TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.2, February 2014, pp. 947 ~ 952
DOI: http://dx.doi.org/10.11591/telkomnika.v12i2.4180

A New Naive Bayes Text Classification Algorithm

Duan Li-guo, Di peng*, Li Ai-ping
Department of Computer Science and Technology, Taiyuan University of Technology, China
Taiyuan City of Shanxi Province, Yingze West Street, No.79, Taiyuan University of Technology, 030024,Ph:03516010071
Corresponding author, e-mail: tyutdlg@163.com, 353967364@qq.com

Abstract

Aiming at the phenomenon that in text classification the calculation of prior probability is time-consuming and has little effect on the classification results and the error propagation of posterior probability affects the accuracy of classification, this paper improves the classical naïve bayes algorithm and proposes a new text classification algorithm which accelerates the speed by removing the calculation of prior probability and reduces the accuracy loss of error propagation by adding an amplification factor. The experiments prove that removing the calculation of prior probability can accelerate the classification speed obviously and has little effect on the classification accuracy, and adding an amplification factor in the calculation of posterior probability can reduce the effect of error propagation and improve the classification accuracy.

Keywords: *naive bayes, amplification factor, error propagation, prior probability, posterior probability*

1. Introduction

With the development of technology and the increase of information on internet, text classification has become a hot research topic in modern information processing. As an important content of text mining, it refers to determining a category for each document in the collection of documents in accordance with the subjects which have been defined earlier, enabling people to find the required information and knowledge. With the rapid growth of text information, especially the explosion of online text information, text classification has become a key technology in organizing and processing large amount of document data, being widely applied in areas such as spam filtering, e-government and information retrieval, etc [1].

Bayes, together with LLSF, SVM, KNN and Decision Tree, are the most well-known text classification method. And in the current study of text classification, Liu Congshan of the Shanghai Jiao Tong University proposes a novel algorithm based on K Nearest Neighbor [2]. The algorithm defines a class imbalance factor and uses NCA to learn a Mahalanobis distance measure and has a good experiment result. Zhang Chunying of the Hebei University of Technology proposes a new weighted naïve bayes classification algorithm [3]. The algorithm uses the identifiability matrix of Rough Set and assigns different weights in different conditions. The experiment proves that the calculation of this algorithm is easier and more effective. Zhong Jiang of Chongqing University proposes an algorithm based on normalized vector. This algorithm makes the three-dimensional feature space of training samples projected onto the two-dimensional feature space, which can decrease the dimension of features and improve the classification accuracy [4].

In the present study, nobody improves the calculation of prior probability and no one finds the error propagation in the calculation of posterior probability. Therefore, the main purpose of this study is improving the calculation of prior probability and reducing the effect of error propagation in the calculation of posterior probability.

2. Related Theories of Bayes

Before the introduction of the application of naïve Bayes algorithm in text classification, it is necessary to know some related theories of Bayes in order to fully understand the Bayes algorithm.

Received July 19, 2013; Revised August 296, 2013; Accepted September 24, 2013

2.1. Total Probability Formula

The first concept needed to be introduced is conditional probability [5]. Given two events of a random experiment A and B with P(B)>0, the conditional probability of A given B is defined as the following equation:

$$P(A \mid B) = \frac{P(AB)}{P(B)} \tag{1}$$

And the definition of total probability formula is as follows:

Given an event A in a random experiment E and a finite partition B_1 ,B_2 ,…,B_n of its sample space Ω with P(Bi)>0 (i=1, 2, …n), the probability of any event A of the same probability space is [5]:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + ... + P(A \mid B_n)P(B_n) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i) \tag{2}$$

According to this formula, if event A happens with any of the Bn events, each with a given probability itself, the total probability of event A can be calculated by adding each probability together.

2.2. Bayes Formula

Given an event A in a random experiment E and a finite partition B_1 ,B_2 ,…,B_n of its sample space Ω, the definition of Bayes formula is [5]:

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)} \tag{3}$$

2.3. The Principle of Naïve Bayes Algorithm

P(h) refers to prior probability, which is the initial probability of h without data training and reflects what is known about h as an established assumption. Without this knowledge, a same prior probability can be given to each assumption in actual processing. Similarly, *P(D)* indicates the prior probability of the training data D which will be observed. And *P(D/h)* is the probability of the appearance of data D when h is established. In machine learning, *P(h/D)* is what we are interested in, which is to calculate the posterior probability, probability of h given a training data D. It reflects the confidence about the establishment of h based on the training data D.

