Acta Technica *62* No. 2A/2017, 143–152

Application of improved Bayesian model based on cosine similarity weighted in prediction of disease classification[1]

Zhaochun Ran[2]

Abstract. With the development of information technology, Bias's classification prediction function has been gradually applied to finance, medical and other fields. Therefore, the application of Bayesian model based on cosine similarity weighted modified classifier in disease classification prediction was studied in this paper. By constructing an improved Bias model based on cosine similarity weighting and performing it on the Spark platform, the accuracy of the improved model was compared with that of the traditional Bias model. Taking hypertension and hyperlipidemia as an example, the prediction results of Bias classification under single machine and cluster mode were compared and analyzed. The results show that under the cluster Spark platform, the improved Bayes model has the highest efficiency in disease prediction classification.

Key words. Cosine similarity, weighted improvement, Bayesian model, disease classification prediction.

1. Introduction

With the continuous development of medical diagnosis, it has fully formed a reciprocal relationship with information technology. The development of computer technology provides a new solution for the development of the medical and health industry, which can be used in medical research, clinical medicine and basic medical treatment [1]. It helps to speed up the development of medicine, promote medical information, reduce medical expenses, and fully protect people's health. At present, many field experts have used computers to build a model for predicting specific diseases, and obtained good prediction results [2].

Classification is the most important part of machine learning research. The model

[1]This work was supported by the Hainan Natural Science Foundation of China, Project number: 20156231, project name: Study on health expert system model based on Bayes algorithm.

[2]Haikou College of Economics, Haikou, 571127, China

established by classification method can analyze unknown input model according to known classification knowledge, and finally determine the attribution category of input model. As an effective and practical forecasting model, Bias has been widely used in people's production and life [3]. In order to improve the classification effect of Naive Bayesian algorithm, scholars have considered the independence assumption condition weakening property. Bias algorithm based on cosine similarity has improved the simple Bias classification algorithm from the point of view of local learning and structure expansion. The frequent item sets mining algorithm is implemented by using cloud framework for large-scale sample data. Finally, the improved algorithm is applied to the problem of disease prediction [4].

2. State of the art

As a classification algorithm, the Naive Bayes algorithm has obvious advantages with a solid theoretical basis, high computational efficiency and high accuracy [5]. In order to reduce the complexity of the model, it is assumed that the Naive Bayes algorithm is independent of each other, which can effectively reduce the complexity of the computation process. In the independent assumption of attributes, the decision attributes for each delegate weight are equal (both are 1) [6]. However, this assumption of independence can be satisfied in very few cases. The conditional attributes are not exactly the same as the weights of the decision attributes, which can lead to lower classification accuracy [7]. To solve this problem, we use weighted Naive Bayes model to assign different weights to each conditional attribute, and relax the independent hypothesis, so as to improve the performance of the classifier on the basis of maintaining the original model. A weighted Bayesian algorithm based on cosine similarity is proposed, which takes different training samples into account to classify the decision weights [8]. Use cosine similarity to measure the distance of samples, and select the best subset of training samples. Moreover, use similarity values as training samples to train and modify Bayesian models.

3. Methodology

3.1. Improved Bayesian model based on weighted cosine similarity

For the traditional Naive Bias classification algorithm (NB), the conditional independent assumption is difficult to satisfy in practical applications. The association between attributes always exists, and it has a bearing on the results. The weighted Naive Bayesian classification algorithm is that the final classification is conducted based on different attributes, and then the corresponding weights are extended for the original algorithm to improve the performance of the Naive Bayes algorithm [9]. The weighted Naive Bias classification algorithm reduces the influence of conditional attribute independence by assigning different weights for different conditional attributes. The accuracy of the posterior probability is calculated as follows

