Eng Int Syst (2025) 3: 345–350

Engineering
Intelligent Systems

The analysis of student achievement using data mining in practical teaching informatization

Ying Zhong[1] **and Juncheng Mo**[2,*]

[1] *Office of Educational Administration, Guilin University, Guilin, Guangxi 541006, China*
[2] *Guilin Medical University, Guilin, Guangxi 541000, China*

This paper offers a brief introduction to the decision tree model in data mining technology. A clustering algorithm was employed to discretize the continuous data samples to facilitate processing. Then, the one-semester English course scores of students studying at Guilin Medical University were used for the case study. The improved decision tree model was compared with the ID3 decision tree model and the unmodified decision tree model. The results showed that the improved decision tree model was faster to build and, moreover, had higher classification accuracy than the other two models. The final exam score had the most impact on the overall score, followed by the scores for in-class tests, completion of daily homework, and number of lateness to class.

Keywords: educational informatization, data mining, decision tree, K-means

1. INTRODUCTION

Education is, or should be, a priority of any country as it affects a nation's future development. At the same time, with the development of the times and social changes, the types of talents needed in the workplace will also change constantly. Therefore, education, with its main goal of cultivating talents required for social and economic development, will also evolve (Jahan & Shahariar, 2020). The rapid development of information technology also promotes educational informatization, among which the more intuitive are all kinds of educational materials that are paperless. For example, students' academic performance, daily learning behavior, and other data can be stored in a school's teaching database. With the progress of technology, the storage capacity of the database is also increasing, and the records of students' learning outcomes are more detailed (Cheng & Hsu, 2017). Traditional mathematical statistics cannot effectively capture the information hidden in big data derived from teaching and learning (Nguyen & Jung, 2016). The data mining technology made possible by more sophisticated computers can mine important data knowledge from such big data and provide a valuable reference for teaching. Ahmed et al. (2018) proposed a framework for predicting the performance of freshmen majoring in computer science. Knowledge extracted from the prediction model was used to identify and analyze students' learning outcomes. Yao et al. (2015) proposed a method to predict changing trends in student achievement based on a C4.5 decision tree. Agus et al. (2017) used the ID3 algorithm to predict students' achievement response and found that personal attitudes towards learning had a significant impact on their responses. This paper briefly introduces the decision tree model in data mining technology. A clustering algorithm was adopted to discretize the continuous data in the sample, and a case study was carried out using one semester of English course scores of students at Guilin Medical University.

*Corresponding address: Guilin Medical University, No. 1, Zhiyuan Road, Lingui District, Guilin City, Guangxi Zhuang Autonomous Region 541000, China Email: jc_mojc@hotmail.com

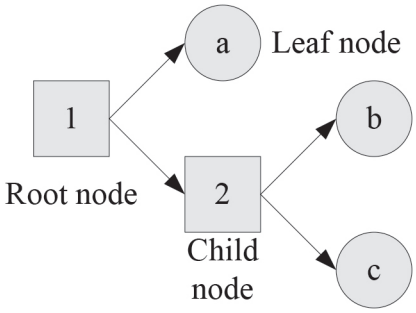

Figure 1 Basic structure of a decision tree model.

2. DATA MINING ANALYSIS OF STUDENT ACHIEVEMENT

The development of information technology has led to the informatization of the education sector. This means that some student data, such as academic performance outcomes, will be recorded and stored in the school's teaching database (Putri, 2020). The recorded data allows teachers to analyse students' learning outcomes, thus providing a basis for the formulation of effective teaching strategies. Traditional mathematical statistics provide relatively superficial information, such as the trend of students' performance in a subject and the areas of weakness that need to be addressed. Although this relatively superficial information can, to a certain extent, assist teachers with lesson planning and teaching strategies, teachers need to further analyze the information based on their experience, which is not intuitive (Nuankaew et al., 2019).

2.1 Decision Tree Model

The decision tree model adopted in this paper is a data mining technology, which has a simple structure and can visualize the analysis process (Hooshyar & Yang, 2021). The decision tree model is a tree-like structure as shown in Figure 1. It is comprised of three types of nodes: the root node and the child node are used for classification, the edge is the judgment condition dependent on the classification attribute, and the leaf node is the classification result (Rajamoni et al., 2022). Similarly, taking Figure 1 as an example, when the decision tree model is used, the samples are first classified using the classification function represented by root node 1; if the conditions are met, the samples are divided into a leaf node representing a classification result; otherwise, the samples are passed down to child node 2 and classified using the classification attribute of child node 2. If the conditions are met, the samples are divided into leaf node b; otherwise, they are divided into leaf node c.

