The research on data mining technology in the analysis of students scores

Jiahua Wan[1,2], and *Lei* Liu [1,*]

[1]School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, Anhui, China
[2]College of Computing and Information Technologies, National University ,Manila, Philippines

Abstract. On the basis of being familiar with the traditional decision tree algorithm, some improvements are made to the C4.5 algorithm to reduce the mining time. After preliminary processing of students' scores, the improved algorithm is used to mine the management association rules among students' scores, and the results of association rules are analyzed and interpreted.

1 Introduction

Today's computer technology and information technology are developing very rapidly, so we must use the computer and information technology in line with the current era to strengthen the work of student performance analysis[1]. This is not only an inevitable means of the times, but also can simplify management and facilitate daily work. This paper applies data mining technology to student achievement analysis, and fully develops the role of data mining in student achievement analysis[2]. Thoroughly improve students' performance information management and analysis ability, mainly for the following groups: college students, educators or managers of higher vocational colleges, etc. The application of data mining technology will certainly improve the efficiency of student performance analysis, and will also bring great help to the improvement of teaching quality, and build a very effective bridge between college graduates and the job market[3].

2 Data description

The research object of this paper is the data of undergraduates majoring in finance and management. Starting from the collection of a series of relevant data of undergraduates majoring in finance and management from 2019 to 2021, this paper combs the relevant data of undergraduates majoring in finance and management, so as to lay a good foundation for data mining, Due to the large number of academic performance analysis data of Finance and management undergraduates, it is not suitable to select a large number of research, so I only selected some data as the research goal, including 553 students majoring in economics and finance and financial management. Among them, the largest number of students are financial management, reaching 291, 174 students majoring in economics and finance, and

[*] Corresponding author: leiliu89@mail.ustc.cn

88 students majoring in information management and information system. After the corresponding integration processing of the data, the next step of the operation of the data is to analyze and process the correlation of the data. By analyzing the correlation of data, we can effectively distinguish the corresponding attribute set of data, eliminate the basically useless or invalid attributes, and pave the way for data mining. Through correlation analysis, we can find many noises that affect the accuracy of mining. If these noises are reasonably filtered out, the mining effect of data mining will be further increased. Therefore, before mining, it is necessary to eliminate the items that have nothing to do with decision support, such as student number, student cadres and so on. Secondly, some special screening can be carried out according to the specific actual situation. For example, the financial management undergraduates in the target universities were originally planned students who must obtain the computer level before they can graduate[4]. However, through the observation and analysis in recent years, although the general financial management graduates have improved their ability to operate the computer, However, it has not reached or rarely reached the level of obtaining the technical certificate, which makes the students quite confused and makes the school feel powerless. Therefore, after repeated weighing, it is necessary to remove the condition of obtaining the computer level before graduation.

2.1 Data analysis

The comprehensive score (student average score) is divided into four discrete values: excellent score (score ⩾ 85), good score (score between 70 and 84), medium score (score between 60 and 69) and failed score. In the regulations of the target universities, all undergraduate graduates of financial management must achieve a medium score before they can be considered as qualified, and those below this score will not graduate. Therefore, in this case, the segments below the average score will not be treated as redundant items through data preprocessing.

2.2 Data processing

Table 1. Summary of student achievement information.

No	**English level**	**Comprehensive achievements**	**practical ability**	**Employment quality**
1	CElb	good	*low*	high
2	CE14	good	medium	medium
3	CElb	excellent	low	high
……	……	……	……	……
150	CElb	good	high	high
151	CE14	medium	low	low
……	**……**	……	……	……

After data conversion, the data should be cleaned. Redundant data shall be cleaned, useless or duplicate data shall be eliminated, and the overall structure of data shall be simplified. In the performance analysis, since students who fail to meet the medium standard will not graduate, these students do not belong to the investigation scope of performance analysis, so the data of these students must be cleaned from the employment analysis data. In the process of cleaning, the quality evaluation of students will be carried out at the same time. For example, some students are not employed but go to higher education, Such students are also rated as high quality. After cleaning and analyzing all sample data, some sample data are selected for data preprocessing for future formal data mining and decision analysis (see Table 1).

3 Algorithm and data mining

3.1 Decision tree algorithm

The current mainstream decision tree algorithms can basically be summarized into two categories: one is the original decision tree algorithm[5], ID3 algorithm[6], and the other is the evolutionary advantage algorithm C4.5 algorithm[7]. The principles of the two algorithms are basically the same. Both of them use the special index of information gain as the attribute selection metric ID3 algorithm.There is uncertainty in the processing of large gain value information. After each intermediate node, the gain value information can be found and processed through uncertainty. Then the rule classification is measured according to the degree of information increase.