According to Bayes formula,

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \tag{4}$$

As a constant variable independent of h, *P(D)* can be removed. Therefore, *P(h|D)* is determined by *P(D|h)P(h)*, which is the core of naïve Bayes algorithm. The next thing is, given training data, to find the assumption h with the biggest probability from the candidate sets H. In other words, it is about, given a training data, how to learn from this sample data so as to classify new data into a certain category.

Usually a data contains multiple attributes, for example, data D may have attributes a1,a2…an. Naïve Bayes algorithm is based on such an assumption: the given target value's attributes are independent of each other. This assumption indicates that when a target value is given, the joint probability of the attributes equals to the product of each attribute's individual probability. Then we have the following formula:

$$P(D \mid h) = P(a_1, a_2 ... a_n \mid h) = \prod_{i=1}^{n} P(a_i \mid h) \quad (5)$$

So the posterior probability in the naïve Bayes algorithm can be expressed as [6]:

$$P(h \mid D) = P(h) \prod_{i=1}^{n} P(a_i \mid h) \quad (6)$$

3. Research Method

Based on the above theory, the design process of naïve Bayes classifier will be introduced in detail from the following aspects: segmentation of Chinese words, feature extraction and the design of naive Bayes classifier.

3.1. Segmentation of Chinese Words

When given a Chinese text, the computer can only identify it as a long string. So the first problem is to enable the computer to identify the text more accurately. This requires some preprocessing of the text so that the computer can identify each word more accurately, which will play a vital role in the following work [7].

Unlike English, there are no spaces between Chinese words, which will necessitate word segmentation before further processing [8]. In the design process of classifier, various techniques are adopted, including JAVA and Myeclipse8.5, particularly the ICTCLAS which were developed by the Institute of Computing Technology of Chinese Academy Of Sciences. ICTCLAS1.0 has won the first prize in the evaluation activities organized by domestic 973 expert groups, and ICTCLAS2.0 has also won the first prize in the evaluation of SigHan. Therefore, ICTCLAS may be the best Chinese lexical analyzer around the world, with a word segmentation speed of 500KB/s, accuracy rate of 98.45%, API of less than 100kb and compressed dictionary data of less than 3M.

3.2. Feature Extraction

Feature extraction is to extract the words that can obviously represent the category of the text and to remove the useless words. Since word segmentation will result in a large vector with each word representing an attribute of the text, Naïve Bayes algorithm assumes that given the target value, the attributes are independent of each other. This ignorance of the dependent relationship among lexical items makes feature extraction extremely important, since if feature extraction cannot return the lexical items which can identify a certain category, the efficiency of naïve Bayes algorithm will be damaged [9]. Usually there are many stop words in Chinese text, which refer to words with a high frequency but little significance, including auxiliaries, adverbials, functional words and prepositions, such as “de”, ”zai”. If these words can be removed after word segmentation, the dimensions of text vector will be greatly lowered and the efficiency of calculation will be greatly improved.

At present, the main methods of feature extraction include the ones that are based on evaluation function, on correlation and on semantic meaning [10, 11]. Different from the past feature extraction which usually involves a very complex algorithm, this study applies a Chinese stoplist which were provided by HIT-CIR. After the word segmentation, each word is compared with the stoplist for removal, a simple and convenient process with greatly improved efficiency.

3.3. Design of Naïve Bayes Classifier

Then we will discuss how naïve Bayes algorithm is applied to text classification. According to Bayes algorithm:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \quad (7)$$

The task of Naïve Bayes classifier is to classify the text into the closest category $C(C_1,C_2 \dots C_j)$ according to the text vector $X(x_1,x_2 \dots x_n)$. $X(x_1, x_2 \dots x_n)$ represents the feature vector of the text to be classified, while $C_1,C_2, \dots C_j$ are the categories which have been defined. Therefore, the computation involved is about the probability $(P_1, P_2 \dots P_n)$ when $X(x_1, x_2 \dots x_n)$ belongs to $C_1, C_2 \dots C_j$, with P_j being the probability when $X(x_1, x_2 \dots x_n)$ belongs to C_j. Then max $(P_1, P_2 \dots P_n)$ is the demanded result.