According to the decision tree model described above, the selection of the classification attributes of the root node and child node will directly affect the classification effect of the model. Therefore, the decision tree model needs to build the classification attributes of root nodes and child nodes through training samples. In general, three steps are used to generate a decision tree model. Firstly, samples are collected and pre-processed. Secondly, the standard attributes of relevant data are used to establish the classification feature attributes of the root node and the child node (Berrar & Dubitzky, 2013), i.e., a classification feature is selected as the root node or the child node to build a decision tree model. Thirdly, the decision tree model is pruned.

The standard attributes of relevant data used to build the decision tree model in the above steps are:

$$\begin{cases} H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \\ Gain(T) = H(S) - H(S/T) \\ GainRatio(T) = \frac{Gain(T)}{IV(T)}, \\ IV(T) = \sum_{v=1}^{V} \frac{|S_v|}{|S|} \log \frac{|S_v|}{|S|} \end{cases} \tag{1}$$

where $p(x_i)$ is the probability of category x_i, X is the set of x_i (there are n categories), $H(X)$ is the entropy of all categories (Lakkakula et al., 2015), $H(S)$ is the information entropy of dataset S, $H(S/T)$ is the information entropy obtained after dividing S by feature T, $Gain(T)$ is the information gain when feature T is adopted, $IV(T)$ is the information gain rate when feature T is adopted, v is the value of feature T (there are V types), $|S|$ is the number of samples in dataset S, and $|S_v|$ is the number of samples with value characteristics.

2.2 Decision Tree Model Improved by the Clustering Algorithm

In this paper, a decision number model is employed to analyze students' grades. For the purpose of analysis, only part of the feature attributes used to indicate the performance are discrete, while most of them are continuous. As a result, when constructing a decision tree (Prasad & Naidu, 2013), it is difficult to determine the feature values that should be used as the segmentation boundary when the feature attributes with continuous values are used to divide the sample set. If the number of value types in the sample set is used, too many dividing lines will make the calculation cumbersome. Therefore, it is necessary to discretize the feature attributes with continuous values (Cai & Ding, 2023), i.e., to divide the values of feature attributes into different intervals, and regard each interval as a value of the feature attribute. For the partitioning of discrete intervals, the traditional approach relies on teachers' experience (Wang, 2023), which is both dependent on their expertise and lacking in dynamism. Therefore, in this study, the K-means clustering algorithm was used to divide the interval. Suppose there are m samples, and each sample has n feature attributes pointing to the final

Table 1 Partial records of a student's academic file for a semester in the English course.

Student number	Name	Class	Daily homework completion degree	Score achieved for in-class tests	Final exam score	Number of lateness to class	Overall performance
33 * * * * 11	Hu **	Class A, Grade One	60%	85	89	3	Excellent
32 * * * * 24	Chang **	Class B, Grade Two	40%	67	75	0	Acceptable
32 * * * * 25	Money **	Class C, Grade Two	50%	75	68	2	Unacceptable
…	…	…	…	…	…	…	…

score. In this paper, the scores achieved by students for different assessment items are used as feature attributes, and they belong to continuous values. The discretization process can be summarized in the following steps.

① The score values of each feature attribute are arranged in descending order, and then according to the normal distribution law, each feature attribute is allocated to a region represented by "excellent", "good", "medium", or "poor".

② The mean value of scores for each region of each feature attribute is calculated, and then the score value nearest to the mean value is taken as the initial clustering center of this region. Thus, there will be four initial clustering centers in each feature attribute.

③ K-means (Cleghern et al., 2017) is employed to classify the data in each feature attribute, and the minimum and maximum values in each score set are the judgment interval of this classification.

④ According to the classification results of the previous step, the continuous values of the feature attributes in the sample set are discretized. For example, if the score value of a feature attribute of a sample is in the judgment interval of the classification of "excellent", the score value of the feature attribute will be replaced with "excellent".

After discretizing the values of the feature attributes of the sample through the above steps, a decision tree model is generated.

