A New Probabilistic Model for Bayes Document Classification

Ya-Shu Liu
Department of Computer Science
Beijing University of Civil Engineering and Architecture
Beijing, P.R.China
ly_s8020@163.com

Han-Bing Yan*
National Institute of Network and Information Security
National Computer network Emergency Response technical Team/Coordination Center of China
Beijing, P.R.China
yhb@isc.org.cn
*: Corresponding Author

***Abstract*—In this paper, we propose a new probabilistic model of naïve Bayes method which can be used in text classification. This method not only takes account of the frequency of feature words, but also considers those important words do not appear in the test document, which overcomes the shortcoming of the Multi-variate Bernoulli event Model(MBM) and Multinomial event Model(MM). Experiments show that the method proposed in this paper has better classification result than those traditional methods.**

Keywords- naïve Bayes; text classification; Multi-variate Bernoulli event Model; Multinomial event Model

I. INTRODUCTION

With the tremendous growth in the volume of digital text documents, the problem of text classification has gained great importance in the area of web search, spam filtering，content based recommendation and many other fields.

The task of automatic text classification is assigning text documents to pre-specified classes(topics or themes) of documents. There have been a lot of text classification methods that have been carefully studied, such as: decision trees, rule sets, probabilistic classifiers, support vector machines, relevance feedback, linear classifiers, etc. Among these methods, the probabilistic approach of Bayes method is important and popular.

There are two kinds of typical Bayes models used in text classification: Multi-variate Bernoulli event Model(MBM) and Multinomial event Model(MM). MBM model define the feature word occurrence in the document as the document attribute, it does not care the frequency of the word appears. MM model considers the feature word frequency, but the classification result will not be influenced by those words that do not appear in the document. Obviously, MBM model and MM model all have limitations.

We propose a new probabilistic model in this paper, which combines both the merits of MBM model and MM model.

II. RELATED WORK

Using machine learning methods to solve text classification problems have been studied since 1980s. For example, Lewis tried to automatically catalog news articles[1] and electronic mail[2], Craven studied web pages category problem[3] and Pazzani learn the reading interests of users[4].

There have been a lot of methods for text classification[5][6][7][8][9][10][11][12], which can be divided into two general categories. The first category is machine learning algorithms, such as decision trees, rule sets, instance-based classifiers, probabilistic classifiers, support vector machines, etc. The other category contains specialized categorization algorithms developed in the Information Retrieval community, including relevance feedback, linear classifiers, generalized instance set classifiers, etc.

Among these methods, many research results show that the probabilistic approach is one of the most excellent algorithm in text classification with accuracy and efficiency both take into account [13][14][15][16] .

Bayes method is a kind of probabilistic approaches that have strong mathematics backgrounds. It uses a collection of labeled training examples to estimate the generative model. Then the generative model is applied to judge the new examples. The Bayes method is widely used in many fields, since it is fast to train, fast to use and easy to be implemented.

Bayes models have two typical kinds[17]: Multi-variate Bernoulli event Model(MBM) and Multinomial event Model(MM). The multi-variate bernoulli event model can be looked as an object-attribute model, which species that a document is represented by a vector of binary attributes indicating which words occur and do not occur in the document. The Multinomial event Model specifies that a document is represented by the set of word occurrences from the document, which is focused on the word frequency of a document.

Many people have studied the Cones and Proes of these two Bayes models[18]. The experiments of [18] show that the MM model always has a average of 27% reduction in error than the MBM model in long document classification, while the MBM model has a better result in short document classification.

In this paper, we propose a new bayes method combined the merits of both MBM model and MM model, which is more flexible than the original methods and suited to be used in many environments.

In the following of this paper, the bayes framework used in text classification will be first introduced, and the Proes and Cones of MBM model and MM model is discussed.

Published by Atlantis Press, Paris, France.

Then we propose the combined model. Finally, the test corpus and experiments results are given.

III. TEXT CLASSIFICATION USING BAYES THEORY

A. Bayes Framework

Traditionally, Bayes method is a probabilistic approach that minimizes the posterior expected value of a loss function. For two events C and D, Bayes method gives a way of calculating P(C|D) from a knowledge of $P(C), P(D)$ and $P(D \mid C)$. Equivalently, it maximizes the posterior expectation of a utility function.

