ATLANTIS PRESS

Advances in Computer Science Research, volume 93

2nd International Conference on Mathematics, Modeling and Simulation Technologies and Applications (MMSTA 2019)

Prediction Model Hadoop-based for High-risk Students

Jichao Yu[1] and Xiaogao Yu[2,*]

[1]School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan China 430081

[2]School of Information Management and statistics, Hubei University of Economics, Wuhan China 430205

[*]Corresponding author

***Abstract*—In order to effectively predict high-risk students, this paper proposed a weighted voting combination prediction model based on Hadoop for high-risk students. Firstly, the design idea of the model was given, and the data of students were stored and processed by Hadoop platform. Secondly, according to the prediction method of students' characteristics and selection, a specific weighted voting combination prediction model Hadoop-based for high-risk students was constructed. Finally, the prediction model was evaluated and the rationality of the model was proved. The model can analyze the collected data of students and predicts the high-risk students in big data environment.**

Keywords—prediction model; high-risk students; big data; Hadoop platform

I. INTRODUCTION

In order to satisfy the need of processing students data and improve the efficiency of high-risk students prediction in big data environment [1], combined prediction model based on the weighted voting (WCVP) is given, the Hadoop platform [2] is used to divide students data into several data subsets, and the weighted voting combined prediction method is used. The optimal prediction results of each subset are obtained by parallel computation of each data subset, and then the prediction results are summarized. The weighted voting combined prediction method is used to process the final prediction results. This method is similar to the weighted voting election. First, the excellent representatives of each region are selected, and then these excellent representatives are aggregated. If the aggregated data is still huge, the iterative method can be used to conduct multiple zonal weighted voting elections, gradually reducing the amount of data, and finally, the global weighted voting elections are carried out, and the final result is obtained. Excellent representative. This method meets the requirements of big data, reduces the cost of computing overhead, and effectively improves the computing efficiency.

II. DATA PARTITIONING,MAP() AND REDUCE()

The parallelization of weighted voting combined prediction method is to divide the data set into N data blocks [3]. The data blocks of each processing node train the corresponding weighted voting combined prediction model, and the training results of each part are summarized as the training set of the aggregate prediction model to train the model [4]. Practice and get the final prediction results. The precondition of weighted voting combination prediction method based on large data parallel is that there is no interdependence between data, and student data can meet this requirement.

MapReduce [5] is divided into Map and Reduce phases when performing distributed computing tasks [6]. The input and output of each phase are keys/values. MapReduce framework can be used to solve different problems by defining map() function and reduce() function[7]. The data set (or task) processed by MapReduce should have the following two characteristics: (1) the data set to be processed can be divided into several small data sets; (2) each small data set can be processed in parallel[8]. Aiming at the characteristics of weighted voting combined prediction method and the parallel computing framework of MapReduce, this paper designs a parallel algorithm of weighted voting combined prediction based on MapReduce, which is as follows.

A. Data Partitioning

The student data set can be partitioned only after being cleaned because a clean student data set is the precondition for the data set to be partitioned. Otherwise, singularity and noise data will weaken the generalization ability of the prediction model for high-risk students. When using MapReduce framework to process, it does not need to divide the student data manually. Hadoop platform can automatically divide the student data set into blocks, and distribute the data blocks to the corresponding processing nodes for calculation [9]. The number of partitions N equals the number of maps, and the size of N is determined by the following three factors. (1) Block size: The size of data block in HDFS, which is 64M by default.(2) total size: the total size of the data set.(3) inputfile_num: the number of files entered.

B. Map() and Reduce()

Student data sets are partitioned and stored distributed after HDFS. Map() function is used to process data blocks in parallel [10]. Each data block is required to be processed by a map() function, and each data block is an input to the map() function. In the map() function, the training process of the voting combination prediction model is defined, and the prediction results are sent to the Reducer as its input as the output of the map() function. Reduce() function analysis and aggregation of

all map() function output, and then weighted voting combination prediction to get the total predict results, if the total predict results data is too large or not meet the requirements, then continue to iterate, otherwise the total predict results will be output as the final predict results.

III. STRUCTURE OF COMBINATION PREDICTION MODEL

The model structure is shown in Figure 1.

