DOI: 10.23977/curtm.2023.060818
ISSN 2616-2261 Vol. 6 Num. 8

Application of Decision Tree Algorithm in Educational Data Mining

Sa Chen*, Xiankun Lin

University of Shanghai for Science and Technology, Shanghai, 200093, China
**Corresponding author: chensaa@163.com*

Keywords: Decision Tree, Student Classification, Educational Data Mining, University Teaching

Abstract: Decision tree algorithms in Educational Data Mining (EDM) emerges as a powerful method for student classification according to their learning station. This article emphasizes the necessity to classify students due to the changing scenario of China's higher education system, which has shifted from elitist to popularization, and lead to university students' differences in learning abilities. For the specialty characteristics of educational data, this study adopts decision tree algorithm based on ID3 to classify students. This article presents an analysis of the application of decision tree algorithm in the course "Hydraulic and Pneumatic Technology Course Design," which is a mandatory course for Mechanical Design Manufacture and Automation Major. This study identifies five splitting features that affect the students' ability to achieve success in this practical course, and build the decision tree model. The learning samples for this algorithm mode are collected from Bachelor students of the University of Shanghai for Science and Technology (USST). According to the categorization results teachers make informed decisions based on the insights provided by the algorithm to improve the learning experience and academic performance of students. Similarly categorization results also provide personalized guidance for students, which is beneficial in ensuring their success and ultimately improving overall educational outcomes.

1. Introduction

With the progress of information technology, various intelligent teaching media and tools have been integrated into the education system, facilitating the collection and analysis of diverse educational data. This process of extracting and analyzing data from educational systems to obtain valuable insights into students' learning and subsequently make informed decisions is referred to as Educational Data Mining (EDM).[1] EDM constitutes a cross-disciplinary field that combines techniques from statistics, computer science, and education to extract significant information from educational data sets. [2] These data sets can be obtained from numerous sources, such as student assessments, attendance records, clickstream data, and additional educational information systems. The application of EDM makes it feasible to identify causal factors that impact student performance, predict student outcomes, evaluate instructional efficiency, and enhance the overall educational experience. The insights garnered through EDM can enable educators to make data-driven decisions that augment the quality of teaching and learning.

In educational data mining, classification is a key technique. Well-known classification models include Logistic Regression, K-Nearest Neighbor model, Decision Tree model, Naive Bayes and Support Vector Machine model, among others. Decision tree model is one of the most classic models due to its simple principle, high accuracy, and effective data processing.

Several studies have reported the successful application of decision tree algorithm in EDM. There are two main applications in higher education: First is applied to student performance prediction. These research articles such as Jeff, C. F., Tony, C. Y.(2020) [3],Wang, F., Zhao, C. (2018) [4], Alaee, S. , Silberglitt, R. (2020) [5] , Patakamuri, R. D., & George, B. C. (2018) [6] demonstrated an approach for analyzing student performance using decision tree algorithm and prediction students' future performance by inputting basic information about the student and their historical performance, so as to help educators to develop student-specific educational programmes and to improve student performance and learning ability.

Second is applied to Student classification. The research articles A. El Khalfi, A. Aqqal (2017)[7], Yang, J.Y., Guan, l. X., Yu, L.(2016) [8], Hussain, A., Khan, M. (2022) [9] deal with using decision tree algorithm to classify students according to their learning ability, interests and other aspects, so that teachers can personalize teaching and management.

Another study about applications of decision tree algorithm in EDMs uch as Subitha, S. Siva, kumar, V.(2016) [10] develop an improved decision tree algorithm based on ID3 to predict whether the students continue or drop their studies and improved decision algorithm on educational dataset is greater.

This study aims to provide a case of student classification by decision tree algorithm. The paper is structured as follows: Section 2 describes the need to classify students in university teaching for personalizing learning. Section 3 outlines the implementation of Decision Tree Algorithm. Section 4 provides applications of the algorithm in a practical courses for Mechanical Design Manufacture and Automation Major and the dataset used is collected from Bachelor students of the University of Shanghai for Science and Technology(USST).