3. CASE STUDY

3.1 Data Sources

This paper collected the English course learning file records, for one semester, of students at Guilin Medical University, some of which are shown in Table 1. In the collected samples, "overall performance" was the final classification result of the decision tree model, and the other indicators are the feature attributes used to indicate the final classification result. Among the above feature attributes, the student's name, class, and student number had little impact on the overall English performance, so they were not used in the subsequent construction process.

3.2 Relevant Parameters

After discretizing the continuous data of the sample, the values of the feature attributes were arranged in descending order according to the steps mentioned above, and the normal distribution law (Malik & Khan, 2017) was used for preliminary division. According to the normal distribution law, the values of the feature attributes in descending order are classified as "excellent", "good", "medium", and "poor" respectively according to the density distribution of 16%, 34%, 48%, and 2% respectively. The best neighbor point of the mean value of each class was used as the initial clustering point of the class, the K value of K-means was set at 4, and the maximum number of iterations was set at 500.

3.3 Evaluation Index

Precision, recall rate, and F-number were used to measure the classification performance of the model. The formulas are:

$$\begin{cases} P = \frac{TN}{FN+TN} \times 100\% \\ R = \frac{TN}{FP+TN} \times 100\%, \\ F = \frac{2\times P\times R}{P+R} \times \% \end{cases} \tag{2}$$

where TP is the number of true positive cases, FN is the number of false negative cases, FP is the number of false positive cases, and TN is the number of true negative cases. In addition to the improved decision tree model, the ID3 decision tree and the unmodified decision tree model were also tested.

3.4 Experimental Results

After the three decision tree models were constructed using the sample data of the training set, the classification performance of the three models was verified by the sample of the test set; the results are shown in Figure 2. The ID3 decision tree had the worst classification performance in terms of students' overall performance in the English course, the unmodified decision tree model had relatively higher classification performance, and the improved decision tree model had the best classification performance. Moreover, the F value of the improved decision tree model was 94.1%.

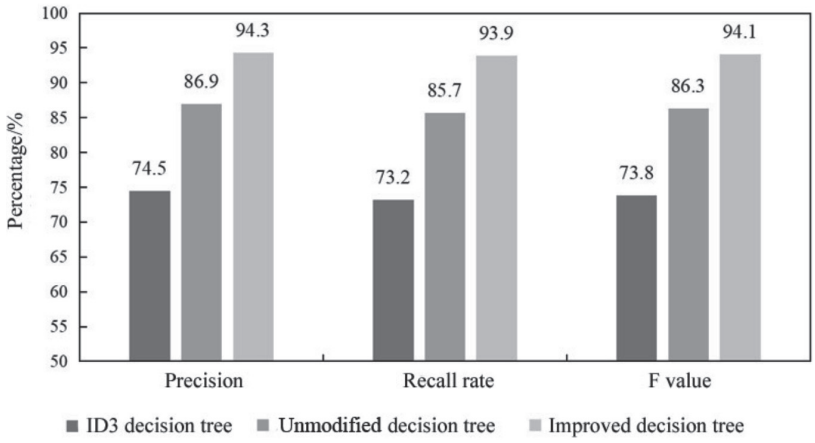

Figure 2 Classification performance of three decision tree models.

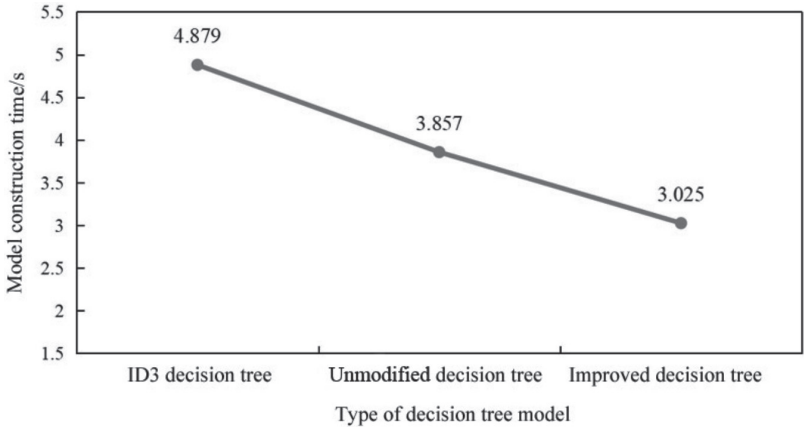

Figure 3 Construction time of three decision tree models.

