The Application of Naive Bayes in Text Categorization

Xie Hongjie[1,a,*]

[1]Department of Science, North Carolina State University, 3001 Hillsborough Street, Raleigh, North Carolina, the United States
a.Hongjiexie97@gmail.com
**corresponding author*

Keywords: naive Bayes, text categorization

Abstract: The thesis reviews the application of naive Bayesian classification method in various text categorization. Each experiment has adopted naive Bayes classifier or combined naive Bayes with other categorization methods to classify huge and tedious data, which reflects the effectiveness of naive Bayesian method.

1 Introduction

Text categorization is a rather significant branch of natural language classification, which has very important applications in the field of Internet public opinion, commodity review and analysis, as well as news. A typical feature of text data is that the dimension is very large. For instance, in an article, there may be thousands or even tens of thousands of words. However, documents of various types or topics are quite different in vocabularies. It is not necessary to consider the order in which vocabularies appear, namely employ the bag of words model. Assuming that the emergence of each word in the text is independent, naive Bayesian method can be used at this time to solve the problem.

2.1 Text Categorization Steps

The first is to carry out word segmentation. Chinese language is different from Western languages. Chinese language has no natural separation. Hence, word segmentation ought to be conducted. Meanwhile, there are many function words in Chinese language that need to be filtered and ignored. Secondly, divide training set and test set. We hope to achieve the best effect of predicting real data values by training this huge data set through a trained model, that is, the prediction error of the real data should be minimized. The third is to count the word frequency, which is the frequency of each word in all training sets, and then sort them in a descending order. The fourth is the stop list. Fifthly, extract text features. Select several feature words, and delete words as well as numbers in the stop list. The sixth is training and prediction.

2.2 TF-IDF

TF is the number of times that appears in a document, while IDF is a measure of the universality of a word. DF is the frequency of occurrence for a certain keyword, and |D| is the total number of files. The more common a word, the larger the denominator in the formula, and the smaller the IDF.

$$IDF = \log\left(\frac{|D|}{1+DF}\right) \qquad (1)$$

2.3. Naive Bayesian Algorithm

$$P(Y|X) = P(X|Y) * P(Y)/P(X) \qquad (2)$$

If X is regarded as a text vector, Y is a text corresponding sign.

Spam classification

After acquiring the relevant data, we can first mark the mail as spam and non-spam.

Carry out processing and word segmentation of the training data. Filter the stop words.

Use the tf-idf method for text vectorization.

Train the model.

Obtain the test results for the body content of the mail.

3.1. Naive Bayesian Model Application on Weibo Topic

By using the naive Bayesian model, Feng Junjun designed and implemented a tracking system of Weibo topic, in the research of Weibo topic tracking technology in 2017. Experiments show that naive Bayes is of simple classification and high efficiency. Consequently, it is particularly suitable for tracking hot topics on Weibo. In terms of the 2015 CnPameng's Weibo corpus, 50 of the 300 Weibo on specific topics were selected as training samples, and another 250 as test samples. The 50 training samples included five different topics, and each topic corresponded to 10 Weibo. Then he used NLPIR Chinese word segmentation system to separate the words and got the dates, places, people, numbers and proper nouns in each Weibo. The tracking results of Weibo can be obtained from the false alarm rate and missing report rate of naive Bayes model as well as their weighted results.

3.2. Chinese Text Categorization Based on Naive Bayes

Jiang Tianyu proposed in Chinese categorization based on naive Bayes in 2019 that the construction and interpretation of naive Bayes and its good performance can be widely used to solve categorization and sorting problems. Experiments demonstrate that naive Bayes can effectively distinguish and identify the cluttered information and quickly meet the needs of users. Jiang Tianyu selected 1,596 e-mails from Xinhua News Agency as experimental materials, which were divided into six categories, that is, environment, transportation, military, economy and sports. He performed word segmentation on the 1596 e-mails, removed the meaningless words such as adverbs and function words, and recorded the remaining words with high frequency of occurrence to form a matrix, namely, the word frequency matrix. After the TF-IDF feature selection, several words with the highest TF-IDF were extracted. The text was classified by this method, and it is found that the precision and reconciliation rate of each article is very high. The rate is also very fast, reaching up to 800 articles per minute.