2. Necessary to Classify Students for Personalizing Learning in University Teaching

Dou to the change of China's higher education from elitist to popularization, the enrollment scale of colleges and universities is increasing, and individual students become more and more different. Students have diverse backgrounds and different levels of knowledge and skills. Some students may have extensive prior knowledge of the subject matter and need more advanced, challenging coursework, while others may need more foundational instruction to build up their understanding.

In order to better meet the learning needs and developmental needs of different students, it is necessary to accurately classify students by analyzing student data and identify patterns in their learning behavior. And by accurately classifying students stratified teaching can be achieved. Thus, it is sure that all students receive instruction and assignments that are appropriate for their level of understanding, as well as more personalized attention from instructors. It also allows students to move at different paces and work on assignments that are tailored more closely to their needs.

Furthermore, stratified teaching creates a learning environment that is more inclusive and equitable. Students who require more support can receive a targeted approach that helps them overcome their challenges and become successful, while high-performing students can receive more advanced instruction that helps them reach their full potential.

This paper adopt decision tree classification algorithm to categorize students according to mined data from their learning activities. It is useful for students to select appropriate learning content and achieve personalized learning. Also, teachers can tailor teaching content at different levels to better match students' learning abilities and increase their motivation and initiative based on the classification results.

3. Implementation of decision tree algorithm

3.1. Description of decision tree algorithm

Decision tree algorithm is a machine learning methods which is a predictive modeling technique that builds a tree-like model to map the input features to the target variable. The decision tree algorithm follows a divide-and-conquer approach to recursively split the data based on the attributes with the highest information gain or the best split. The algorithm is built in a top-down fashion, As shown in the figure 1, where each internal node (such as node B, C, D) represents a decision based on a feature, and each leaf node(such as y1,y2,……y6) represents a prediction. The top node A is called root node which contains the entire training set of samples.

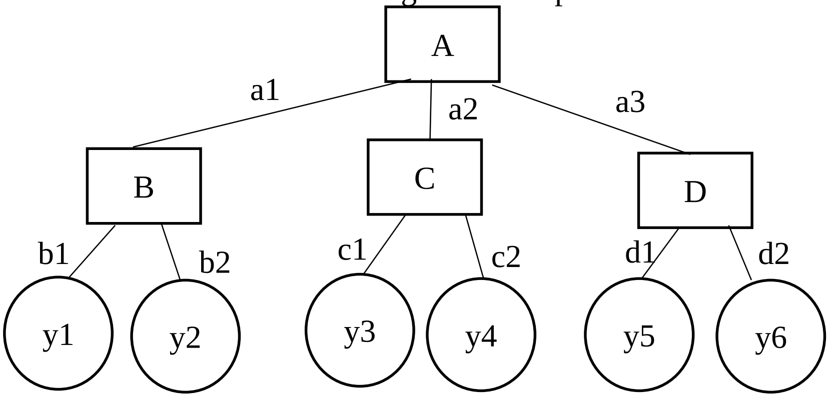

Figure 1: Description of tree-like model in the decision tree algorithm.

At each internal node, the algorithm chooses the attribute that best splits the data into purest subsets by maximizing the information gain, Gini Index, or entropy. The purest subsets are those that contain the same class of labels, making it easy to make accurate predictions.

The decision tree algorithm continues to split the data until all leaf nodes are pure or when a predefined stopping criterion is met. Once the tree is built, it can be used to make predictions on new unseen data by traversing the tree from the root to a leaf node based on the values of the input features, and returning the class label stored at that leaf node as the prediction. ID3, C4.5 and CART are the classic tree classification algorithms. In this study according to the characteristics of educational data, ID3 algorithm is more appropriate.

3.2. Implementation for Algorithm based on ID3

"Information entropy" measures the impurity or randomness of a set of samples. It can be calculated using the following formula:

$$Ent(\mathrm{S}) = -\sum_{i=1}^{|k|} p_{\_\mathrm{i}} \log_2 p_{\_i} \tag{1}$$

Where Ent(S) is the entropy of the set S, and p_i is the proportion of samples in S that belong to the i-th class.