ID3 decision tree algorithm is processed by analyzing information gain value. The uncertainty can be eliminated by obtaining different qualitativeness and then determining whether it is a choice content according to the content. The main advantages of ID3 algorithm are: simple operation, branch and branch without too much restraint, very easy to understand, the amount of calculation is relatively small, the use is also very simple and fast. However, ID3 algorithm has weak ability to deal with continuous data and cannot deal with complex sets well.

For this defect, for C4. 5 metric attributes of the algorithm are modified to solve the disadvantages of ID3 algorithm, so that the shortcomings mentioned in ID3 algorithm can be made up.

The C4.5 algorithm assumes that the information set to be mined is A, which has N discrete attributes, and then downloads the number of included data sets S1, S2... Sn, the information gain calculation is the same as the data sample classification method calculated by ID3 algorithm. Formula (1) of the following formula is used for conversion through a series of information. We can get the calculation formula (2).

$$\mathrm{Split}(S,A) = \sum \frac{s_i}{s}\log_2\frac{s_i}{s} \quad (1) \quad \mathrm{Cainradio}(A) = \frac{\mathrm{Gain}(A)}{\mathrm{Split}(S,A)} \quad (2)$$

In order to simplify the amount of computation to a certain extent, entropy calculation needs to be changed to improve the efficiency of entropy calculation, as shown in Formula (3).

$$\mathrm{Info}'(X) = -\sum_{i=1}^{m} \frac{|A_i|}{|X|}\mathrm{Info}(A_i) \quad (3)$$

In the balanced environment of attribute A, formulas (4) and (5) can be obtained after optimized entropy calculation:

$$\mathrm{Gain}'(A) = |\mathrm{Info}(x) - \mathrm{Info}'(A)| \quad (4) \quad \mathrm{Gainradio}'(A) = \frac{\mathrm{Gain}'(A)}{\mathrm{Split}'(A)} \quad (5)$$

Although the core of the improved C4.5 algorithm is still ID3 algorithm, on the basis of the information gain rate selection attribute, the processing ability of discrete continuous attribute is completed. The C4.5 algorithm is described below:

Step 1: Determine the class (class tag) attributes.

Step 2: Classify the data to be mined by rules.

Step 3: By further dividing the data after rule classification, nodes are divided by category, and N is divided into leaf nodes.

Step 4: then the data set of candidate attributes can not be considered, then N also becomes a leaf node, and the leaf node tag (class tag) can be obtained according to the majority principle.

Step 5: For each attribute to be labeled, judge, can not distinguish will be separated.

Step 6: Set the limit value of Gainradio in the formula.

After get the rules of classification properties, its specific node values can be handled as a branch of the decision tree, data collection of related branches of data collection and data acquisition) branch node, branches of classification data set of attributes, when the data set for real number, branch is established, if not, then to the above steps.

C4.5 algorithm is very good at mining series of information and can be widely used by multi-level rule classification. C4.5 algorithm can be used to analyze attributes well and collect values of continuous attributes and discrete attributes within the range.

Taking the above basic attributes as an example, Naive Scale algorithm performs discrete calculations and arranges values a_1, a_2,... ,a_n,Then calculate the average value of adjacent values a_i and a_{i+1} in order to get the average value a, then the average value a is a node, this method can get n-1 sub-nodes, then the sample set S will be separated by n-1 sub-nodes to form a corresponding number of different subsets. Where, $A \leq a$ and $A > a$ are the corresponding discrete values of the continuous attribute a, through which the segmentation calculation can be carried out to obtain the information gain. The calculated information gain shows that the node A 'is the maximum, and a' is used to divide the sample set S into two new partial subsets. Is A ≤ a 'and A > a', then the increase degree of information on node a 'can be obtained, through which the final judgment and analysis can be made, but it cannot exceed the limit value within the range.

3.2 Data mining

Among the algorithms of data mining, the relatively simple decision tree algorithm is chosen, because users are not professionals in data mining, the cognition of decision tree and the way of system expression are very simple and easy to understand, and easy to be accepted by ordinary users. The calculation method is simpler and more efficient than other algorithms.