3.3. Chinese News Screening

In the research on Chinese toponymy extraction algorithm for Chinese event news in 2019, Liu Jiaqi put forward the extraction method of location where the event news occurs, based on gazetteer and Naive Bayesian algorithm. The experiment also suggests that naive Bayesian classification combined with a gazetteer has a precision and recall rate of more than 90%. In this application, the attributes of the place where the event news occurs include whether there are adjacent common words in a sentence, whether the place name appears at the beginning of the sentence and the occurrence rate of the place name. Liu Jiaqi selected 1300 training sets and 800 test sets to test the effectiveness of adjacent words and algorithms. It can also be informed from this experiment that the gazetteer can extract 64% toponymy keywords, and the precision rate is above 96%. As a result, this algorithm can be used to extract toponymy keywords.

3.4. Fire Text Categorization

Based on the uneven distribution of various fire texts in August 2019, Ge Jike proposed a fire classification algorithm that improved naive Bayesian categorization. This method extracts and represents the feature entries of fire text in a standardized way, which solves the problem that few category features are covered and classification accuracy is greatly affected by the scale of feature word set when the number of training sets is small and the distribution of categories is uneven. The raw data set of this study is from the national fire yearbook of the Public Security Fire Bureau from 2008 to 2016, and 755 investigation reports of major fires across the country are selected. From the perspective of causes, these data are divided into six categories, including electric, production operation, domestic fire accidental, human error fire play, arson and other reasons. It is shown that these data samples are quite different and unevenly distributed. After the training set text is preprocessed, it combines ICHI method and T-IDF weighting algorithm to absorb and represent feature words. Then through modeling, a classifier based on the improved naive Bayesian algorithm is created to classify the trained text, so as to obtain the classification results. After introducing Kappa coefficient, it is found that the classification model based on naive Bayes has better accuracy and Kappa coefficient. Thus, it improves the accuracy of fire classification and enriches the research content of safety management engineering.

3.5. Naive Bayes to Solve the Problem of Electric Power Infrastructure Construction

In 2019, Xie Zhiwei presented a text categorization method based on naive Bayes to solve the problem of electric power infrastructure construction overawed by the difficulties and retreat, aiming at the huge data. It realizes automatic classification and greatly improves the convenience. He has a total of 1,000 construction problem sets in the infrastructure department of a power supply bureau. Via data preprocessing, redundant and repetitive data is removed. Finally, 800 texts are selected as experimental samples. The classification categories include human factors, equipment factors, regulatory factors, and environmental factors. Additionally, numbers are labeled. He used Jieba word segmentation tool in python for word segmentation, then converted the text into a feature vector space, quantified the text data with word frequency, and finally built a classification prediction model. In order to prove its better performance, SVM model and KNN classification model were respectively established for comparison. The classifier method based on naive Bayes solves the huge problem data of construction and low efficiency of monitoring and management, which improves work efficiency to some extent, and provides technical support for the refined management of power grid.

3.6. Navie Bayes and Python on Custom Classification of Commodity

In 2019, Lv Yan took advantage of the efficient and easy training features of naive Bayes, and applied python language and its program to realize the custom classification of commodity. He collected half a million labeled raw data sets of questions from the 10th China university innovation competition. Firstly, he preprocessed the raw data, filtered out the missing data and removed Chinese stop words. Then, the counter vectorizer feature words in sklearn database were used for quantitative calculation, and the fit_transform function in python was used to convert the information of feature words into a word frequency matrix. The article presents the prediction results after fitting with polynomial Bayesian classifier according to the obtained standardized data. The mean value accuracy is relatively high, and the prediction effect is favorable. Experiments show that naive Bayesian classification is of stable efficiency and high prediction accuracy, and it is insensitive to the missing data. However, it is not suitable to use this method when the number of feature attributes is large and the correlation between them is high.

3.7. Navie Bayes Application on Text Sentiment Classification

In the study of text sentiment classification in 2019, Liang Ke compared the methods of ordinary logistic regression and naive Bayesian classification, finding that the recall rate and accuracy rate of naive Bayesian classifier model were higher than those of logistic regression classification. He selected 25,000 English film reviews on Kaggle net as raw data for case analysis, dividing them into positive and negative categories through investigation. The first step is to preprocess the data, remove the html tags, cut into words, take out stop words and reorganize the data. Then, extract the feature vectors by using countvectorizer in sklearn. In the use of logical classification and naive Bayesian classification of bag of word and word2vec, it is found that the accuracy rate of naive Bayesian classifier is 0.027% higher than that of logistic regression classifier, and the recall rate is 0.028 % higher too.