"Information gain" measures the reduction in entropy achieved by splitting the set S based on a certain feature. It can be calculated using the following formula:

$$Gain(D,A)=Ent(D)-\sum_{v=1}^{n}\frac{|S\_v|}{S}Ent(S\_v) \tag{2}$$

where Gain(S, A) is the information gain achieved by splitting the set S based on the feature A, |S_v| is the number of examples in subset S_v that correspond to the feature value of A, |S| is the total number of samples in S, and Ent (S_v) is the entropy of subset S_v. Information gain measures how much uncertainty or randomness is removed from the set S by splitting it based on the feature A. The higher the information gain, the better the attribute is for splitting the data, and it should be selected as the split node for the decision tree algorithm.

From root node, at each recursive call, Selects the feature with the highest information gain as the splitting feature, and creates a new node with this feature. Then recursively calls itself on each subset of instances that correspond to each possible value of the splitting feature. The process continues until all instances in a subset have the same class label, or until a stopping condition is met.

4. Algorithm applied to practical courses in university teaching

This article takes a practical courses "Hydraulic and Pneumatic Technology Course Design" which is a mandatory course for Mechanical Design Manufacture and Automation Major as a study case to specifically analyze how to use decision tree algorithm to classify students in the teaching process.

"Hydraulic and Pneumatic Technology Course Design" is a professional practice course. Its key characteristics are comprehensiveness, independence, and diversity. Therefore, it is crucial to classify students enrolled in this course. Teachers can assign various design projects to various pupils based on the classification results.

The main task of the course is to design a special machine tool for machining spool of hydraulic Reversing Valve. Students are required to complete the following: design of the hydraulic system oil circuit, drawing of the oil circuit schematic, completion of the structure design of the tailstock, longitudinal and transverse tool holder mechanism, rear tool holder mechanism and integrated block. The spool of hydraulic reversing valve is shown in Figure 2.

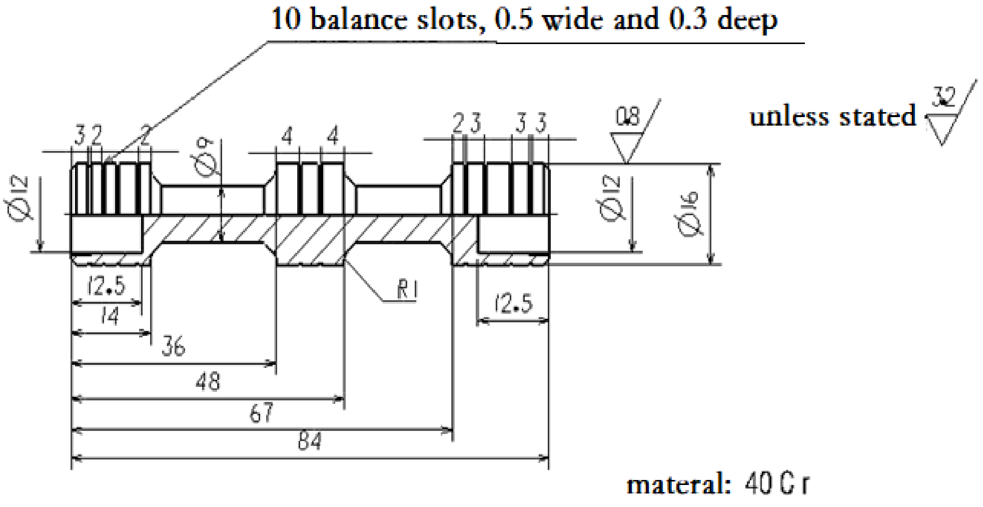

Figure 2: Hydraulic reversing valve spool to be machined

4.1. Identify splitting features of student classification

First step in the algorithm is to determine the splitting features of the elements in the set. In this course "Hydraulic and Pneumatic Technology Course Design", splitting features are identified based on course syllabus and core aims.