According to naïve Bayes algorithm, it is possible to get the following formula:

$$P(C_j \mid x_1, x_2,..., x_n) = P(x_1, x_2,..., x_n \mid C_j)P(C_j) \quad (8)$$

In this formula, $P(C_j)$ is the prior probability when the text belongs to C_j, $P(x_1,x_2 \dots x_n|Cj)$ is the posterior probability that C_j contains the text vector $(x_1,x_2 \dots x_n)$ when the text to be classified belongs to C_j. So max $(P_1,P_2,\dots,P_n)$ is equal to the maximum value of the following formula:

$$\arg\max_{C_j \in C} P(x_1, x_2,..., x_n \mid C_j)P(C_j) \quad (9)$$

According to the assumption of Bayes, the attributes $(x_1,x_2,\dots,x_n)$ are independent of each other. Then the joint probability is equal to the product of the probability of each attribute. So the final classification function is:

$$\arg\max_{C_j \in C} P(C_j) \prod_{i=1}^{n} P(x_i \mid C_j) \quad (10)$$

In this formula, $P(C_j) = \frac{N(C = C_j)}{N}$, $N(C = C_j)$ is the quantity of text which belong to Cj in the training set, and N is the total quantity of the text in the training set. $P(x_i \mid C_j) = \frac{N(X_i = x_i, C = C_j) + 1}{N(C = C_j) + M}$, $N(X_i=x_i,C=C_j)$ is the quantity of text with the attribute xi of C_j, $N(C=C_j)$ is the quantity of text which belong to Cj, and M is the dimension of text vector.

The above is the main design process of naïve Bayes classifier, in which two places have been improved.

3.4. Improvement and Innovation

Usually, when given a Chinese text, we do not know which category it belongs to, which requires the text to get the same prior probability for each category according to the principle of fairness. It is unfair and unreasonable to regard the prior probability different because that the amount of different kinds of texts in training sets are different. Therefore, it is reasonable to remove the calculation of prior probability and equals to assign the same prior probability. Then we can obtain the following classification function:

$$\arg\max_{C_j \in C} \prod_{i=1}^{n} P(x_i \mid C_j) \quad (11)$$

Because the maximum probability is what is needed, removing the calculation of prior probability will not affect the final classification result, but greatly speed up the calculation.

In the study, a variable of double was defined to hold the posterior probability. Sometimes the text to be classified is very long, making the dimension of the text vector much more and the posterior probability very small. And after all the attributes' probabilities are multiplied, the result may be zero, so an error propagation appears. To solve this problem, we enlarge the posterior probability of each feature attribute to a certain multiple. This will not affect the experiment results, because what matters in the end is the comparison of probability between categories and it is more convenient if the probability is multiplied. At first, we

magnified 10 times; but in the experiment, we found that sometimes the posterior probability will beyond the scope of the variable of double, which will greatly affect the experiment results. Finally, we added an amplification factor K in the function to reduce the effect of error propagation. And how to determine the value of K will be shown in the next part. So the classification function is:

$$\arg\max_{Cj \in C} \prod_{i=1}^{n} K * P(xi \mid Cj) \tag{12}$$

4. Results and Discussion

The design of naïve Bayes classifier needs a lot of training sets and test sets. We download a corpus from the Sogou Laboratory and use the corpus of its reduced version. It contains 9 categories: finance and economics, IT, health, sports, tourism, education, job wanted, culture, military affairs, each with 1990 texts. Some texts were randomly selected from each category as the training texts and some as the test texts. Accuracy rate was used as the evaluating indicator.

$$Accuracy\ Rate = \frac{correctly\ classified\ texts}{total\ texts\ numbers} \tag{13}$$

The computer we use in the experiment is hp compaq 6515b, with the cpu of AMD Turion 64, memory capacity of 1G, and CPU Clock Speed of 2.2GHz. The experiment results are shown in form 1. And the accuracy rate before and after improvement are shown.

