Application of English Score Management System Based on Spark-Decision Tree Algorithm

Yisha Zhang
Jiangxi University of Engineering, Jiangxi Xinyu, 338000, China
313129396@jxue.edu.cn

Abstract. This paper is based on the classification decision tree algorithm, based on the Spark cloud platform, through the construction of the decision tree algorithm, and setting AUC parameters, comprehensive comparison of the support vector machine algorithm, and Logistic regression. The results show that the Spark cloud platform has the best robustness. Better realize data forecasting.

Keywords: Decision tree algorithm; prediction; AUC; Spark cloud platform.

1. Introduction
The annual English four-level exam is not only concerned about college students, but also a major event that all institutions are concerned. How to improve the students' four-level pass rates into an important part of teaching management of all institutions. This pass rate has a certain relationship with the students' usual English scores, and can you predict the four-level passing situation of each student from usual scores. Through data mining technology, you can explore implicit, unknown, user-interested in data, and provide potential value for decision-making in a large amount of data of the student score management system. Data mining should pass through a series such as data acquisition, pretreatment, data analysis, and results representation. The algorithm used is mainly: related rules, decision tree methods, artificial neural network, genetic algorithm, rough set method, fuzzy method, Bayesian model, etc. [1] This article will adopt a classification decision tree algorithm to analyze the average score of higher vocational students, to find a valid classification rule forecast student through the English four-level exam.

2. Decision tree algorithm
The decision tree can be based on a set of data records without any order rules, and a classification rule representing the form of decision tree is introduced. The decision tree method adopts the self-turning recursive mode, and the decision tree is compared to the point of property value, and the downward branch is determined according to the different attribute values, and the leaf junction of the decision tree is conclusively. Each node of the decision tree represents an attribute, and each branch represents the judgment condition of this attribute.[3]

Decision trees and decision tree rules belong to the classification method of output, mainly for solving classification and prediction issues. The most critical place in this algorithm is the construction of decision trees. An effective decision tree is established to include both steps and twigs. The first step
is to use the training set to establish a decision tree to establish a decision tree model rule; the second step is to classify the input data using the generated decision tree. For the input record, the recorded property value is tested from the root node until a certain leaf node is reached, thereby finding the class in which the record is located. The decision tree must be trimmed during the decision tree. The purpose of the twig is to reduce the undulations generated due to noise in the training set. There are two of our commonly used trimming techniques: one is a pre-repair, the other is the post-trim. The conditions of the decision tree stop segmentation have: data on one node belongs to the same category; no attributes can be used to segment the data. [2]

Decision tree algorithm mainly solves three optimization problems: the minimum leaf is generated; the depth of each leaf generated is minimal; the generated decision tree leaves are minimal and the depth of each leaf is minimal. [3]

This article combines the characteristics of aviation data and the development of big data technology, uses Spark cluster as a memory computing framework, uses decision trees to build models, and conducts predictive analysis on English four and six levels. The experimental results can provide data support for efficient organizations.

3. Spark platform

3.1. Introduction to Spark

Spark is a big data parallel computing framework based on memory computing, which was developed by the University of California, Berkeley in 2010. At present, a complete biosphere has been formed, which takes SparkCore as the core and can use different components, such as: Spark ecosystem Streaming, SparkMLlib, GraphX and so on. It mainly tries to realize the platform of big data application through large-scale integration among algorithms, machines and people.

