Research on Text Combination Classifier Based on KNN and Bayesian Algorithm

Tongke Fan

Engineering college, Xi'an International University, Xi'an, China

Keywords: KNN; Bayes; Text classification; classifier

Abstract: Text classification is the basis of data mining, which can provide an important guarantee for us to effectively and accurately mine valuable information from a large number of text information, so how to quickly and accurately classify a large number of text is a key problem in data mining. Based on the research of text classification algorithms, this paper designs a combined classifier based on KNN and naive Bayes for the features and feature extraction methods of Chinese text dataset, which effectively improves the accuracy of feature vectors and the accuracy of classification methods, and makes up for the shortcomings of existing text classification methods. The experimental results show that the recall and precision of the text classification system based on the combined model are significantly improved.

1. Introduction

With the development of Internet and the explosive growth of all kinds of information, sending some information not only provides us with rich information, but also brings us troubles. How to quickly and accurately extract valuable information from massive information is very important. Text classification is the basis of data mining, which can provide an important guarantee for us to effectively and accurately mine valuable information from a large number of text information, so how to quickly and accurately classify a large number of text is a key problem in data mining [1].

2. Application scenarios of text classification

2.1 Document usefulness judgment

For binary text classification, the common application is to judge the usefulness of documents. For example, in the era of information security, users are disturbed by a large number of network information every day, and they will receive a large number of short messages or e-mails basedon advertising, cheating, pyramid selling and harassment. Users are affected by the information in their life and work, and at the same time they are faced with huge economic risks. Therefore, it is very necessary to identify and screen the above meaningless SMS or e-mails.

Based on text classification technology, electronic documents can be automatically analyzed and distinguished, and SMS or e-mail can be automatically divided into two categories: "useful" or "useless". Text classification can help users to better manage information, so that users' limited energy can focus on valuable information, and then improve the user's life and work efficiency [2-3].

2.2 Emotional analysis of word of mouth

In the field of e-commerce, users can browse a lot of online word-of-mouth information about products or services, which can help users make purchase decisions for online products. However, the number of online reviews of many products is tens of thousands, so it is difficult for users to browse and comprehensively analyze the comments in a limited time, and they often can only randomly select some comments to feel the overall view of the consumer group.

The method of manual browsing online word-of-mouth is time-consuming and labor-consuming,

  DOI: 10.25236/iccem.2021.029

and the conclusion is not objective enough [4]. Therefore, we can use text classification technology to divide the overall view of online reviews into two basic views: "positive" and "negative". Based on the classification results, users can generate more objective and accurate cognition of products or services according to online information.

2.3 Identification of negative information

Text classification technology can also be used to identify potentially harmful content and objects in the network environment. For example, text classification technology can analyze the content of a web page, so as to determine whether the web page contains pornography, violence, crime and other negative content. Classifying and filtering these negative content, on the one hand, can play a role in purifying the network environment, on the other hand, can help government managers identify the crisis in advance and take corresponding measures [5].

At present, text classification technology has been widely used in public opinion management. By monitoring the information of mainstream social media, the relevant departments can find the content and objects that are potentially harmful to society, and take social management programs and measures.

3. KNN algorithm

KNN uses the local information of samples instead of the global information of categories to judge the classification results. When classifying the samples, K samples nearest to the target sample are selected from the sample set as the reference, and then the main category of the K samples is taken as the classification result - the main category refers to the category that is marked the most times in the data set.

KNN method is very intuitive, but the number of adjacent reference samples K needs to be determined in advance. K-nearest neighbors is the classification result based on K nearest neighbor samples. KNN is defined based on vector distance, so it is necessary to realize the standardization method of vector and the definition of vector distance function in the algorithm.

When selecting K value, if the value is too small, the classification result will be unstable, and the classifier will be easily interfered by "noise" samples; If the K value is too large, the complexity of the algorithm is too high. In addition, in order to ensure that the main category is unique, the value of K is usually odd. In general, K is usually 3 or 5[6].