One of the core aims of this this course is to augment and deepen students' fundamental theoretical expertise in hydraulic and pneumatic technology, building upon prior knowledge acquired. In this regard, they will be expected to achieve proficiency in comprehending the structure and operational principles of fundamental hydraulic and pneumatic components, as well as the design process and calculation of essential hydraulic systems. Attainment of sound theoretical understanding constitutes a pivotal prerequisite for successfully completing the course design, so one of the splitting features used to classify students are basic theoretical knowledge mastery ability. The degree of proficiency in this aspect of the curriculum will be assessed and categorized at four levels: excellent, good, fair, and poor, commensurate with the students' grades in the prerequisite-course--Hydraulic and Pneumatic Technology.

Another core aims of this course is to develop students' ability to design, manufacture and process planning for hydraulic and pneumatic equipment through comprehensive training. So the students' ability to independently design, innovate and plan is also one of the splitting features used to classify students. This attribute can be obtained from whether students have participated in relevant professional competitions or innovative experimental projects, and is categorized into: yes or no.

This course requires students to draw assembly and component diagrams of hydraulic stations, or non-standard integrated block component diagrams and assembly outlines. Therefore, the ability of engineering drawing is one of the splitting features used to classify students. Students' engineering drawing ability can be categorized into four levels: Excellent, Good, Fair and poor, based on their grades in the prerequisite Fundamentals of Engineering Drawing course.

This course also requires students to have experimental skills. Through disassembly and assembly experiments, students can more intuitively observe the internal structure of components and understand their processing requirements. Through circuit control experiments, students can deepen their understanding of the advantages of fluid as a transmission medium. Therefore, students' experimental operation ability is also one of the classification attributes and is categorized into three levels: excellent, medium and poor.

This course requires students to access various design manuals and product samples, and to select available hydraulic components such as hydraulic pumps, hydraulic cylinders and various hydraulic valves. Therefore, the ability to access literature is also one of the splitting features used to classify students. This attribute is categorized into three levels: Excellent, medium and poor.

This course necessitates that students possess the capability to utilize numerous design manuals and product samples, while selecting from a diverse range of hydraulic components such as hydraulic cylinders, hydraulic pumps, and assorted hydraulic valves. Consequently, the ability to access literature is also one of the splitting features used to classify students. This attribute is categorized into three levels: Excellent, medium and poor. Based on the above analysis, the splitting features of algorithm are as follows: basic theoretical knowledge mastery ability, engineering drawing ability, creative and design ability, experimental operation ability, and literature retrieval ability. The students are classified two categories: higher level and lower level.

Decision tree algorithm need training samples and learns from samples to build a model that can generalize the patterns and relationships in the data set, making it possible to apply the model to classify new, unknown observations. Nowadays, a variety of online learning platforms and teaching management systems are extensively implemented within universities. Vital learning data such as encompassing assignments, exams, and classroom performance metrics and so on can be readily accessed from these teaching management systems. In this study, this learning sample is collected from Bachelor students of University of Shanghai for Science and Technology. The learning data of students who have attended the course in the previous academic semester can access directly from the teaching platform database. The sample of students’ data are shown in Table 1:

Table 1: Sample of students’ data based on splitting features.

NO.	basic theoretical knowledge	engineering drawing ability	creative and design ability	experimental operation ability	literature retrieval ability	Classification results
1	Excellent	Good	No	Excellent	Average	Higher level
2	Good	Fair	No	Average	Average	Lower level
3	Fair	Excellent	No	Average	Average	Lower level
4	Fair	Good	No	Average	Average	Lower level
5	Good	Excellent	No	Average	Excellent	Higher level
6	Poor	Fair	No	Average	Poor	Lower level
7	Fair	Good	No	Excellent	Excellent	Higher level
8	Good	Good	No	Excellent	Average	Higher level
9	Excellent	Excellent	Yes	Excellent	Excellent	Higher level
10	Excellent	Fair	No	Average	Average	Lower level
11	Fair	Fair	No	Average	Average	Lower level
12	Good	Fair	No	Average	Excellent	Higher level
13	Good	Good	No	Excellent	Excellent	Lower level
14	Fair	Poor	No	Poor	Average	Lower level
15	Excellent	Excellent	Yes	Excellent	Average	Higher level
16	Fair	Fair	No	Average	Average	Lower level
17	Fair	Good	No	Average	Average	Lower level
18	Good	Excellent	No	Average	Excellent	Higher level
19	Poor	Poor	No	Poor	Poor	Lower level
20	Fair	Fair	No	Average	Excellent	Lower level