![Figure 1 Spark ecosystem](image)

3.2. Description of the environment

In the environment installation, we first configure three virtual machines to install the Centos7Linux system, configure the SSH keyless login and system time synchronization between the virtual machines, and install the Java environment and the Hadoop environment. After the Hadoop cluster is successfully configured, install the Spark cluster in SparkonYarn mode. [4,5]
Table 1. Description of cluster nodes

| Host name    | IP address       | Effect                      |
|--------------|------------------|-----------------------------|
| Chenor01     | 192.168.57.255   | NameNode, Master            |
| Chenor02     | 192.168.57.253   | DataNode, Worker            |
| Chenor03     | 192.168.57.254   | DataNode, Worker            |

4. Data introduction

4.1. Data source

This project research mainly analyzes the statistical data of a certain high-efficiency English Band 4 and 6 grades. The data set size is 556M. There are 8 columns of data, including college, gender, college entrance examination results, class category, first grade of English at school, and second grade of English at school. The second grade, the third grade of English at school, the fourth grade of English at school, and other relevant information. [6]

4.2. Data cleaning

English scores may be affected in various ways during the collection process, such as data recording errors, data recording facts, etc., so the data may be inaccurate, such as null values in the data, and special characters in some data. In order to avoid affecting the results of later data analysis, the data needs to be cleaned. [7]

After processing the dataset, the final dataset categories are as follows:

Table 2. Data format table after processing

| Column name                          | Data type | Data description |
|--------------------------------------|-----------|------------------|
| College                              | Int       | 0-26             |
| gender                               | Int       | 0-1              |
| SAT score                            | Int       | 0-750            |
| Class category                       | Int       | 25               |
| The first grade of English at school | Int       | 0-100            |
| The second grade of English at school| Int       | 0-100            |
| The third grade of English at school | Int       | 0-100            |
| The fourth grade of English at school| Int       | 0-100            |

5. Implementation of Prediction algorithm based on Spark

5.1. CART classification tree

CART is the conditional probability distribution of the output random variable Y under the condition of a given input random variable X. the CART algorithm generates a binary tree, and there are only two categories: "yes, no". [8]

The CART classification tree selects the optimal features by GINI index. The GINI coefficient indicates the impurity of the model. The smaller the GINI coefficient is, the lower the impurity is and the better the characteristic is. Therefore, the classification condition can be obtained by selecting a smaller GINI coefficient, which can be assigned as a syncopation point in turn. In the classification problem, suppose that there are k categories, and the probability of the k category is, then the expression of the GINI coefficient is:

\[ Gini(p) = \sum_{k=1}^{k} p_k (1 - p_k) = 1 - \sum_{k=1}^{k} p_k^2 \]  

(1)

For a given sample D, suppose there are K categories, and the number of the Kth category is, then the expression of the GINI coefficient of sample D is
\[ Gini(D) = 1 - \sum_{k=1}^{K} \left( \frac{|C_k|}{|D|} \right)^2 \]  
\[ (2) \]

For sample D, if D is divided into D1 and D2 according to a value an of feature A, under the condition of feature A, the expression of GINI coefficient of D is as follows:

\[ Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \]  
\[ (3) \]

The greater the GINI coefficient, the greater the uncertainty of the sample set.

5.2. Evaluation indicators

In the classification model, the quality of the classifier is mainly based on confusion matrix, ROC and AUC.

Confusion matrix is used to express the accuracy evaluation, and is mainly used to distinguish whether the classification is good or bad. Take two categories as an example, 0 and 1 represent two types of samples, in which TP: is true positive, prediction is 1, in fact it is also 1 position FP: false positive, prediction is 1, fact is also 0 position TN: true negative, prediction is 0, reality is also 0: false negative, prediction is 0, fact is also 1. [9]

| Actual \ forecast | 1   | 0   |
|-------------------|-----|-----|
| 1                 | TP  | FN  |
| 0                 | FP  | TN  |

Table 3. Confusion matrix

Accuracy: the proportion of correctly classified samples to the total number of samples, indicating the accuracy of the evaluation.

\[ \text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \]  
\[ (4) \]

Accuracy: indicates the proportion of positive samples that are actually positive examples.

\[ \text{Exact ratio} = \frac{TP}{TP + FP} \]  
\[ (5) \]

Sensitivity: it is actually a positive sample and is divided into the proportion of positive samples in the positive samples, indicating the proportion of all positive samples to be paired, in order to measure the ability of the classifier to recognize positive examples.