$$P(C_i|H) = P(C_i) \prod P(ak|C_i)\,. \tag{1}$$

Improved Bayesian model based on weighted cosine similarity uses the cosine similarity as the weights to optimize the weights of attributes. In the training set A, each sample A_i contains n conditional attribute fields, each corresponding to a category. The training set can be represented by multiple conditions, attributes, and classes. Treat each condition attribute field as a random variable X_i corresponding to the conditional attribute fields, and treat the class as a random variable Y. The random variables are corresponding to the conditional attributes, and then the distributions of the two values are corresponding, that is to say, for the random variables, all the values contained in them are a_{ij} [10]. Through this transformation, the original training set can be represented as a set of random variables that satisfy a particular probability distribution. Then, the correlation between the training set attributes and categories (A_i, C) is determined by measuring the correlation between the two random variables (X_i, Y). The class attributes C and the characteristic attribute A_i of the training set can be represented by the decision attribute Y and the characteristic attribute X_i. Cosine similarity is introduced as a measure of the correlation degree between two random variables. Cosine similarity is a measure of the difference between two sample vectors. Firstly, the vector space is used to represent the attributes of the sample, and then the similarity between the two vectors is measured by calculating the spatial angle cosine of the two vectors [11]. The smaller the angle of the two vectors (the closer to 0), the greater the cosine of the vector is, which indicates the higher the similarity between them. For random variables X_i and Y, the similarity of two vectors is calculated by the cosine similarity formula.

$$\cos\theta = \cos < Xi, Y >= \frac{\sum_{i=1}^{n} a_{ij} \times y_j}{\sqrt{\sum_{1}^{n} a_{ij}^2} \times \sqrt{\sum_{j=1}^{n} y_j^2}}\,. \tag{2}$$

Correlation analysis is used to measure the degree of correlation between the two variables. Two variables need correlation in the correlation analysis. In machine learning, the correlation of two vectors is usually measured by similarity or distance.

The key to the weighted Bayesian algorithm based on cosine similarity is to find the cosine similarity weights between the conditional attributes and the category, which is denoted as NNB. Firstly, the data sets that need to be processed are discretized and filled with missing values. Secondly, for the classification phase, it is needed to jump to the classification step directly. If it is the training phase, the training sample data set is carried out. The third is the statistical table learning: according to the training sample in the data set, the number of training samples, attributes and their categories is counted.

The fourth is the cosine similarity weight learning: all training samples are traversed, the cosine similarity between each conditional attribute and classification class is calculated according to formula, and the value of formula (2) is taken as the weight coefficient of a_i [12]. Then the attribute weight statistics constitutes, and finally the classification is carried out. The training phase of the Bayesian model for

disease classification prediction is as follows.

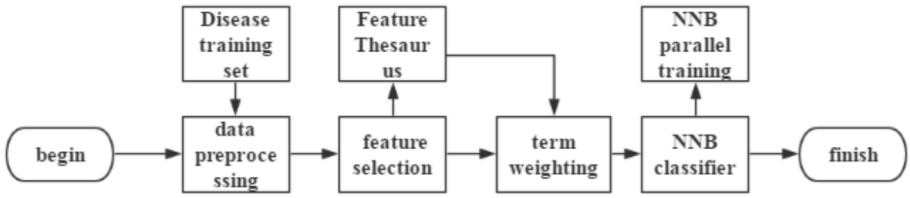

Fig. 1. Bayesian model training phase for disease classification prediction

The application phase is shown in Fig. 2.

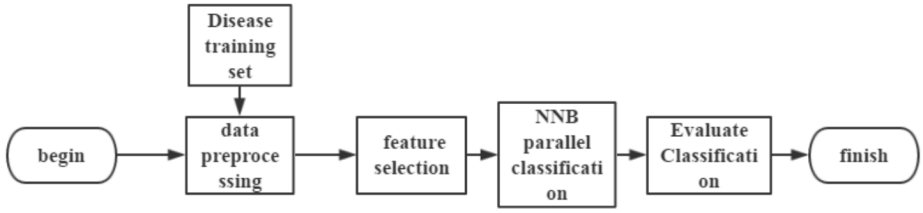

Fig. 2. Application phase of the Bayesian model test for disease classification prediction

3.2. Bias classification algorithm based on cloud computing platform Spark

The parallel Bias classification algorithm is implemented using cloud computing platform Spark. In parallel design ideas, it is similar to the "Map" and "Reduce" ideas in MapReduce. Firstly, we calculate the number of different characters in each category by "Map". Secondly, we calculate the parameters by "Reduce" superposition calculation. The parallelization process of Bias classification algorithm is realized by the Spark kernel scheduling of cloud computing platform. In the process of scheduling, the data sets are automatically allocated according to the number of nodes in the cluster, and then the tasks are executed in parallel to realize the parallelization of the Bias algorithm [13].