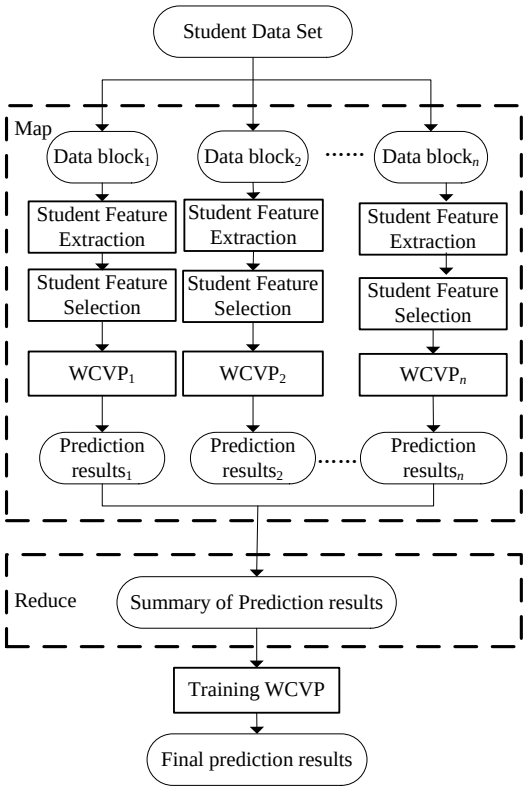

FIGURE I. STRUCTURAL CHART OF COMBINATORIAL PREDICTION MODEL

In Figure I, the training set of student data that needs to be processed is stored in HDFS on Hadoop platform. Then, the training set is divided into blocks to form several data blocks. Then, the weighted voting combination prediction method (WCVP) in each Map is trained in parallel. Finally, the weighted voting combination prediction model after training is obtained. The prediction model after training is used to process the predicted student data set in the same block. That is to say, each data block to be predicted is stored in each node of Hadoop platform, and then processed by parallel computing. Then, the predicted results of each node are processed by unified statistics and analysis. Finally, the whole data set to be predicted can be obtained. The final prediction result of the predicted student data set.

IV. WORKFLOW OF COMBINED PREDICTION MODEL

The workflow of the model is divided into two parts: (1) model training; (2) prediction of unknown student data. As shown in Figure II

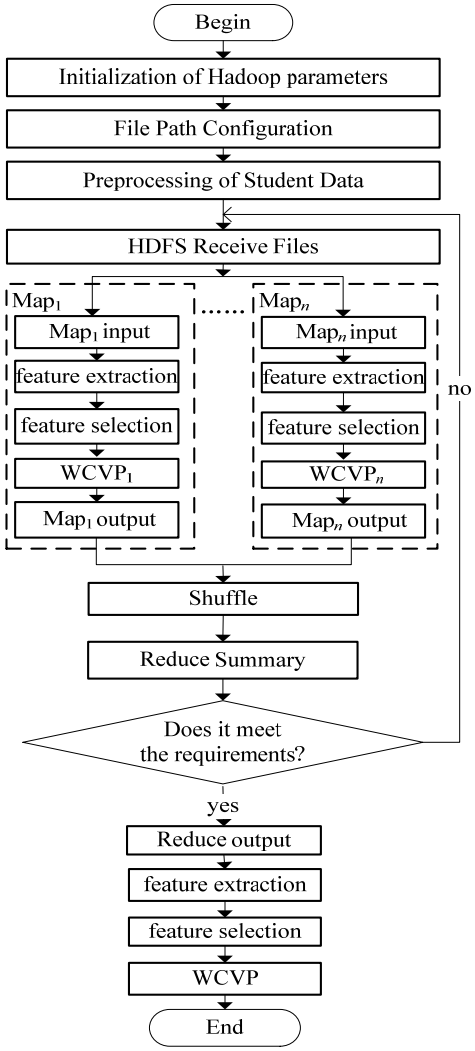

FIGURE II. WORKFLOW OF COMBINED PREDICTION MODEL

A. Model Training

There are four steps in the process of model training.

- Read the student data training set, preprocess, eliminate noise data and singularities, and block the data into HDFS.
- In each Map, students' characteristics are extracted according to the corresponding data blocks.
- Choosing the corresponding students' characteristics.
- WCVP composed of classification decision tree, logical regression and naive Bayesian is trained in corresponding Map.

B. Predicting Unknown Student Data

- Initialization of Hadoop parameters. Each MapReduce task is initialized to the corresponding job. At this stage, the job is initialized, and related operations such as naming and setting the input and output types, as well as reading the Hadoop file configuration are carried out.
- Configuration of input and output paths. In the configuration process, the input and output paths need to be manually configured. The input and output paths of the prediction model are all HDFS paths. The data of the input prediction model and the results of model processing are stored on HDFS.
- Data preprocessing for unknown students. The data sets of students are cleaned and denoised, and the features of students are extracted and selected.
- HDFS storage of data. The data of students to be predicted after processing is uploaded to HDFS for distributed processing by MapReduce.
- Map stage. According to the data blocks on HDFS, feature extraction and feature selection are carried out in corresponding maps, and high-risk students are predicted by trained WCVP, and the prediction results of each prediction model in each map are obtained.
- Shuffle stage. Shuffle process pulls map task-side data to reduce end and ensures that unnecessary bandwidth consumption is minimized.
- Reduction stage. In reduce stage, the prediction results of map stage are aggregated to determine whether the aggregated data sets are too large, and if they are too large, they are iterated. Otherwise, WCVP is used to generate the final reasonable prediction results.