\[ \text{Recognition ability} = \frac{TP}{TP + FN} \]  
\[ (6) \]

ROC is a curve connected by the points of FPR and TPR. The horizontal axis is FPR, and the vertical axis is TPR.

\[ TPR = \frac{TP}{TP + FN} \]  
\[ (7) \]

\[ FPR = \frac{FP}{FP + TN} \]  
\[ (8) \]

The curve area of AUC: ROC is the AUC value, and AUC is mainly used to measure the performance or generalization ability of machine learning algorithms in binary classification problems. The ROC curve is closer to the upper left corner, and the larger the value of AUC is, the better the effect is.
6. Experimental results and analysis
Cross-validation is used in the experiment to randomly divide the test set and the training set, in which the training set accounts for 70% of the data and the test set accounts for 30%. The decision tree algorithm is realized by Spark programming, and the performance of the algorithm is analyzed. At the same time, compared with the other two algorithms, the results are shown in the following table: [10]

| Method               | Accuracy rate | Precision rate | Sensitivity | AUC     |
|----------------------|---------------|----------------|-------------|---------|
| Decision tree        | 97.83%        | 98%            | 96.92%      | 97.72%  |
| Support vector machine | 92.01%    | 87.22%         | 95.46%      | 92.43%  |
| Logistic regression  | 88.56%        | 82.63%         | 82.61%      | 89.11%  |

From the results in the above table, we can see that the accuracy and sensitivity of using the decision tree algorithm to predict the classification of English through data are relatively high. The high classification accuracy rate indicates that the model has passed accurate judgments on levels four and six, and the high sensitivity indicates that the model has a low false detection rate. Therefore, it can be concluded that the decision tree algorithm is more effective in the process of performance prediction and judgment.

7. Conclusion
Through the classification model formed by the above rules, the students' English level 4 exam can predict that when the above IF conditions can be met, the prediction conclusions of the THEN can be obtained. With this prediction module, policymakers can strengthen English teaching work in higher vocational orders. As a high-time recruitment, since the students who are not like high schools have experienced high school English, their English ability is weak, and should pay more attention to English learning.

References
[1] Han Jiawei. Data mining concept and technology [M]. Beijing: Machinery Industry Press, 2008.22, 66.
[2] Jiawei Han. Concept and technology of data mining [M]. Beijing: Machinery Industry Press, 2001
[3] Wu Xia. Research on the Application of Data Mining Technology in Teaching Management [D]. Harbin Engineering University, 2007:10.
[4] Xu Haiwen, Han Songchen, Liu Xiaodong. Influencing factors of flight delays from the perspective of passengersAnalysis[J]. China Market, 2015(52):221-224.
[5] Gao Yanjie. Spark big data processing technology. Application and performance optimization [M]. Beijing: MachineMechanical Industry Press, 2015.
[6] Liang Yan. Combination of data mining algorithms based on distributed platforms Spark and YARNXinghua Research[D]. Guangzhou: Sun Yat-sen University, 2014.
[7] He Yang, Zhu Jinfu, Zhou Qinyan. Airport flight delay based on support vector machine regressionForecast[J]. Journal of Civil Aviation University of China, 2018, 36(1):30-36,41.
[8] Luo Yunqian, Chen Zhijie, Tang Jinhui, etc. Flight delays using support vector machine regressionForecast Research[J]. Transportation System Engineering and Information, 2015, 15(1): 143-149,172.
[9] Liu Xiaoping. The cause mechanism and anticipation of passenger group emergencies in the context of flight delays. Research on Police Mechanism[D]. Wuhan: Wuhan University of Technology, 2013.
[10] Zhuang Rong, Li Lingjuan. Research on Parallelization of CVFDT Classification Algorithm Based on Spark[J]. Computer Technology and Development, 2018, 28(6): 35-38.