In the model training phase, it first reads the file from the HDFS (the file already processed), creates a new RDD, and performs caching operations locally to cache the RDD data. Then, by Map operations, mark each row, stack the value of the StackByKey operation, and count the frequency of each class, the number of classes as well as the frequency of each feature of each class. Finally, the model is trained by these parameters [14]. In the classification test phase, the mapping function is used to predict the classification of each test sample in parallel and calculate the final result. In the model training phase, the file is first read from the HDFS (a data set file that has been processed). Through the set function to obtain classification information, the data file is imported into the program. In the classification prediction phase, the calculation process of each classification can be computed one by one using Map, and then mapped directly to the output of the result, because the calculation of the

sample does not affect each other in the classification prediction stage. In addition to the classification of the prediction process, the training process, query and processing that are generated mainly by two frequency statistics tables [15].

4. Result analysis and discussion

4.1. Comparison and analysis of improved Bias algorithm based on cosine similarity weighting and traditional Bias algorithm

Different data sets (Letter, Lymphography, Segment, Credit-g data sets) were selected, the higher the average accuracy was, and the better the classification performance was. In the contrast experiment, the used classifiers included Naive Bayesian classifier algorithm (NB), Bayesian Network (BayesNet), and the improved Bias classification algorithm (NNB) proposed. The correct rate of each classifier after running was recorded.

Table 1. Accuracy comparison of experimental results based on different algorithm classifiers

Letter	0.7323	0.7393	0.7554
Lymphography	0.8022	0.7814	0.8443
Segment	0.8895	0.9141	0.9046
Credit-g	0.7542	0.7856	0.8343

The improved Bias algorithm based on cosine similarity weighted (NNB) has good performance in accuracy, and has a certain improvement compared with the traditional Bias classification algorithm and Bias network. Credit-g data sets have great dependence on each attribute, so the accuracy of NNB is obviously better than that of the former two algorithms. The improved Bayesian algorithm based on cosine similarity weighted (NNB) adds weights to the conditional attributes to improve the accuracy, but it needs to calculate the time cost of the weights. Therefore, the execution time of the improved algorithm is longer than that of the Naive Bayes algorithm. It can be concluded that the attributing weight based on the traditional Naive Bayesian algorithm can be carried out, so that the correlation between features and categories can be established by cosine similarity, which can alleviate the impact brought by the conditional independence assumption and improve the accuracy of disease classification to some extent.

4.2. Prediction analysis of disease classification under single machine condition

The disease classification prediction phase includes preprocessing, feature selection, classifier classification, and final result output. The classifier is constructed by training phase. The predicted categories of diseases are classified according to the classification and classified into the most relevant categories. Finally, the output is

evaluated according to the classification and evaluation criteria.

The experiment was done by single machine and cluster. A single experiment completed the calculation of the accuracy of the improved algorithm. The cluster mainly completed the cloud computing platform environment to improve the classification speed experiment.

A single NNB classifier was used, and the traditional Decision stump, C4.5, REPtree and Bayesian classification method based on cosine similarity weight prediction model were selected to forecast two kinds of common diseases in the elderly (high blood lipid and hypertension). The classification prediction accuracy and related error statistics are shown in Table 2.