Table 1. Classification Result

category	Training texts	Test texts	Before	after	Trend
Finance	700	400	0.9000	0.9125	↑
IT	650	370	0.8973	0.8973	—
Health	610	330	0.9152	0.9242	↑
Sport	580	310	0.9129	0.9355	↑
Tourism	540	290	0.9103	0.9172	↑
Edu	510	270	0.8704	0.8778	↑
Job	530	300	0.8633	0.8867	↑
culture	550	310	0.8774	0.8903	↑
Military	570	320	0.9375	0.9375	—

Table 2. Calculation Time

Test texts	Before	after	increase
Finance	1 M 36 S	1 M 27 S	9 S
IT	1 M 28 S	1 M 12 S	16 S
Health	2 M 30 S	2 M 13 S	17 S
Sport	2 M 2 S	1 M 48 S	14 S
Tourism	1 M 18 S	1 M 02 S	16 S
Education	2 M 5 S	1 M 48 S	17 S
Job	1 M 35 S	1 M 21 S	14 S
Culture	1 M 2 S	0 M 47 S	15 S
Military	1 M 44 S	1 M 26 S	18 S

It can be seen from the experiment that the accuracy rate of naïve Bayes classifier is higher except when the category is IT and Military. Besides, we selected a text from each category as test text and calculated the time before and after improvement. The experiment results are shown in form 2. According to it, the time of calculation is obvious shortened after improvement, changing with the number of the words in the test text.

The determination process of the amplification factor K is shown in form 3. In the experiment, we choose 100 test texts randomly from each kind of the texts. And if the

calculation result of posterior probability of a test text is 0, we will mark the text. K is assigned to 3, 4, 5, 6, 7, 8, 9, 10 successively. And we record the quantity of text which are marked.

Table 3. Amplification Factor Experiment

multiple / kind	3	4	5	6	7	8	9	10
finance	9	8	5	5	5	6	7	7
IT	8	7	6	5	6	7	7	7
health	10	7	4	5	5	7	8	9
sport	8	7	6	6	7	8	8	8
tourism	9	8	5	6	6	7	7	8
education	9	7	7	6	7	8	9	9
job	8	6	4	5	6	6	7	8
culture	7	6	5	5	6	7	8	9
military	6	4	4	5	5	6	6	7

According to the experiment, we can see that when the value of amplification factor K is 4 or 5 or 6, the number of the texts which are marked is less and the experimental result is better.

5.Conclusion

Naïve Bayes algorithm is a simple and effective method to design a text classifier with high accuracy rate and fast speed. However since naïve Bayes is an algorithm based on machine learning, the accuracy rate is largely dependent on the training set. Therefore, how to establish the training set and how to determine the value of amplification factor more precisely will be a future research area for us. Besides, when the test text contains some complex sentences, the accuracy rate will be lower. So this is also a research area for us in the future.

Acknowledgement

This work is financially supported by the Natural Science Fund Project in Shanxi Province (2013011015-2) and State key lab of software engineering open subject project (SKLSE2012-09-30).

References

[1] Chen Ye-wang, Yu Jin-shan. An Improved Text Classification Method Based on Bayes. *Journal of Huaqiao University (Natural Science)*. 2011; 32(4): 401-404.
[2] Liu Cong-shan, Li Xiang-bao, Yang Yu-pu. Text Classification Algorithm Based on Neighborhood Component Analysis. *Computer Engineering*. 2012; 38(15): 139.
[3] Zhang Chun-ying, Wang Jing. A new Weighted Naïve Bayesian Classification Algorithm. *Microcomputer Information*. 2010; 26(10-3): 222-223.
[4] Zhong Jiang, Sun Qi-gan, Li Jing. Text Classification Algorithm Based on Normalized Vector. *Computer Engineering*. 2011; 37(8): 47.
[5] Sheng zhou. Editors. Probability theory and mathematical statistics. Beijing: Higher Education Press. 2010.
[6] Tom M Mitchell. Zeng Hua-jun, Zhang Yin-kui. Machine Learning Beijing. China Machine Press. 2003.
[7] Wei Xiao ning, Zhu Qiao ming,Liang Xing yan. Using Bayesian in Text Classification with Participle method. *Journal of Suzhou Vocational University*. 2008; 19(1): 104-107.
[8] Fakir FM. Segmentation and Recognition of Arabic Printed Script. *IAES International Journal of Artificial Intelligence(IJ-AI)*. 2013; 2(1).
[9] Yu Fang, Jiang Yun fei. A Feature Selection Method for NB_based Classifier. *Acta Scientiarum Naturalium Universitatis Sunyatseni*. 2004; 43(5): 118-120.
[10] Zhang xiao yan, Song Li ping. A Survey on the Method of Feature Selection in Text Categorization. *Journal of Modern Information*. 2009; 29(3): 133.
[11] Enikuomehin T, Sadiku JS. Text Wrapping Approach to natural Language Information retrieval using significant Indicator. *IAES International Journal of Artificial Intelligence (IJ-AI)*. 2013; 2(3).