V. EVALUATION AND COMPARISON OF PREDICTION RESULTS

A. Experimental Design

The experimental computer is configured with i7 7700K CPU and DDR4 16G memory. The data set is 316 students' data since enrollment. 60% of the data are randomly selected to form the training set and the remaining 40% to form the test set. The main data sources used in the experiment are as follows:

- Students' basic archives and students' basic archives data are managed by the student management department.
- Campus card consumption data.
- Data on access control in student dormitories.
- Access control and borrowing data of libraries.
- Achievement data. The data of students' academic achievements are managed by the Academic Affairs Department, which records the students' normal grades, final grades, comprehensive scores and credit scores in each semester.
- Web log. Students access the campus network by means of student identification number. The format and identifier of students' visiting web pages are recorded in the log file. This information are mainly encrypted student identification number, requested URL, time, source IP address, target IP address, etc. Web logs contain a lot of noise data, which need to be filtered and converted effectively.

B. Experimental Result

Table1 lists whether students can pass the course examination in Parallel_CART, Parallel_Logistic, Parallel_NB and WCVP on the student data set, and evaluates the sensitivity, specificity and accuracy of the model by the retention method.

TABLE I. COMPARISON OF MODEL PREDICTIONS

Prediction method	Precision	Sensitivity	Specificity
Parallel_CART	97.5%	96.1%	97.8%
Parallel_Logistic	97.9%	96.1%	98.2%
Parallel_NB	98.2%	98.0%	98.2%
WCVP	99.2%	98.1%	99.5%

According to Table1, the results of each prediction model are as follows:

- The accuracy of Parallel_CART is 97.5%, which means that the prediction model can accurately predict students'performance with a sensitivity of 96.1%. It means that most students who fail in the course will be correctly predicted with a specificity of 97.8%. It means that students who pass the course have a high probability of correctly predicting their performance.
- The accuracy of Parallel_Logistic is 97.9%, which is slightly higher than that of Parallel_CART, indicating that the prediction model can predict students'performance more accurately; the sensitivity is 96.1%, which is consistent with that of Parallel_CART, indicating that the effect of predicting failed students is similar to that of Parallel_CART; the specificity is 98.2%, which is higher than that of Parallel_CART. The specificity is slightly higher, indicating that the probability of correct performance of students who pass the Parallel_Logistic prediction course is slightly higher than that of Parallel_CART.
- The accuracy of Parallel_NB is 98.2%, and Parallel_Logistic is slightly higher, indicating that the prediction model can predict students'performance more accurately; the sensitivity is 98.0%, which is slightly higher than that of Parallel_Logistic, indicating that the probability of correctness of Parallel_NB in predicting students' performance failing courses is slightly higher than that of Parallel_Logistic. 98.2%, which is consistent with the specificity of Parallel_Logistic, indicates that the effect of predicting passing students is similar to that of Parallel_Logistic.
- The accuracy, sensitivity and specificity of the three single prediction models and WCVP are all above 96%, and the fluctuation is not large, but the accuracy,

sensitivity and specificity of WCVP are the highest. Therefore, selecting excellent features and adopting different prediction methods can achieve better prediction results. WCVP is used to optimize the performance of the prediction model, which can effectively improve the effect of the prediction model and achieve the desired prediction purpose.

ACKNOWLEDGMENT

This research was financially supported by Hubei University of Economics international courses construction special project (No. [2015]3).

REFERENCES

[1] Xiaogao YU. Prediction of High Risk Students Based on Big Data. xiamen university press, 12 May 2019.

[2] M Alkasassbeh, “An empirical evaluation for the intrusion detection features based on machine learning and feature selection methods”, Journal of Theoretical & Applied Information Technology, vol. 22, pp. 5962-5976, November 2017.

[3] X Cheng,H Cai,Y Zhang,et al, “Optimal combination of feature selection and classification via local hyperplane based learning strategy”, Bmc Bioinformatics, vol. 16, pp. 1-10, January 2015.

[4] RJ Palma-Mendoza,D Rodriguez,L De-Marcos, “Distributed ReliefF-based feature selection in Spark”, Knowledge & Information Systems, vol. 19, pp. 1-20, December 2018.

[5] Sun Z, Song Q, Zhu X, et al, “A novel ensemble method for classifying imbalanced data”, Pattern Recognition, vol. 48, pp. 1623-1637, May 2015.

[6] X Liu,L Wang,J Zhang,et al, “Global and Local Structure Preservation for Feature Selection”, IEEE Transactions on Neural Networks & Learning Systems, vol. 25, pp. 1083-1095, June 2017.

[7] R Behera,K Das, “A Survey on Machine Learning: Concept,Algorithms and Applications”, International Journal of Innovative Research in Computer & Communication Engineering, vol. 2, pp. 1301-1309, February 2017.

[8] NX Vinh,S Zhou,J Chan,J Bailey l, “Can high-order dependencies improve mutual information based feature selection? ”, Pattern Recognition, vol. 5, pp. 46-58, March 2016.

[9] S Egea,A Rego,B Carro,et al, “Intelligent IoT Traffic Classification Using Novel Search Strategy for Fast Based-Correlation Feature Selection in Industrial Environments”, IEEE Internet of Things Journal, vol. 3, pp. 1616-1624, May 2018.

[10] L Wang,C Wu, “A Combination of Models for Financial Crisis Prediction: Integrating Probabilistic Neural Network with Back-Propagation based on Adaptive Boosting”, International Journal of Computational Intelligence Systems, vol. 10, pp. 507-520, October 2017.