Table 2. Prediction accuracy of classification and related error statistics

Classification method	Accuracy rate	Kappa statistics	Mean absolute error	Root mean square error	Relative absolute error	Relative square root error
Decision stump	58.97 %	0.4019	0.2541	0.3594	69.79 %	84.24 %
C4.5	82.05 %	0.7536	0.0914	0.2739	25.09 %	65.47 %
REPtree	82.05 %	0.7522	0.1326	0.2843	36.40 %	66.63 %
Improved Bayesian prediction model based on cosine similarity weighting	88.72 %	0.8071	0.1682	0.3013	40.76 %	64.38 %

The classification of diseases in this section was realized based on the traditional Bias classification and the improved classifier based on the improved classifier. The accuracy, recall rate and F1 value of disease category prediction under different feature dimensions were investigated. From the above table, the accuracy of Bayesian prediction based on cosine similarity weighted was higher, which reached 88.72 %. The predictive accuracy of Decision stump was low with only 58.97 %, and its predictive performance was poor because of the poor classification performance of hypertension or hyperlipidemia. In addition, the improvement of Bayesian NNB based on cosine similarity weighting was similar to that of NB in parallelization. Additional parallel computation of weighting coefficients was needed to realize NNB parallel algorithm.

It was assumed that the disease data set size was 200 thousand lesion data. In feature selection, the number of feature dimensions from 7000 to 11000 was set at intervals of 500, which was realized respectively through NB and NNB classification algorithm. Accuracy, recall rate, and F1 values are as follows.

Table 3. Classification prediction results of different feature categories of disease category

Characteristic dimension	NB			NNB		
	Accuracy	Recall rate	F1value	Accuracy	Recall rate	F1value
7000	0.8463	0.7935	0.8191	0.8979	0.8310	0.8632
7500	0.8501	0.8021	0.8254	0.9068	0.8367	0.8703
8000	0.8568	0.8095	0.8325	0.9132	0.8425	0.8764
8500	0.8621	0.8136	0.8372	0.9191	0.8487	0.8825
9000	0.8678	0.8186	0.8425	0.9234	0.8531	0.8869
9500	0.8724	0.8213	0.8461	0.9203	0.8501	0.8838
10000	0.8620	0.8203	0.8406	0.9168	0.8456	0.8797
10500	0.8603	0.8186	0.8389	0.9132	0.8413	0.8758
11000	0.8534	0.8103	0.8312	0.9086	0.8326	0.8689

From the above table, in the classification of diseases, the improved Bias classification algorithm based on cosine similarity was superior to the traditional Bias classification algorithm NB in NNB. For feature selection of different feature dimensions, the accuracy, recall rate and F1 value of NNB were higher when the feature dimension was 9000, while NB was higher when the feature dimension was 9500.

The accuracy rate, recall rate and F1 value of NNB and NB in the highest accuracy rate were demonstrated through the histogram form, so as to observe the experimental results more intuitively, as shown in Fig. 3.

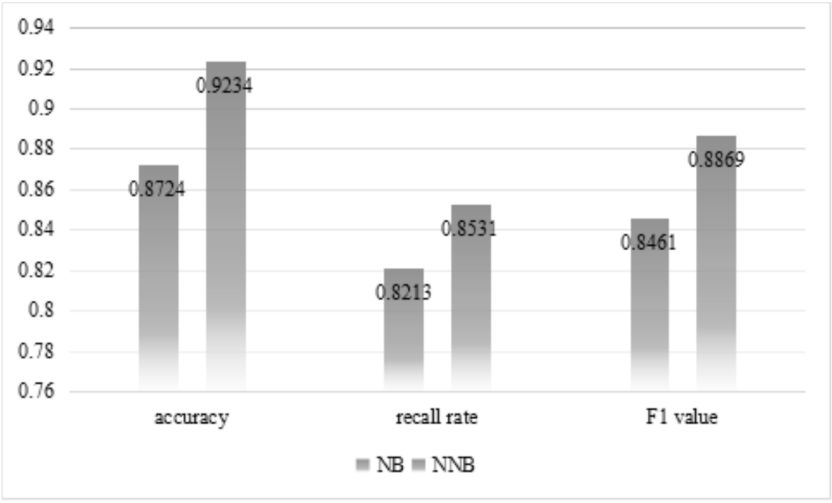

Fig. 3. Accuracy, recall rate and F1 value

Compared with the traditional algorithm NB, the cosine similarity weighted improved algorithm proposed in this paper has certain advantages in accuracy, recall

and F1 value of NNB. The performance of the classifier is improved by means of cosine similarity weighted attributes, and the accuracy of disease classification prediction is improved.