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)} \quad (1)$$

In real application, D always is not a single event, but a combination of many events or attribute. For example, to judge whether a patient is diabetes, many factors should be taken into account, including age, weight, appetite, goiter degree, blood test results and so on. If D is a complex event, it can be viewed as a vector composed by different factors: $D = (d_1 \quad d_2 \quad ..... \quad d_n)$, where n is the vector dimension。If the dimension of event D is very high, it is almost impossible to calculate $P(D \mid C)$ directly, while $P(d_i \mid C)$ is easy to calculate. To calculate the $P(D \mid C)$, we always assume that given a class, each attribute d_i is conditional independent of other attributes. Then the following can be got,

$$P(D \mid C) = P(d_1 \quad d_2 \quad ... \quad d_n \mid C) = \prod_{i=1}^{n} P(d_i \mid C) \quad (2)$$

Substitute eqation (2) into (1) and we get equation (3),

$$P(C \mid D) = \frac{\prod_{i=1}^{n} P(d_i \mid C)P(C)}{P(D)} \quad (3)$$

By equation 3, the class has the largest value should be the class which event D should belonged to, which can be written by:

$$\arg\max_C P(C \mid D) \quad (4)$$

Because $P(D)$ is not change with different C, it need not compute in application. Equation (4) can be written by:

$$\arg\max_C \prod_{i=1}^{n} P(d_i \mid C)P(C) \quad (5)$$

If we want to calculate $P(D)$, which can be computed by:

$$P(D) = \prod_{i=1}^{n} P(d_i \mid C) + \prod_{i=1}^{n} P(d_i \mid \overline{C})$$

where $\overline{C}$ is the complement collection of class C。

B. MBM Model and MM Model used in Text Classification

The goal of text classification is to determine whether a given document belongs to a predefined category. In supervised text classification, a set of predefined categories are predefined and two sets of documents for training and testing are given. We use training documents to determine the text classification model, and use testing documents to verify the effectiveness of the trained classification model.

Documents are composed of different words. In text classification, documents are always tackled as a space of words and selected words are used as document features. According to the different statistic model, the naïve Bayes method used in text classification can be divided to two different kinds: Multi-variate Bernoulli Model and Multinomial Model.

Given a document D and a set of predefined categories $(C^1 \quad C^2 \quad ..... \quad C^q)$ (q is the categories number), the two models are discussed as following.

1) Multi-variate Bernoulli Model

In Multi-variate Bernoulli Model(MBM), a document is expressed by a binary vector over a space of feature words. Given a set of feature words $(f_1 \quad f_2 \quad ..... \quad f_V)$ with dimention V, a document D can be expressed by a vector $(t_1 \quad t_2 \quad ..... \quad t_V)$, where t_k means whether the word f_k appears in document D,the value of t_k is 0 or 1. As we all know, this is a binary Bernoulli Model in statistic theory. Under this model, the equation 5 can be expressed by:

$$\arg\max_{C^j} P(C^j)\prod_{k=1}^{V} P(f_k \mid C^j)^{t_k}\left(1 - P(f_k \mid C^j)\right)^{1-t_k} \quad (6)$$

$P(f_k \mid C^j)$ is the probability of feature f_k appears in training documents, which can be calculated by *m-estimate* method: $P(f_k \mid C^j) = (1+n_k) \Big/ \left(q + \sum_{i=1}^{V} n_i\right)$, n_i is the frequency of feature f_i appears in training documents.

In MBM Model, each feature word contributes to the classification result , no matter if it appears in the test document.

2) Multinomial Model

In Multinomial Model(MM Model), a document is expressed by a sequence of word events, which is the “bag of words” representation for documents. Based on naïve assumption, the word events satisfies multinomial distribution. A document D contains a set of feature words $(w_1 \quad w_2 \quad ..... \quad w_U)$ with the dimension U, a document D can be expressed by a vector $(n_1 \quad n_2 \quad ..... \quad n_U)$, where n_m is the frequency of word w_m appears in document D. Using the multinomial distribution theory, the equation 5 can be expressed by,

$$\arg\max_{C^j} P(C^j) n! \prod_{m=1}^{U} \frac{P(w_m \mid C^j)^{n_m}}{n_m!} \quad (7)$$

Where $n = \sum_{m=1}^{U} n_m$. Using m-estimate method, $P(w_m \mid C^j)$ can be expressed by $P(w_m \mid C^j) = (t_m + 1)/(t + U)$, where t_m is the number of w_m appears in training documents, t is the total number of feature words appears in training documents.