3.8. Naive Bayesian to Analyze the Data Resources of Archives Bureau of Gansu Province

In 2018, Liu Peixin exploited the naive Bayesian classification method to analyze and study the data resources of Archives Bureau of Gansu province. It is proved that the classification model is appropriate for the classification of archive text resources, which realizes the function of automatic classification. Firstly, he selected 900 test anthologies, which can be divided into six subject categories, and then filtered 2700 training sets. Then the data was preprocessed by using TF-IDF. After the experiment, it is found that naive Bayesian model has a better classification efficiency in the classification of archive texts, with about 85% recall rate and around 80% precision rate. Among them, after the rough sort based on reference number, the Naive Bayesian algorithm is used for disaggregated classification, which reduces the calculation amount of TF-IDF algorithm and naive Bayesian algorithm, thus saving the classification time.

3.9. Navie Bayes Application of Network Question Answering Feedback

In 2016, Jiang Liqun used the naive Bayesian method to exploit a network question answering feedback system for java courses. It assists teachers in answering questions, and it can also classify and present students' questions to teachers, which greatly improves the teaching quality and questions answering efficiency. In the article, a search engine based on Lucene full-text search and dictionary-based Chinese word segmentation technology is used to design a question answering system. In the classification algorithm, the naive Bayesian classification algorithm is used for classification. The system collected 60 documents for each category of Java knowledge point. Each document covers point description, definition, questions answering, as well as other contents of knowledge points, which is enough to improve information classification. Then use the sample data to extract, train and classify feature words. After marking, students can retrieve the answer to the question by key words. Through this experiment, it is also concluded that the network question answering feedback system adopting the naive Bayesian text classification algorithm can achieve rather accurate text classification. The application in the Java network question answering system is of great help to practical teaching.

3.10. Web Page Pre-Classification Method Based on Naive Bayesian Classification

In the research on user behavior entailment in 2018, Qin Peng proposed a web page pre-classification method based on naive Bayesian classification, specific to the characteristics of low accuracy rate, low recall rate and low classification efficiency in traditional classification methods. The algorithm extracts relevant websites according to users' online activities, analyzes web content and keywords, and classifies them by using naive Bayesian method. The results show that this method is more accurate, which can fully explore users' hobbies and interests. It can be used as a data algorithm of user behavior analysis for commercial promotion and judicial forensics. The training text data he used was the web text set provided by SouGou, with a total of 10 categories, including culture, email, music, etc. Data sets such as tests were related web page texts downloaded by users when browsing the URL. For each category of training set, 2000 web page texts were used as training sets. During the test, 100 URLs were selected from each category for test. The total URLs were 1000, and the results were obtained.

3.11. Naive Bayes-Based Cultural Tourism Text Categorization

In 2018, Wang Xiangxiang studied the naive Bayes-based cultural tourism text categorization technology. According to the characteristics of cultural tourism texts, a cultural special subject thesaurus is firstly constructed. The scenic spot description text is transformed into a vector by using the vector space model. Lexical feature selection is carried out through information gain, and weight assignment is conducted by using the word frequency-inverse document frequency. The classifier model is constructed to realize the automatic classification of tourism texts. The experiment selected 1447 scenic spot description texts, which is classified according to Minnan, red-tourism, Hakka and ecological culture. Based on the acquisition of corpus, this research adopts Scikit-Learn for experiments, and it uses Grid Search to select the optimal parameters. The accuracy rate of naive Bayes is as high as 91.07%. Compared with SVM, DT, LR, NB, and NB2 algorithms, naive Bayes is the optimal algorithm. In addition, after improving the feature weight assignment, the classification accuracy of the method is improved by 1%. Through this classification technology, and according to the characteristics of different cultural thematic categories, it provides convenience for the construction of cultural thematic databases and the design of cultural thematic tourism products, promoting the development of cultural tourism.