4.3. Prediction analysis of disease classification under cluster condition

The data set size was set, and the training time of different data sets in the cloud computing Spark platform was compared. The running time of the program was observed. Spark adopted 1 main node and 4 slave nodes. Run time results are shown in Table 4.

Table 4. Running time of different data in single machine and Spark

Data quantity (strip)	Single machine running time (s)	Spark run time (s)
5000	2.89	18.56
10000	5.65	23.81
50000	11.46	34.25
100000	24.51	41.39
200000	70.36	52.85
500000	273.75	75.54

The experimental results corresponding to the line chart are shown in Fig. 4.

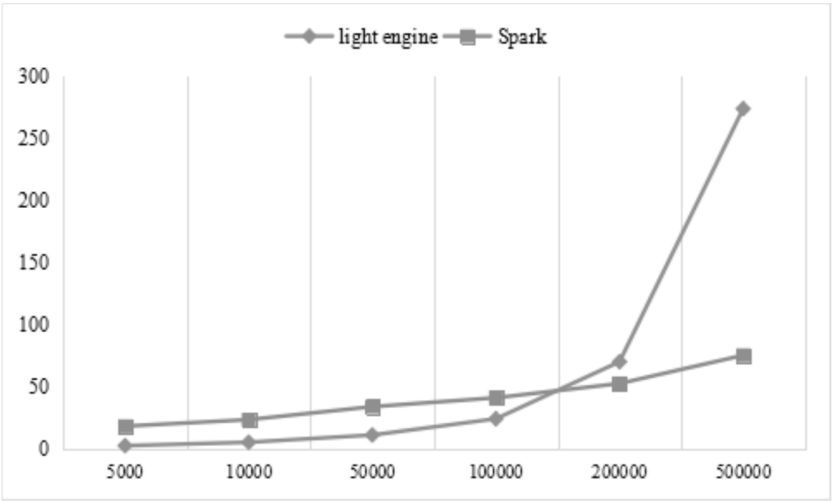

Fig. 4. Line chart of run time results

As can be seen from Fig. 4, in the case of small data (<10 million), the independent running speed of the disease classification program is significantly faster than that of the cloud computing platform Spark. It is mainly because in the cost of node communication, the communication and scheduling costs of the cloud computing platform Spark have relatively large proportion of the total running time with

a small amount of data. When the amount of data increases to some extent, the cloud computing platform Spark is more prominent at runtime than stand-alone. As can be seen from the diagram, when the amount of data reaches 200 thousand, the speed of cloud computing platform Spark is slightly higher than independence. When the amount of data increases to 500 thousand, the running time of program is obviously increased, and the number of cloud computing platform Spark is not increased. Therefore, the cloud computing platform Spark has obvious advantages in dealing with a large number of text classifications. When the data set increases to a certain extent, a single machine environment, such as a memory overflow problem, will appear. And because Spark cache mechanism and RDD conversion operation will greatly accelerate the implementation time, the parallel operation advantages of the Spark cluster are reflected. From the analysis of the current experimental results, it can be seen that the improved Bias classification algorithm based on cosine similarity weighted method has good performance in disease classification based on Spark platform. The improved algorithm is well ported to Spark. In the case of large amounts of data, the Spark based disease classification program can show good running speed, which indicates that the disease classification process works effectively in Spark.

