Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Research on Text Classification Based on Feature Selection Algorithm

Kebin Cui [a] , Yaqian Huang [b, *]

Department of Computer Science, North China Electric Power University, China

[a]ncepuckb@163.com, [b] yaqian_huang@163.com

Abstract: With the development of information technology, text information data is growing explosively. Effectively obtaining useful information from numerous text data is a problem worthy of study. Aiming at this task, this paper proposes a text classification model based on feature selection algorithm, and integrates data mining technology to classify, select, and verify text data, and use various methods to filter and summarize new useful information from the original information. In the end, the internal laws of information that people expect are obtained. This article mainly introduces the feature selection from the naive Bayes classifier and the Filter algorithm, and introduces the specific implementation of the chi-square algorithm in detail, so as to realize the corresponding feature algorithm, obtain comprehensive features about the text, and then classify the text based on this feature. Experimental results show that this method can extract text features more accurately and has higher classification accuracy.
Keywords: Text; feature selection; Bayesian classifier; text classification.

---

1. INTRODUCTION

1.1 Research background and significance

The practical significance of text classification in engineering is huge, including in many fields such as mail classification, digital library, content search, information filtering, etc. Therefore, the research of text classification has very important practical significance. Incorporate data mining technology into the research of text classification, and find the rules between data from a large amount of text information. Data mining technology is to process this information according to certain rules from a large amount of information containing noise and redundancy. These rules include a series of operations such as classification, selection, verification, and discarding, and finally get the internal laws of information that people expect [3-4], in fact, it is to filter and summarize new useful information from the original information through various methods, scientific analysis and processing, and finally obtain the real effect hidden in the data, thereby digging out new and undiscovered Effective information.

Classification is an important branch in the field of data mining. Its significance lies in the feature learning of a sample prepared in advance to obtain a subset. In turn, the learned model is used to predict the class label of the unknown case. For model optimization, the feature dimension is directly

proportional to the number of subsets. In real life, the size of the subset is regulated. The classification model is used to extract features that can represent group information to improve the efficiency and accuracy of classification. However, many feature algorithms face the problem of "dimensionality disaster". There are many irrelevant or redundant features in the higher-dimensional feature sample set, which will make the problem more and more complicated, which makes it more difficult to understand and solve text classification. Corresponding question in.

Feature Selection (Feature Selection), also known as Feature Subset Selection (FSS), or Attribute Selection (Attribute Selection), refers to selecting a feature subset from all features to make the constructed model better. In order to obtain a model with a better structure, it is necessary to perform statistics and analysis on the original sample, perform grouping and learning, and extract the features of the subset with the same characteristics. At the same time, it is also necessary to ensure that the features are as small as possible under certain accuracy conditions. The dimension is reduced. Once the dimension is reduced, it will greatly reduce the complexity of the calculation and speed up the calculation. In order to reduce the dimension, it is necessary to ensure that the data is separable, and at the same time, some closely related features are discarded and separated , You can make the feature selection work faster and more accurate. The final use of the inductive learning algorithm based on the feature selection method in the feature subset generation process is mainly composed of the following two aspects: one is the Filter model, when it is looking for features and learning, its data set is irrelevant to others; two It is the Wrapper model. The efficiency of this model is used as the criterion for feature selection when the criterion for feature selection is performed. This paper mainly focuses on the study of the Filter model, designing feature selection algorithms such as Bayesian classification, chi-square and information gain, and on this basis, the optimal feature selection algorithm is selected according to the actual application.

For example, text classification can be understood as a process of categorizing a certain product into a certain category, and computer code can indicate the automatic classification of a certain event. The application range of text classification is very wide. For example, here is an item, and some of them will associate with the meaning of this item. For a text classification, this is actually of little use, and this article can understand it. It is a classification based on the theme, but in fact, the classification of the text can also determine the shape, color, placement, and year of the item. It can be understood that if an event is not only related to classification, text or features, it is called text classification. In the field of text classification, keywords are still the most used. Searching based on keywords is widely used in search giants such as Baidu, Google, and Sogou. Text classification can only be searched based on keywords, and cannot be classified based on content such as time, address, and source. This requirement is very effective for these content in the website.

1.2 Research summary

The history of the discovery of feature selection should be traced back to after 1960. At that time, most experts and professors would study this issue through computing. However, due to the development of economy and computer technology at that time, the characteristic field is not only the research object of experts in the field of statistics, but also experts in other disciplines such as computer science, algorithm field, and engineering field. It is not only based on algorithms. The core of is also a pre-processing applied to training.

Filter-type feature selection method is a very important type of feature selection method. This method is divided into two different types: single feature evaluation and feature subset evaluation. Single feature evaluation includes calculation and mutual information, information gain and other feature selection algorithms; feature evaluation is the feature evaluation obtained by the smallest feature subset that satisfies the important metric. In general, feature subset evaluation has greater advantages than single feature evaluation [13].