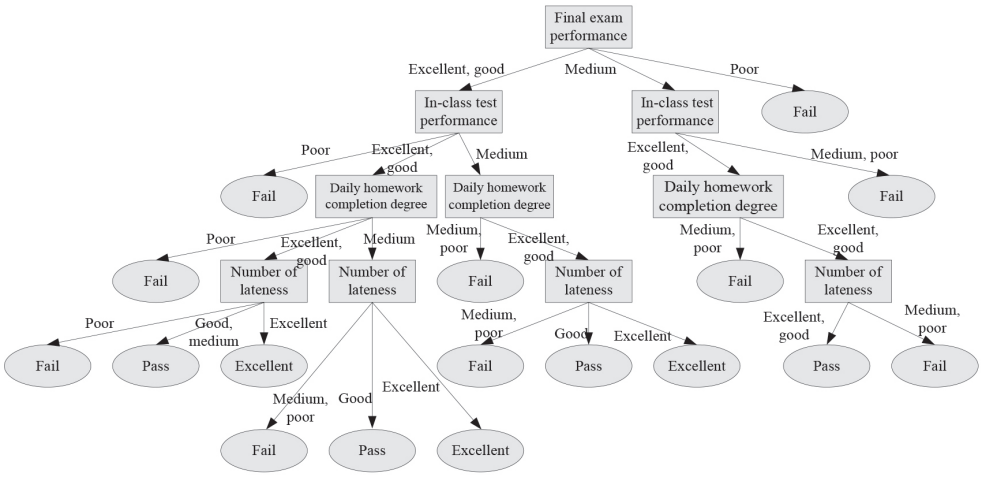

Figure 4 Construction results of the improved decision tree model.

The time required for the construction of the three decision tree models using the samples in the training set is shown in Figure 3. The ID3 decision tree model took the most time to build, followed by the unmodified decision tree model, and the improved decision tree model took the least time to build.

The construction results of the improved decision tree model proposed in this paper using the samples of the training set are shown in Figure 4. Starting with the feature attribute of the root node, and following the path representing the feature attribute classification standard, the sample classification result was reached. The path is the classification rule in the decision tree. For example, the classification rules can be extracted from the figure: if final exam performance = excellent or good, in-class test performance = excellent or good, daily homework completion degree = excellent or good, and number of lateness to class = excellent, then overall performance = excellent.

4. DISCUSSION

The progress of information technology also promotes the evolution of education informatization. This means that colleges and universities collect a large amount of information about students, in addition to data related to students' academic performance. The analysis of the latter can reveal the factors that affect students' performance, enabling teachers to apply more effective teaching strategies. However, the informatization of education has produced big data which cannot be processed efficiently by traditional mathematical statistics; nor can traditional methods derive more than just superficial patterns from the data. Teachers are required to analyze these patterns in order to obtain more in-deep information. With the development of more sophisticated computers and related technologies, data mining is possible and is being widely used in many industries; it can also be used for the analysis of student performance data. In this paper, the decision tree model was used to analyze students' performance. At the same time, in order to solve the problem that the decision tree model is unable to deal efficiently with continuous data, this paper used a clustering algorithm to discretize the data. Then, a case study was carried out on the data related to the English course scores of students at Guilin Medical University over one semester. The optimized decision tree model was compared with the ID3 decision tree model and the unmodified decision tree model. The decision tree model improved by the K-means algorithm was not only faster to build, but also had greater classification accuracy after construction than the other two models.

In addition, it can be seen from the generation steps of the decision tree model that the key to the construction of the decision tree is the selection of the feature attributes of the root node and the child node. In this paper, the decision tree model used the information gain rate to compare feature attributes, so as to select the node order that can make the classification result more closely aligned with the actual classification. The order in which nodes are arranged indicates the importance of the feature attribute represented by the node. For instance, in the proposed decision tree model, when the feature attribute "final exam score" was taken as the root node, it had a higher information gain rate than other feature attributes. In other words, this feature attribute can be more suitable for actual classification in the initial sample division and had a greater impact than the other feature attributes. Similarly, the feature attribute "final exam performance" had the greatest influence on the overall performance, followed by in-class test performance, daily homework completion degree, and number of lateness to class.