5. Conclusion

Data mining has been widely used in many fields since it was put forward, and its technology is quite mature. As an important technology of data mining, classification technology plays an important role in practice. Nowadays, with the development of information technology, the hospital has gradually realized paperless office, and the database has accumulated a large amount of data. The use of data mining methods for disease classification prediction and clinical decision-making services has a special significance for chronic disease research. In this paper, the application of Bayesian model based on cosine similarity weighted improvement in the prediction of disease classification was analyzed. Firstly, the current situation of medical information mining technology and the research situation of Bayesian prediction algorithm in China were introduced. Secondly, the improved Bias model based on cosine similarity weighting and the Bias classification algorithms based on cloud computing platform Spark were introduced. Finally, the improved Bias algorithm based on cosine similarity weighted was compared with the traditional Bias algorithm, and the disease classification under single machine and cluster condition was predicted. The experimental results show that in the methods of using single classifier and ensemble classifier for disease prediction, the improved Bayesian model ensemble based on cosine similarity weighting has high classification accuracy of classifier, which still needs to promote the application of cloud computing platform in diversity.

References

[1] K. O. Akande, T. O. Owolabi, S. O. Olatunji: *Investigating the effect of correlation-based feature selection on the performance of support vector machines in reservoir characterization.* Journal of Natural Gas Science and Engineering *22* (2015), 515–522.

[2] G. C. Garriga, R. Khardon, L. D. Raedt: *Mining closed patterns in relational, graph and network data.* Annals of Mathematics and Artificial Intelligence *69* (2013), No. 4, 315–342.

[3] J. Nahar, T. Imam, K. S. Tickle, Y. P. P. Chen: *Association rule mining to detect factors which contribute to heart disease in males and females.* Expert Systems with Applications *40* (2013), No. 4, 1086–1093.

[4] F. Nori, M. Deypir, M. Hadi, K. Ziarati: *A new sliding window based algorithm for frequent closed itemset mining over data streams.* International eConference on Computer and Knowledge Engineering (ICCKE), 13–14 October 2011, Mashhad, Iran, IEEE Conference Publications (2011), 249–253.

[5] M. Cusumano: *Cloud computing and SaaS as new computing platforms.* Communications of the ACM, Technology strategy and management *53* (2010), No. 4, 27–29.

[6] S. Maurya, S. K. Shrivastava: *Kalman filter based flexible sliding window algorithm for mining frequent itemset over data stream.* International Journal of Computer Applications *111* (2015), No. 9, 13–19.

[7] M. Deypir, M. H. Sadreddini, M. Tarahomi: *An efficient sliding window based algorithm for adaptive frequent itemset mining over data streams.* Journal of Information Science and Engineering *29* (2013), No. 5, 1001–1020.

[8] J. Chen, B. Zhou, L. Chen, X. Wang, Y. Ding: *Finding frequent closed itemsets in sliding window in linear time.* IEICE Transactions on Information and Systems *E91-D* (2008), No. 10, 2406–2418.

[9] V. B. Wicaksono, R. Saptono, S. W. Sihwi: *Analisis perbandingan metode vector space model dan weighted tree similarity dengan cosine similarity pada kasus pencarian informasi pedoman pengobatan dasar di puskesmas.* Jurnal ITSMART *4* (2015), No. 2, paper 73.

[10] J. Zhu, Y. Ma, Q. Qin, C. Zheng, Y. Hu: *Adaptive weighted real-time compressive tracking.* IET Computer Vision *8*, (2014), No. 6, 740–752.

[11] Y. Wu, N. Jia, J. Sun: *Real-time multi-scale tracking based on compressive sensing.* Visual Computer *31* (2015), No. 4, 417–484.

[12] K. S. Lin: *Fuzzy similarity matching method for interior design drawing recommendation.* Review of Socionetwork Strategies *10* (2016), No. 1, 17–32.

[13] K. Khan, B. Baharudin, A. Khan, A. Ullah: *Mining opinion components from unstructured reviews: A review.* Journal of King Saud University - Computer and Information Sciences *26* (2014), No. 3, 258–275.

[14] A. Triana, R. Saptono, M. E. Sulistyo: *Pemanfaatan metode vector space model dan metode cosine similarity pada fitur deteksi hama dan penyakit tanaman padi.* Jurnal Teknologi Informasi ITSMAR *3* (2014), No. 2, paper 90.

Received July 12, 2017