Embedded feature selection algorithm is based on a specific classification of feature selection, similar to Wrapper type feature selection method. Compared with the Embedded feature selection algorithm, the evaluation object of the Wrapper feature selection method is a specific classification learning algorithm. The corresponding feature evaluation function is established for these feature subsets, and then specific analysis is performed through transparent classification. At present, the Wrapper-type feature selection algorithms that are widely used and frequently used are as follows: RFE and its improved algorithm [12-14] (supports elimination of backward features of the vector machine), HITON [16] algorithm (any specified classifier) , Markovblanket [15] algorithm, forward/backward Wrapper algorithm (Forward/Backward Wrapper).

In view of the characteristics of the Wrapper-type feature selection method and the Filter-type feature selection method, related scholars have studied a hybrid feature selection algorithm that combines the characteristics of the two.

Bayesian algorithm is concerned with the probability that a document belongs to a certain category. The probability is equivalent to the formula for the probability of words belonging to this category in this document. And this probability can be calculated by the number of times in the word training document, so it is possible to use this algorithm in the entire calculation process. When Bayesian algorithm is used, its task in the training phase is to estimate these values.

1.3 Paper arrangement

The second section of this article mainly describes the overview of Bayesian networks and Filter-type algorithms; the third section mainly introduces the experiments and core code implementation based on text classification, mainly using java to implement the chi-square feature algorithm and related experimental results; Section 4 mainly analyzes and summarizes the results of this model.

2. SUMMARY OF NAIVE BAYES AND FILTER ALGORITHMS

Bayesian network specifically analyzes and processes information that cannot be accurately defined or cannot be accurately targeted. Graph theory is a way of expressing conclusions by Bayesian network. In the field of knowledge discovery and data mining for a long time, Bayesian network has been regarded as an important research topic. It is a comprehensive theory based on probability theory, statistics and graph theory. The core of the theory is the Bayesian formula.

2.1 Bayesian theory foundation

Conditional probability formula: Let A and B be two random events of random trial E, and P(B)>0, say

$$p(A \mid B) = \frac{p(AB)}{p(B)}$$

The above formula refers to the conditional probability of event A when event B occurs.

Prior probability: Assuming that An, A1, A2... is an event in a sample space, if P(Ai) can be obtained based on prior knowledge estimation, then P(Ai) is called the prior probability.

Posterior probability: Posterior probability is one of the basic concepts of information theory. In a communication system, after receiving a certain message, the probability that the receiving end knows that the message is sent is called the posterior probability.

The posterior probability refers to the probability of re-correction after obtaining the information of the "result", as in the Bayesian formula, it is the "cause" in the problem of "finding the cause", the prior probability and the posterior probability are inseparable The calculation of the posterior probability should be based on the prior probability.

Bayesian formula description: Suppose the sample space of random experiment E is Ω, Bi (i=1,2,...,n) is a finite division of Ω, and P(A)>0, P(Bi)>0 , There are

$$p(B_i \mid A) = \frac{p(B_i)p(A \mid B_i)}{\sum_{j=1}^{n} p(B_j)p(A \mid B_j)}$$

2.2 Naive Bayes Classifier

Naive Bayes classifier can actually be understood as the eigenvalue is not related to the elements of the Curry, so it can be known that this limit is very large, in reality it is difficult to find the above conditions, but once the naive Bayes classifier is applied If so, the classification is very easy, and the effect is very good, and the price is high. In simple terms, the naive Bayes classifier assumes that each feature of the sample is not related to other features, that is, the selection of each feature is independent.

The naive Bayes classifier relies on the natural probability model to obtain good classification results in the training and learning samples. In many practical applications, the Naive Bayesian model parameter estimation uses the maximum likelihood estimation method. In other words, the Naive Bayesian model can work without Bayesian probability or any Bayesian model.

Naive Bayes Classification Model

In theory, the probability model classifier is a conditional probability model.

$$p(C \mid F_1,...F_n)$$

The independent categorical variable C has several categories, and the conditions depend on several characteristic variables. But the problem is that if the number of features n is large or each feature can take a large number of values, it becomes unrealistic to list probability tables based on probability models. So this article modifies this model to make it feasible. Bayes' theorem has the following formula:

$$p(C \mid F_1,...,F_n) = \frac{p(C)p(F_1,...,F_n \mid C)}{p(F_1,...,F_n)}$$

The numerator in the above formula is equivalent to the joint distribution model.