When determining the value of K, the main category can be obtained without direct counting, and the weighted value based on the sample distance can be considered

$$\mathrm{score}(c, d) = \sum_{d' \in s_k} I_c(d)\,\mathrm{Dis}(v_{d'}, v_d) \tag{1}$$

Among them, score (C: D) is the score of category C corresponding to document D. The score was calculated based on all the categories

$$c = \mathrm{argMax}\ \mathrm{score}(c, d) \tag{2}$$

KNN method does not need to model the training samples in advance, but only analyzes the training samples when predicting the classification results of samples. Therefore, KNN is also known as memory based learning or case-based learning. The disadvantage of KNN algorithm is that the response time is long, because a lot of algorithm work is delayed until the prediction stage of samples, The time complexity of KNN algorithm is a linear function of the size of training set [7].

From the function structure of the model, both naive Bayesian model and Rocchio method are linear classifiers, while KNN belongs to nonlinear classifiers. In many cases, the linear classifier is lack of expression ability and easy to be interfered by noise, so the nonlinear classifier can usually obtain higher accuracy. Therefore, KNN classification method is better than Nb and Rocchio.

4. Bayesian model

Bayesian classification is a statistical classification method, which is based on Bayesian theorem. It can be used to predict the possibility of class membership and give the probability that the text belongs to a specific category. When classifying, the text can be divided into the category with the highest probability according to the prediction results.

Naive Bayes assumes that in a case with many attributes, the influence of one attribute of text on classification is independent of other attributes, that is, the attributes of text are not related. This is the assumption of class condition independence introduced to reduce the computational overhead. Text D is represented by the characteristic words it contains, that is, d = (t_ 1,t_ 2,••,t_ j,· · ·,t_ n) , n is the number of characteristic words of D, t_ J is the j-th feature word.

$$P(d|C_i) = P\left(\left(t_1, t_2, \cdots, t_j, \cdots, t_n\right)\middle|C_i\right) = \prod P(t_j|C_i) \tag{3}$$

Where: P (D ⏐ C_ I) indicates that the classifier predicts the word t_ J in class C_ I is the probability of occurrence in the text. Therefore, equation (2-26) can be converted to:

$$P(C_i|d) \propto P(C_i) \times \prod P(t_j|C_i) \tag{4}$$

The training process of naive Bayes classification model is actually the process of statistics of every feature appearing in various types, with good speed and accuracy. Its success in many fields makes it widely used [8].

5. Algorithm adaptive model design based on KNN

The process of machine learning can be understood as the learning process of knowledge and the accumulation process of experience. Every algorithm traversal and selection process can be regarded as the accumulation of experience. When the number of accumulated training times is large enough, the system can be used as the accumulation of experience, The machine learning classification algorithm is applied to get the classification model by supervised learning according to the characteristics of the data set and the best algorithm corresponding to the data set[9]. When there is a data set, it is not necessary to traverse all the algorithms again to select the best one. Only by applying this model to analyze the dominant and recessive characteristics of the data set, the corresponding optimal algorithm can be obtained. This model is called algorithm adaptive model. In this paper, KNN is used as the training algorithm of adaptive model

For the training of a model, data set and machine learning algorithm are necessary, and adaptive model is not necessary. The data set for model training should not be too few, so the initial establishment of adaptive model can be completed when the user's data set is accumulated to a certain extent. The process is shown in Figure 1.