The specific modeling process of C4.5 algorithm is as follows:

First, according to the ID3 algorithm, the expected value of the classification is calculated according to the attributes of the training sample set by using the formula $I(S_i) = P(S_i) * \log_2 \frac{1}{p(S_i)}$. There are 410 samples of sample data, and employment quality is taken as the attribute of classification setting.

$$I(128,165,117) = \frac{128}{410}\log_2\frac{128}{410} - \frac{165}{410}\log_2\frac{165}{410} - \frac{117}{410}\log_2\frac{117}{410} = 1.5691$$

According to each attribute, it can be classified as "advanced", "intermediate" and "low". Through the specific analysis of these three items, make discrimination, data classification rules and attribute determination. Get the information of each subset of attribute profession.

$$I(24,47,43) = -\frac{24}{114}\log_2\frac{24}{114} - \frac{47}{114}\log_2\frac{47}{114} - \frac{43}{114}\log_2\frac{43}{114} = 1.5308$$

$$I(67,77,44) = -\frac{67}{188}\log_2\frac{67}{188} - \frac{77}{188}\log_2\frac{77}{188} - \frac{44}{188}\log_2\frac{44}{188} = 1.5483$$

$$I(37,41,30) = -\frac{37}{108}\log_2\frac{37}{108} - \frac{41}{108}\log_2\frac{41}{108} - \frac{30}{108}\log_2\frac{30}{108} = 1.5733$$

The information decision subset of each attribute value can be obtained through the above calculation, and the decision tree calculation formula $I(s_1,s_2,\dots,s_m) = -\sum_{i=1}^{m} P_i \log P_i$,can calculate the corresponding information entropy value, and then classify the whole data set according to the data attributes, and finally select the mining target of data mining.

$$E(\text{English level}) = \frac{114}{410} I(24{,}47{,}43) + \frac{188}{410} I(67{,}77{,}44) + \frac{108}{410} I(37{,}41{,}30) = 1.5500$$

$$E(\text{Comprehensive achievements}) = \frac{78}{410} I(29{,}19{,}30) + \frac{187}{410} I(51{,}88{,}48) + \frac{145}{410} I(48{,}58{,}39) = 1.5465$$

$$E(\text{Practical ability}) = \frac{124}{410} I(64{,}32{,}28) + \frac{164}{410} I(44{,}93{,}27) + \frac{122}{410} I(20{,}40{,}62) = 1.4407$$

Use the formula (1) for data mining calculation, and use the corresponding formula can be used for data mining decision calculation.

$$\text{Split}(\text{Practical ability}) = -\frac{122}{410}\log_2\frac{122}{410} - \frac{164}{410}\log_2\frac{164}{410} - \frac{124}{410}\log_2\frac{124}{410} = 1.5709$$

$$\text{Split}(\text{Comprehensive achievements}) = -\frac{78}{410}\log_2\frac{78}{410} - \frac{187}{410}\log_2\frac{187}{410} - \frac{145}{410}\log_2\frac{145}{410} = 1.52024$$

$$\text{Split}(\text{English level}) = -\frac{114}{410}\log_2\frac{114}{410} - \frac{188}{410}\log_2\frac{188}{410} - \frac{108}{410}\log_2\frac{108}{410} = 1.5362$$

At this time, formula 2 can be used for data mining decision calculation.

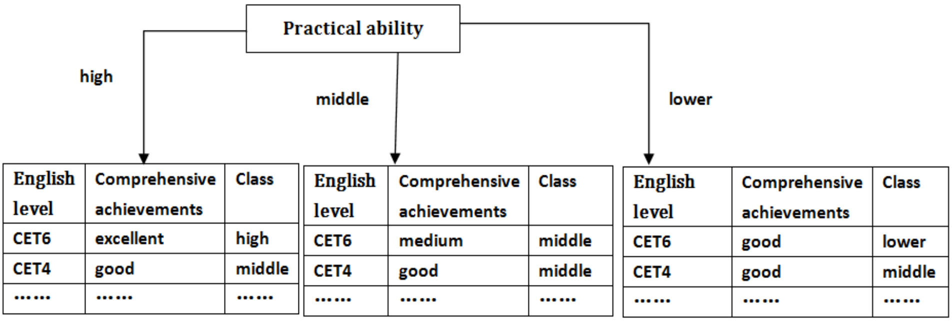
Gairatio(English level)=0.0126, Gainratio(Comprehensive achievements)=0.0150, Gainratio(Practical ability)=0.0821

If the improved C4.5 algorithm is used, formula 5 will be used to calculate the decision information of data mining more efficiently.

Gainratio(English level)=0.0124, Gainratio(Comprehensive achievements)=0.0150, Gainratio(Practical ability)=0.0817

It can be seen that the improved decision attribute values are basically consistent with the normal algorithm.

By comparing decision attributes in the system, the information gain rate is calculated. Here, "practical ability" is taken as the root node, and the attribute with the maximum information gain rate is calculated according to the algorithm. The attribute of practical ability shows that there are important branch decision items, which are divided into three items, each of which is a collection of data. Practical ability decision tree is shown in Fig.1.



English level	Comprehensive achievements	Class
CET6	excellent	high
CET4	good	middle
……	……	……

English level	Comprehensive achievements	Class
CET6	medium	middle
CET4	good	middle
……	……	……

English level	Comprehensive achievements	Class
CET6	good	lower
CET4	good	middle
……	……	……

Fig. 1. Decision tree with practical ability as root node.