The “naive” conditional independence assumption comes into play: Assume that each feature Fi is conditionally independent of other features Fj,j≠i. This means

$$p(F_i \mid C, F_j) = p(F_i \mid C)$$

For i≠j, the joint distribution model can be expressed as

$$p(C \mid F_1,...,F_n) \otimes p(C)p(F_1 \mid C)p(F_2 \mid C)p(F_3 \mid C)... \otimes p(C)\prod_{i=1}^{n} p(F_i \mid C)$$

From the above, it can be seen that the advantages of Naive Bayes classifier are: first, the program needed to implement the algorithm is easy; second, the program is stable; third, it takes a short time.

2.3 Filter-type feature selection algorithm

The feature ranking method in the filter-type feature selection algorithm and the optimal feature selection algorithm based on feature subsets are introduced and explained. At the same time, the characteristics, calculation methods and shortcomings of the two algorithms are clarified.

2.3.1 Filter feature selection algorithm

As a widely used feature selection method, the feature ranking method is a typical representative of the Filter-type feature selection algorithm. The feature evaluation object of the feature ranking method is not group, but single: firstly by evaluating and scoring a specific feature; secondly, sorting according to the score of each feature; finally, the features of the previous position in the sorting Become an important basis for the selection of this feature, and artificially endow this feature with a strong class distinction ability. The feature ranking method in the filter-type feature selection algorithm needs to sort each feature in the feature space, and is not suitable for different feature subsets sorting [4].

At present, the most widely used calculation method of feature sorting is the IG feature selection algorithm [19], which is a calculation method with extremely high operating efficiency. The sorting principle is to determine the information gain of each feature in the feature space and the information gain of the class label set. Since the scale of the training set has a certain linear relationship with the information gain, the calculation of this formula shows that the IG feature selection algorithm has a high execution efficiency, which means that the use of this calculation method can obtain more accurate results in a short time . Therefore, it meets the requirements of the feature sorting method for operating efficiency [9].

In addition to the IG feature selection algorithm, Relief and its series of improved algorithms [20] are also commonly used calculation modes in feature ranking methods. The basic principle of Relief is: first clarify the inter-class spacing and sample intra-class spacing, and secondly perform feature evaluation based on the interaction between the two. Among them, it means that the neighbor samples in the same label have very close feature values; the long distance means that the neighbor samples in different types of labels have very different values. On this basis, judge the contribution of each feature to the sample distance, so as to obtain the corresponding weighted score (the feature distance of the sample and its nearest hit sample is less than the feature distance of the sample and its nearest error sample, then the feature weight will be reduced, and vice versa. Feature weight). After repeated calculations of feature average weights, the final feature weight values can be obtained, and the magnitude of the value indicates the sorting situation [15].

For the feature ranking method, its significant advantages are mainly manifested in the high efficiency of execution. In addition, its shortcomings are also prominent. Since this method ignores the output of feature subsets, the number of selected features must be artificially counted. Therefore, the selection of a reasonable number is very important in the calculation process.

2.3.2 Filter-type optimal feature sorting method

To effectively complete feature selection, the core content is the selection of the selection algorithm for searching the optimal feature through feature subsets. Compared with the feature ranking method, the difference of this method is that it comprehensively considers the correlation between the feature and the class tag set and the redundancy between the feature and the feature, which is intended to be able to obtain the "best m features".

For the selection algorithm based on the feature subset to search for the optimal feature, the main search strategies selected for the feature subset include greedy search (greedy search), random search (random search), and best-first search (best-first search), etc. [12] . The greedy search strategy is further classified, including search strategies such as forward/backward [17] and floating order [8]. At present, the most commonly used strategy is the greedy search strategy [19], and it is also the easiest method to produce local optimal phenomena. In the random search strategy, in order to eliminate the local optimal phenomenon, the method of adding random factors into the search is usually adopted, and the typical one is genetic algorithm. Taking the relevance of feature and class tag set and the redundancy between features and features as the starting point, the main evaluation methods of feature subsets include: comprehensively evaluating the relevance and redundancy of feature subsets, and then independently evaluating feature subsets Relevance and redundancy.

3. TEXT CLASSIFICATION ALGORITHM IMPLEMENTATION

Common text classification algorithms: document frequency, cross-correlation information, information gain and chi-square statistical algorithm. This article uses java language to experiment with chi-square statistical algorithm during implementation, so this part mainly describes the chi-square statistical algorithm.