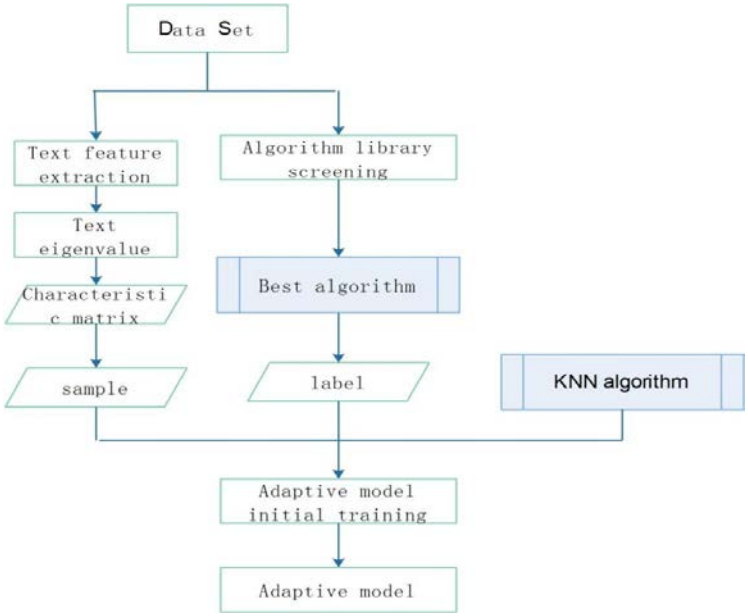

Figure 1 Flow chart of adaptive model initial training

After obtaining the dominant and recessive features of the data set through feature extraction, the dominant and recessive features are normalized and integrated to generate the feature matrix, which is used as the training sample of the adaptive model. Then the best algorithm corresponding to the current data set is selected from the algorithm library, which is used as the label of the adaptive model. In the aspect of algorithm, because the training set is the normalized value, we need to choose the suitable algorithm for the numerical classification. According to the research, SVM algorithm mainly solves the problem of two classification, but the performance of multi classification is poor. The adaptive model mainly classifies five algorithms in the algorithm library, so the selection of SVM algorithm can be excluded first.

6. Conclusion

Aiming at the user's application needs of the best model in the online machine learning education platform, this chapter proposes an automatic prediction scheme of the best algorithm for Chinese text classification, which can output the best algorithm and the best parameters according to the user's input data set, so as to form the best model for users to use. Aiming at the problem of incremental training of adaptive model, an incremental training mode of KNN algorithm based on weighted cluster is proposed. After the first generation of adaptive model, the incremental training of adaptive model can be completed according to the continuous expansion of user data set. At the same time, it also greatly reduces the memory occupancy and shortens the time-consuming of the algorithm prediction. At the same time, the best algorithm is predicted. In this chapter, the parameter prediction model based on Bayesian optimization is applied to predict the local optimal parameters of the best algorithm in a short time.

Acknowledgements

Xi'an International University 2020 "curriculum ideological and political" demonstration course construction project (No.:kcsz202014). The 13th five year plan of Educational Science in Shaanxi Province in 2020. "Research on the ecological mode and innovation path of computer course teaching in Shaanxi private colleges and Universities under the background of big data "(No: SGH20Y1424). Xi'an 2021 social science planning fund project (No.:JX52).

References

[1]Raj, S. Building Chatbots the Easy Way: Using Natural Language Processing and Machine Lcaming Building Chatbots with Python, 2019.

[2] Yu Minghua, Feng Xiang and Zhu Zhiting. The application and innovation of machine learning in the perspective of artificial intelligence. Journal of distance education, 2017, 35 (03): 11-12.

[3]Aladag. C. H. Anew architecture selection method based on tabusearch forarificia neural networks. Expert Systems with Applications, 2011, 38(4):3287-3293.

[4]Jaafra. Y, Laurent. J. L., Deruyver, A. Reinforcement llearning cture search: Areview. Image and Vision Computing, 2019, 89.

[5]Wistuba, M. Deep Leaming Architecture Search by Neuro-Cell-Based Evolution with Function-Preserving Mutations. Machine Learning and Knowledge Discovery in Database, 2018.[6] Wang fengxu, Li Qi, Han Qinglong. Research on MOOC learning performance prediction based on K-nearest neighbor optimization algorithm. Computer and digital engineering, 2019, 47 (4): 785-788.

[7]Wu F L, Zheng Y F. Adaptive normalized weighted KNN text classification based on PSO. Scientific Bulletin of National Mining University, 2016, (1): 109-115.

[8] Yang Y, Pedersen J Q. A comparative study on feature selection in text categorizaition. //Proceeding of the 14 International Conference on Machine Learning (ICML). Nashville, Tennessee, USA: IMLS, 1997. 412-420.

[9]Zhang Yin, Dai Miaolin, Ju Zhimin. Preliminary discussion regarding SVM kernel function selection in the twofold rock slope prediction model.Journal of Computing in Civil Engineering, 2016, (30)3.