On this basis, we further calculate the data decision analysis, calculate the branch set through the decision tree algorithm, and the attribute value of the branch is. Therefore, through this standard value, we can determine the next primary factor, which is the comprehensive score. Therefore, we can choose "comprehensive score" The attribute continues to be the medium subset for the mining algorithm. Among the subsets with low practical ability, the "English level" attribute set will be used for the mining algorithm. The above steps of cyclic operation are pruned to obtain the practical ability decision tree, as shown in Fig. 2.

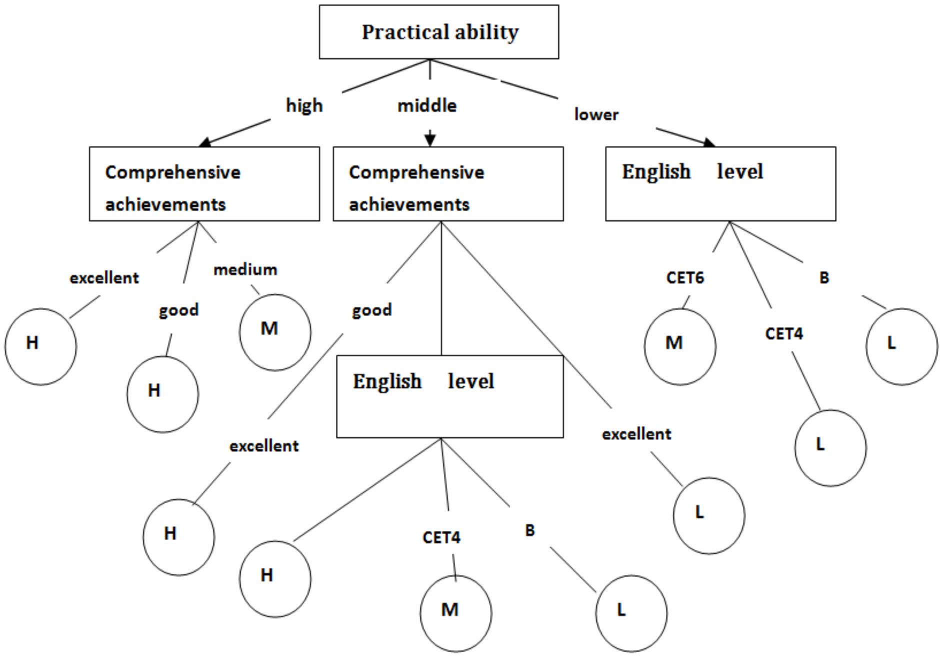


Fig. 2. Employment quality decision tree.

4 Conclusion

In this paper, data mining decision tree algorithm is used to analyze the results of finance and management graduates to help them provide targeted employment guidance. With classified data mining technology to achieve the financial and management undergraduate employment guidance work of the entire data mining process, determine the target of data mining, develop the target data pretreatment scheme, and finally form based on C4.5 decision tree model for the adaptability between financial and management graduates' academic performance and employment quality is established.

This paper is supported by the major teaching reform research project of Anhui Province (No. 2020jyxm0789) and the University blockchain technology Innovation and Application Plan project of Anhui Province (No. 2020qkl21). At the same time, the project is also supported by the Industry-University-Research Horizontal Project of Anhui Xinhua University (No. 2019CXY014) and the Key Supported Discipline Project of Anhui Xinhua University (No. Fcxk202102). In addition,the authors would like to give many thanks to the anonymous reviewers for their contributions to improve this manuscript.

Reference

1. Li Qiaojun, Li Wei. Research on the application of data mining technology in student achievement analysis .J. Microcomputer application. **31**,2(2015)
2. Li Jingyan. How to use data mining technology to build college student achievement analysis system. J. Information and computer (theoretical Edition). **22**,4(2015)
3. Sun Yonghui, Zhou Hong. Research on the application of data mining technology in college performance analysis .J. Science and technology innovation guide. **12** ,3(2015)
4. Liu Jinyi. Application of data mining technology in college students' performance analysis .J. Information recording materials. **7**,3(2021)
5. Xue Yanan, Yang Xiaodong.Application of decision tree algorithm in student achievement.J.SCIENCE & TECHNOLOGY INFORMATION. **36**.3(2019)
6. Yu Shuyun. Research on online teaching learning effect prediction model based on ID3 algorithm. J. journal of lanzhou university of arts and sciences (natural science edition) **35**,5.(2021)
7. Li Chunsheng, Jiao Haitao, Liu Peng & Liu Xiaogang. Improvement and application of decision tree classification algorithm based on C4.5. J. Computer technology and development.**5**,4,(2020)