3.12. Feature Independence Assumptions of Naive Bayesian Algorithm

In 2019, Ren Shichao published the feature independence assumptions of naive Bayesian algorithm, pointing out that traditional TD-IDF weighting algorithm ignores the relationship between features and categories as well as documents, causing the problem that the weights given to features by traditional methods represent their accuracy. It proposes naive Bayesian classification algorithm of two-dimensional information gain weighted. Moreover, it further considers the two-dimensional information gain of the feature. Compared with the traditional method by designing experiments, it is found that the accuracy rate, recall rate, and F1 index performance increase by 6%. The experimental data were selected from the internationally common 20_NewsGroup data set, which included 20 categories. Six of which were randomly selected, and then 100 documents were sorted from each category. Hence, there would be 600 documents. By using the cross-validation method, 360 documents were randomly picked up as the training set, whereas 240 documents as the test set. After the preprocessing of experimental data, using the four indexes of precision rate, recall rate, F1 and macro F1, it is concluded that the naive Bayesian classification algorithm of two-dimensional information gain weighted can classify the text more effectively.

3.13. Thematic Web Page Recognition Based on Improved Naive Bayesian

In 2018, Ma Jin carried out a research on thematic web page recognition based on improved naive Bayesian algorithm, which has better effect on the recognition of text theme. The experimental corpus was 1,443 military and non-military web pages extracted from Sina.com. Naive Bayesian classifiers assume that attribute values are mutually conditional independent of each other when given class tags. That is, in the case of a given instance, the joint probability is the product of the probabilities of each individual feature. The classification formula of naive Bayesian classifier can be obtained as follows:

Their calculating formula is as follows:

$$c(a) = \arg\max_{c \in C} P(c) \prod_{j=1}^{m} P(a_j \mid c) \tag{3}$$

Firstly, extract the subject content of the web page, and then perform word segmentation. Finally, use the selected features to vectorize the segmented text. It is proved that naive Bayesian algorithm can effectively identify the theme of web pages.

4. Conclusion

Based on Naive Bayesian algorithm, this thesis automatically classifies the data in different fields. Compared with other classification methods, naive Bayesian classification method has significantly improved the efficiency of classification results. The problem solved by naive Bayesin introduced in this thesis is very close to life, which reflects that the application of naive Bayesian will be more extensive.

References

[1] Feng Junjun, He Xiaochun, Wang Haipei. Research on Weibo Topic Tracking Technology Based on Naive Bayesian Network. Computer and Digital Engineering, 2017 (11).

[2] [Jiang tianyu, Wang su, Xu wei. Research on Text Catogorization Based on Naive Bayesian. Computer Knowledge and Technology, 2019(23).

[3] Liu jiaqi, Luo yonglian, Research on Chinese Place name Extraction Algorithm for Chinese Event News. China Computer and Communication, 2019(15).

[4] [Ge Jike, Chen Dong, Wang Wenhe, Chen Zuqin, Chen Guorong, Liu Can. Research on improved naïve Bayes classification Algorithmic fire classification, 2019(19).

[5] Xie Zhiwei, Feng Honghuai, Xu Ruiqi, Li Huifu, Research on Text classification on Power Infrastructure Construction Problem. Modern Information Technology, 2019(17).

[6] Lv Yan, Cheng Shuling, Research on Automated Classification Method of Commodity by Naive Bayes and Implementation with Python. Journal Of Anqing University(Science Edition) 2019(25).

[7] Liang Ke, Li jian, Chen Yinxue, Liu Zhigang, Research on Text emotional classification and realization based on Naive Bayes. Intelligent Computer and Applications 2019(5)

[8] Liu Peixin, Yu Hongzhi, Xu Tao, Research on archives text clasification based on Naive Bayes, Journal Of Hebei University(Natural Science Edition), 2018(38).

[9] Jiang Liqun, Java Curriculum Network Answering System Based on the Naive Bayesian Classification. Computer Knowledge and Technology, 2016(12).

[10] Ren Shichao, Huang Ziliang, Naive Bayes Classification Algorithm of Feature Weighting Based on Two-Dimensional Information Gain.Computer System and Application, 2019,28(6).

[11] Ma Jing, Zhu Yanhui, Liu jin, Tian Hailong, Research on web page Recognition Based on Improved Naïve Bayes Algorithm. Imformation and Communication, 2018(183).

[12] Wang Xiangxiang, Fang yun, Chen Chongchen, Classification technique of cultural tourism text based on naive Bayes. Journal of Fuzhou University, 2018(5).

[13] Qin Peng, Cao Tianjie, Behavior derivation of users based on Naive Bayes web page classification. Journal of Shenyang University of Technology, 2018(1).