Chi-square statistics, as a hypothesis testing method, is often used in the field of information processing to perform feature selection in classification problems, to indicate the correlation between a feature and its category, and to evaluate the category membership of a feature. When the chi-square value is large, the feature is more important for category judgment; conversely, the feature is less important for category judgment. The formula for calculating the chi-square statistic of feature word t and category ci is as follows:

$$CHI(t,c_i) = \frac{N_{all}[p(t,c_i) * p(\bar{t},\bar{c}_i) - p(t,\bar{c}_i) * p(\bar{t},c_i)]}{p(t) * p(\bar{t}) * p(c_i) * p(\bar{c}_i)}$$

For the chi-square statistics of all categories of feature t, the average value is often used as the final chi-square statistics, as shown below:

$$CHI(t) = \sum_i p(c_i) CHI(t,c_i)$$

In the feature selection stage of text classification, the null hypothesis is generally used as "the vocabulary t is not related to the category c". The larger the square root value calculated, the greater the deviation from the null hypothesis, and the more we tend to think that the null hypothesis is The negative situation is correct. The selection process calculates the square root value of category c for each word, sorting from large to small (the larger the square root value is, the more relevant it is), and

the top k ones can be used. Therefore, the larger the chi-square value, the more relevant the vocabulary and classification.

In this experiment, the main() method is used as the entry method, which mainly refers to all the methods involved, and constructs the part of the judgment method that realizes the feature selection. Of course, this part also involves the chi-square function to achieve the acquisition of experimental data. The final realization in this experiment: a feature corresponds to a final chi-square value, and this value is used to determine the feature of the text or the value that appears.

4. CONCLUSION

Through consulting related materials, the research in the field of text classification is discussed in depth. This article is a text classification research based on feature selection algorithms. It mainly discusses the Naive Bayes classification and Filter feature selection algorithms in depth. Reasonable selection of the corresponding feature selection algorithm or further improvements to the algorithm will greatly reduce the use Time, improve the efficiency of text classification. Through this study and research, I have a deeper understanding of the mining and utilization of data information, and I hope that I will have further study and research on data mining related fields in the future.

REFERENCES

[1]Y. Aphinyanaphongs, LD Fu, Z. Li, ER Peskin, E. Efstathiadis, CF Aliferis, and A. Statnikov, “A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization,” J. Assoc . Inform. Sci. Technol., vol. 65, no. 10, pp. 1964–1987, 2014

[2]T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprintarXiv:1301.3781, 2013

[3]Bo Tang, Student Member, IEEE, Steven Kay, Fellow, IEEE, and Haibo He”Toward Optimal Feature Selection in NaiveBayes for Text Categorization”VOL. 28, NO. 9, SEPTEMBER 2016

[4] Chen Yujie, Harbin Institute of Technology, master's degree thesis "Research on Feature Selection Algorithms in Text Classification", 2015

[5] Zhang Aihua, Jing Hongfang, Wang Bin, et al. Research on the role of feature weighting factors in text classification[J]. Journal of Chinese Information Processing, 2010, 24(3): 97-104.

[6] Liu Y, Loh H T, Sun A. Imbalanced text classification: A term weighting approach[J]. Expert systems with Applications, 2009, 36(1): 690 -701.

[7]Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents,” arXiv preprint arXiv:1405.4053, 2014

[8]B. Tang and H. He, “ENN: Extended nearest neighbor method for pattern recognition [research frontier],” IEEE Comput. Intell. Mag., vol. 10, no. 3, pp. 52–60, 2015 .

[9] Chen Xiaoyun, Chen Qi, Wang Lei, Li Ronglu, Hu Yunfa. Frequent pattern text classification based on classification rule tree [J]. Journal of Software, 2006, 05: 1017-1025.

[10] Zhang Lixin PhD thesis of Tsinghua University "Feature Selection of High-Dimensional Data and Integrated Learning Research Based on Feature Selection" 2004.

[11] Gu Ping, Chongqing University Ph.D. Thesis "Research on Document Classification and Related Technologies Based on Bayesian Model" 2006

[12 Li Xinjian, Wang Chengtao, Xu Liang, Wuhan Wendao Information Technology Co., Ltd. Research and improvement of feature extraction methods based on Bayesian [J]. Police Technology, 2015, 02: 1017-1025.

[13]Nanjing University Information Science A review of text classification based on Naive Bayes_He Ming[J] 2015

[14]Pinheiro R H W, Cavalcanti G D C, Correa R F, etal. A global-ranking local feature selection method for text categorization[J]. Expert Systems with Appli-cations, 2012, 39(17): 12851-12857.

[15] Feng Guohe, Zheng Wei. Summary of Research on Feature Dimensionality Reduction of Text Classification[J]. Library and Information Service, 2011, 55(9): 109-113.
[16] Song Fengxi, Gao Xiumei, Liu Shuhai, et al. Dimensionality reduction and low-loss dimensionality reduction in statistical pattern recognition[J]. Chinese Journal of Computers, 2005, 28(11): 1915-1922.
[17]Li Gang, Xia Chenxi, Zheng Zhong. Comparison and improvement of local text feature selection algorithms[J]. Journal of Information, 2008, 27(4): 506-511.