-}^{'}$, otherwise same with naïve Bayes. Note that not all of the features have corresponding auxiliary one. If one feature exist more than a auxiliary feature, we choice auxiliary feature which is subject to $\max((p_{i-}^{'} - p_{i+}^{'}) - (p_{i+} - p_{i-}))$.

4. Experiments

The data sets we used here is two mail sets with 45396 junk mails and 18314 normal mails from CCERT[18]. We choose 30000 junk mails and 10000 normal mails as the data sets for our experiment.

There are three cases here: (a) choose 1000 features to represent the document, and it turns out that 96 features have auxiliary feature; (b) choose 1500 features to represent the document, and it turns out that 152 features have auxiliary feature; (c) choose 2000 features to represent the document, and it turns out that 217 features have auxiliary feature.

We conduct 5 times 10-fold cross validation using naïve Bayes method and the proposed method respectively for the three cases. The average results are as follows.

Table1.The classification precise of auxiliary feature method vs. naive Bayes.

	Naïve Bayes	Auxiliary Feature Method
1000 features	0.856	0.871
1500 features	0.859	0.879
2000 features	0.861	0.885

5. Conclusion

After feature selection in text classification, naive Bayes classifier partition the text subspace composed of all document in which d_i present based on each d_i. Because naïve Bayes classifier assigns the document to the class with the highest probability, naïve Bayes classifier is the optimal in probability sense. The auxiliary feature method proposed here partition the text subspace again, so it outperforms the traditional way, and illustrative examples show that the proposed method indeed improves the performance of naive Bayes classifier.

Since the auxiliary feature method need choose features twice, how to give the auxiliary directly is meaningful and can reduce the computation complexity. Meanwhile, the relationship between existing choosing features methods and our method is promising. In addition，due to the sparsity problem in text classification, whether to take the feature total of document into account when adjusting the probability is worth to work other than substitution in this paper.

Acknowledgements

The research is supported in part by the National Natural Science Foundation (60633020, 60802056, 60921003,60905018), National Science Fund for Distinguished Young Scholars (60825202), Key

Projects in the National Science &Technology Pillar Program (2011BAK08B02), 863 High Tech Development Plan (2007AA01Z480, 2008AA01Z415).

References

[1] Duda,R.O.and Hart,P.E.:Pattern Classification and Scene Analysis.New York:John Wiley 1973.

[2] Lewis, D.D. Naive (Bayes) at forty: The independence assumption in information retrieval. MachineLearning: ECML-98, Tenth European Conference on Machine Learning 1998. 4-15.

[3] D.D. Lewis, Representation and Learning in Information Retrieval, PhD dissertation, Dept. of Computer Science, Univ.of Massachusetts, Amherst, 1992.

[4] A.K. McCallum and K. Nigam, Employing EM in Pool-Based Active Learning for Text Classification, Proc. ICML-98, 15th Int'l Conf. Machine Learning, J.W. Shavlik, ed., pp. 350-358, 1998.

[5] McCallum, A., Nigam, K.: A comparison of event models for Naive Bayes text classification. In: Learning for Text Categorization: Papers from the AAAI Workshop, AAAI Press 1998. 41–48 Technical Report WS-98-05.

[6] Eyheramendy, S., Lewis, D.D., Madigan, D.: On the Naive Bayes model for text categorization. In Bishop, C.M., Frey, B.J., eds.: AI & Statistics 2003: Proceedings of the Ninth InternationalWorkshop on Artificial Intelligence and Statistics.2003 332–339

[7] D. Pavlov, R. Balasubramanyan, B. Dom, S. Kapur, and J. Parikh. Document preprocessing for naïve bayes classification and clustering with mixture of multinomials. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), 2004.

[8] Eyheramendy, S., Lewis, D.D., Madigan, D.: On the Naive Bayes model for text categorization. In Bishop, C.M., Frey, B.J., eds.: AI & Statistics 2003: Proceedings of the Ninth InternationalWorkshop on Artificial Intelligence and Statistics. 2003, 332–339

[9] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In International Conference on Machine Learning, pages 412–420, 1997.

[10] Rogati, M.; Yang, Y. High-performing feature selection for text classification. CIKM'02, 2002, pp 659-661

[11] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," Expert Systems with Applications, vol. 36, no. 3, pp. 5432-5435, April 2009.

[12] M. Srinivas, K. P. Supreethi, E. V. Prasad, and S. A. Kumari, "Efficient Text Classification Using Best Feature Selection and Combination of Methods," in Proceedings of the Symposium on Human Interface 2009 on ConferenceUniversal Access in Human-Computer Interaction. Part I: Held as Part of HCI International 2009. Springer-Verlag, 2009, pp. 437-446.

[13] McCallum, A., Nigam, K.: A comparison of event models for Naive Bayes text classification. In: Learning for Text Categorization: Papers from the AAAI Workshop,AAAI Press (1998) 41–48 Technical Report WS-98-05.

[14] D. Koller and M. Sahami. Toward optimal feature selection. In International Conference on Machine Learning, pages 284–292, 1996.

[15] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In International Conference on Machine Learning, pages 412–420, 1997.

[16] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In International Conference on Machine Learning, 2001.

[17] E. P. Xing, M. I. Jordan, and R. M. Karp. Feature selection for high-dimensional genomic microarray data. In Proc. 18th International Conf. on Machine Learning, pages 601–608. Morgan Kaufmann, San Francisco, CA, 2001.

[18] http://www.ccert.edu.cn/spam/sa/datasets.htm#2.