C. Comparison of MBM Model and MM Model

MBM model use feature words as the document attributes, all feature words contribute to the classification result no matter whether the word appears in the document or not. What's more, MBM model only takes into account whether the word appears in the document, but does not care how many times the feature word appears.

MM model takes into account the frequency of the feature word appears in the document. For a feature word does not appears in the test document, it will not contribute to the classification result.

McCallum etc.[18] did a lot of experiments to compare the MBM model and MM model. They conclude that, if the document is short, MBM model performs well; If the document is long, MM model can get better result, where 27% error is reduced than the MBM model.

IV. Combined Model Used in Text Classification

As stated above, MBM model takes into account the contribution of all feature words while frequency is not considered; MM model considers feature words frequency but some feature words are neglected.

Since MM model neglect those feature words that does not appear in the test document, the classification result of MM may be determined by few feature words. It may be unfit in some cases, and that is the reason why MM could not get better results in short text classification cases. What is worse, if the number of feature words are limited and each word is very important, the performance of MM model will be bad even in long text classification.

To overcome these problems, we propose a new model combined both MBM and MM models, which is more flexible than the traditional methods and is fitted to be used in more cases.

A. Combined Model

In this paper, we proposed a model combined both MBM and MM model together, which not only take into account the word frequency, but also considered those important feature words that do not appear in the document.

Given a set of feature words $(w_1 \quad w_2 \quad \ldots\ldots \quad w_U \quad f_1 \quad f_2 \quad \ldots\ldots \quad f_V)$, word $w_1 \quad w_2 \quad \ldots\ldots \quad w_U$ appears in the test document D , whose frequency is $n_1 \quad n_2 \quad \ldots\ldots \quad n_U$. $f_1 \quad f_2 \quad \ldots\ldots \quad f_V$ does not contained in the document. The document is composed by the events of w_i appears n_i times, and f_i does not appear.

The probability of $P(D \mid C^j)$ can be expressed by,

$$P(D \mid C^j) = n! \prod_{m=1}^{U} \frac{P(w_m \mid C^j)^{n_m}}{n_m!} \bullet V! \prod_{k=1}^{V} \left(1 - P(f_k \mid C^j)\right) \quad (8)$$

where $n = \sum_{m=1}^{U} n_m$, n_m is the frequency of word w_m appears in document D. $P(w_m \mid C^j)$ can be expressed by $P(w_m \mid C^j) = (t_m + 1)/(t + U)$, where t_m is the number of w_m appears in training documents, t is the total number of feature words appears in training documents. $P(f_k \mid C^j)$ can be calculated in the same way.

B. Classification process

In our application, two sets of words are selected. The first set of words E are selected by manual experiences, the other set of words W are selected by automatic method. The word number of E is very small, which is decided to be very important to the classification task. The set W is larger than E , it was selected by an automatic algorithm[19]. Words in E has been excluded from the set W .

Document D is composed by two sets of words W and E . According to the assumption of naïve Bayes, W and E are conditional independent, so $P(D \mid C^j)$ can be expressed by

$$P(D \mid C^j) = P(D^W \mid C^j) \bullet P(D^E \mid C^j) \quad (9)$$

Where D^W means document D is represented by words set W , and D^E means the document is represented by words set E .

Then, we use MM model to describe $P(D^W \mid C^j)$, using the main part of equation 7. And we use combined model to describe $P(D^E \mid C^j)$ with equation 8.

Substitute equation 9 into equation 5, the classification result can be got.

V. Corpus and Experiments

We use a corpus from Yahoo(http://news.yahoo.com). On Yahoo, news have been categorized. Six categories news are selected, they are business, entertainment, health, politics, science and sports. The documents in each category are divided into two parts, training documents and test documents. The number news in each category is listed in table 1.

TABLE I. CORPUS

	Training Document	Testing document
Business	15	13
Entertainment	13	7
Health	16	11
Politics	11	13
Science	19	14
Sports	12	11

We use MM method and our method in chapter III to train the document, then test the testing documents as chapter IV. The experiments results are shown in table 2.