5. CONCLUSIONS

This paper provides a brief introduction to a decision tree model for which a clustering algorithm was used to discretize the continuous data in the samples. Then, a case study was carried out on the data related to the one-semester English course scores of students at Guilin Medical University. The performance of the improved decision tree model was compared with that of the ID3 decision tree model and the unmodified model. The proposed decision tree model had the highest accuracy in terms of classifying the overall English scores. The improved decision tree model required the least amount of construction time. The feature attribute "final exam performance" had the greatest influence on the overall score, followed by in-class test performance, extent of daily homework completion, and number of lateness to class.

REFERENCES

1. Agus, N., Lusi, M., & Deasy, P. (2017). Implementation ID3 Algorithm to Predict Children Achievement in Response (Case Study Children Playgroup School). *Journal of Engineering & Applied Sciences*, *12*(2), 204–207.
2. Ahmed, R.M., Omran, N., & Abdelmgeid. (2018). Predicting and Analysis of Students' Academic Performance using Data Mining Techniques. *International Journal of Computer Applications*, *182*(32), 1–6.
3. Berrar, D., & Dubitzky, W. (2013). Information Gain. *Springer New York*.
4. Cai, J. & Ding, Y. (2023). Deep Mining Method of Distributed Data Association Based on Decision Tree Algorithm. *Engineering Intelligent Systems*, *31*(3), 229–273.
5. Cheng, Y.F., & Hsu, Y.S. (2017). Decision tree for investigating the factors affecting graduate salaries. *Journal of Research in Education Sciences*, *62*(2), 125–151.
6. Cleghern, Z., Lahiri, S., Özaltin, O., & Roberts, D.L. (2017). Predicting future states in DotA 2 using value-split models of time series attribute data. *FDG'17: Proceedings of the 12th International Conference on the Foundations of Digital Games*, 1–10.
7. Hooshyar, D., & Yang, Y. (2021). Predicting Course Grade through Comprehensive Modelling of Students' Learning Behavioral Pattern. *Complexity*, *2021*(1), 1–12.
8. Jahan, N., & Shahariar, R. (2020). Predicting fertilizer treatment of maize using decision tree algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *20*(3), 1427–1434.
9. Lakkakula, N.P., Naidu, M.M., & Reddy, K.K. (2015). An entropy based elegant decision tree classifier to predict precipitation. *2014 European Modelling Symposium*, 11–19.
10. Li, L., Yao, S.M., Ou, Z., & Chen, Q.J. (2015). Forecast of Student Achievement Variation Trend Based on C4.5 Decision Tree. *2015 International Conference on Artificial Intelligence and Industrial Engineering*, 383–386.
11. Malik, A.J., & Khan, F.A. (2017). A hybrid technique using binary particle swarm optimization and decision tree pruning for network intrusion detection. *Cluster Computing*, *21*(3), 1–14.
12. Nguyen, H.L., & Jung, J.E. (2016). Statistical approach for figurative sentiment analysis on Social Networking Services: a case study on Twitter. *Multimedia Tools & Applications*, *76*(6), 1–14.
13. Nuankaew, P., Nuankaew, W., & Temdee, P. (2019). Institution recommendation using relationship optimisation between program and student context. *International Journal of Higher Education and Sustainability*, *2*(4), 279.
14. Prasad, N., & Naidu, M.M. (2013). Gain ratio as attribute selection measure in elegant decision tree to predict precipitation. *2013 8th EUROSIM Congress on Modelling and Simulation*, 141–150.

15. Putri, G.A. (2020). Implementation of the C4.5 Algorithm to Predict Student Achievement at SMK Negeri 6 Surakarta. *IJIE (Indonesian Journal of Informatics Education)*, *4*(2), 10–19.
16. Rajamoni, R.N., Kumar, M.S.S., & Leela, B.C. (2022). Factors Affecting the Academic Performance of Students with Hearing Impairment. *Revue d'intelligence artificielle*, *36*(4), 569–574.
17. Wang, K. (2023). Evaluation Model of Police Physical Education Teaching Mode Reform Effect based on IC3 Decision Tree Algorithm. *Engineering Intelligent Systems*, *30*(4), 295–301.