TABLE II. CLASSIFICATION RESULT

Accuracy	MM method	Our method
Business	84.62%	100.00%
Entertainment	71.43%	85.71%
Health	90.91%	90.91%
Politics	92.31%	92.31%
Science	85.71%	92.86%
Sports	90.91%	90.91%

From our results, we can see that, the accuracy rate of testing data is significantly increased by the method described in this paper.

Our corpus should be downloaded from the webpage http://blog.sciencenet.cn/?828693 soon.

VI. CONCLUSION

In traditional MBM model or MM model used in naïve Bayes method, the word frequency or some important words does not play a role in the final classification process, which result some limitation to each method.

In this paper, we proposed a new probabilistic model for text classification. This method not only takes account of the frequency of feature words, but also considers those important words do not appear in the test document. So this method is more flexible to be used in application. Experiments show that the result of this method is more accurate than the traditional MM model in long text classification (MM model has better results than the MBM model in long text classification).

This method can be made more flexible. Some variables can be added to adjust the importance rate of word sets W and E in equation 9, so this method will get better result in different application environment. But how to adjust those variables in different environment is an important problem. This is our work in the future.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China(Project Number 61171193), the Technology Porject of MOUHURD of China(Project Number 2010-K9-25), and the Natural Science and Humanities Society Foundation of BUCEA(Project Number 101102915).

We greatly thank to Bin Xu for his help in this work.

REFERENCES

[1] D. Lewis and Gale. A sequential algorithm for training text classifiers. In Proceedings ofACMSIGIR Conference, 1994.

[2] David D. Lewis and Kimberly A. Knowles. Threading electronic mail: A preliminary study. Information Processing and Management, 33(2):209-217, 1997

[3] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the World Wide Web. In Proceedings of AAAI-98, 1998.

[4] M. J. Pazzani, J. Muramatsu, and D. Billsus. Syskill fc Webert: Identifying interesting Web sites. In AAAI-96, 1996

[5] M. B. Amin and S. Shekhar. Generalization by neural networks. Proc. of the 8th International Conference. on Data Engineering. April 1992

[6] S.M. Weiss and C. A. Kulikowski. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufmann, San Mateo, CA, 1991.

[7] David D. Lewis, Robert E. Shapire, James P. Callan, and Ron Papka. Training algorithms for linear text classifiers. In Proceedings of the 19 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages pages 298–306, 1996

[8] B. Masand, G. Linoff, and D. Waltz. Classifying news stories using memory based reasoning. In SIGIR-92, pages 59–64, 1992.

[9] C. Cortes and V. Vapnik. Support vector networks. Machine Learning, 20:273–297, 1995.

[10] Nguyen H S. Diseretization problem for rough sets methods. Proceeding of lst International Conference of Rough Sets and Current Trends in Computing. 1998:545−552.

[11] Ramoni M, Sebastiani P. Robust Bayes classifiers. Artificial Intelligence, 2001, Vol.125, No.122, 209-226.

[12] Nguyen H S, Skowron A. Quantization of real values attributes, rough sets and boolean reasoning approaches. Proceeding of the Second Joint Annual Conference on Information Science. Wright Sville Beach, NC. 1995: 34−37.

[13] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In SIGIR-94, 1994.

[14] D. Lewis and M. Ringuette. Comparison of two learning algorithms for text categorization. In Proc. of the Third Annual Symposium on Document Analysis and Information Retrieval, 1994.

[15] D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In Tenth European Conference on Machine Learning, 1998.

[16] K Ming, A Chai, HT Ng, HL Chieu. Bayesian Online Classifiers for Text Classification and Filtering. Proceeding SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. 97-104

[17] EH Han, G Karypis. Centroid-Based Document Classification: Analysis & Experimental Results. 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, 424 - 431, 2000

[18] A McCallum, K Nigam. A comparison of event models for naive bayes text classification. In AAAI/ICML-98 Workshop on Learning for Text Categorization, 41-48.

[19] Chen Jing-nian, Huang hou-kuan, Tian Feng-zhan, Qu You-li. Feature selection method used in text Bayes classification. Copmputer Engineering and Applications. 2008, Vol 44, No 13, 24-32.