4.2. Calculate information gain of splitting features

First, calculate the information entropy of root node according to table 1 sample. As can be seen from the table 1, students with higher ability account for 8/20 of the total students and students with lower ability account for 12/20, so the information entropy of root node can be derived as:

$$Ent\left(\mathrm{S}\right)=-\sum_{i=1}^{|2|}p_{\_i}\log_2 p_{\_i}=-\left(\frac{8}{20}\log_2\frac{8}{20}+\frac{12}{20}\log_2\frac{12}{20}\right)=0.971$$

Then, calculate information gain of all attributes {basic theoretical knowledge, engineering drawing ability, innovation and design ability, Experimental operation ability, literature retrieval ability }. The information gain of "basic theoretical knowledge" is calculated below.

The attribute "basic theoretical knowledge" has four possible values: {Excellent, Good, Fair, Poor}. Use this attribute to split the sample set S, and obtain four subsets: S_1 (basic theoretical knowledge = Excellent), S_2 (basic theoretical knowledge = Good), S_3 (basic theoretical knowledge = Fair) and S_4 (basic theoretical knowledge = Poor).

Subset S_1 contains students numbered {1, 9, 10, 15}, where the positive example (higher level students) P1 = 3/4 and the negative example (lower level students) p2 = 1/4; Subset S_2 contains {2, 5, 8, 12, 13, 18}, where the positive example (higher level students) P1 = 4/6 and the negative example (lower level students) p2 = 2 /6; Subset S_3 includes {3, 4, 7, 11, 14, 16, 17, 20}, where the positive example (higher level students) P1 = 1/8 and the negative example (lower level students) p2 = 7/8; And subset S_4 includes students numbered {6, 19}, where the positive example (higher level students) P1 = 0 and the negative example (lower level students) p2 = 2.

According to formula (1), separately calculate the information entropy for four subsets based on attribute "basic theoretical knowledge" as:

$$Ent\left(S_{\_1}\right)=-\left(\frac{3}{4}\log_2\frac{3}{4}+\frac{1}{4}\log_2\frac{1}{4}\right)=0.811,$$

$$Ent\left(S_{\_2}\right)=-\left(\frac{4}{6}\log_2\frac{4}{6}+\frac{2}{6}\log_2\frac{2}{6}\right)=0.918,$$

$$Ent\left(S_{\_3}\right)=-\left(\frac{1}{8}\log_2\frac{1}{8}+\frac{7}{8}\log_2\frac{7}{8}\right)=0.544$$

$$Ent\left(S_{\_4}\right)=-\left(\frac{0}{2}\times\log_2\frac{0}{2}+\frac{2}{2}\times\log_2\frac{2}{2}\right)=0$$

According to formula (2), calculate the information gain for attribute "basic theoretical knowledge" as follow:

$$Gain\left(\mathrm{S}, basic\ \ theoretical\ \ knowledge\right)=Ent\left(S\right)-\sum_{v=1}^{4}\frac{\left|S_{\_v}\right|}{S}Ent\left(S_{\_V}\right)$$

$$=0.971-\left(\frac{4}{20}\times0.811+\frac{6}{20}\times0.918+\frac{8}{20}\times0.544+\frac{2}{20}\times0\right)=0.3158$$

Similarly, the information gain for other attributes are calculated separately:

$$Gain\left(\mathrm{S}, Engineering\ \ Drawing\ \ ability\right)=0.423$$

$$Gain\left(\mathrm{S}, innovation\ \ and\ \ design\ \ ability\right)=0.145$$

$$Gain\left(\mathrm{S}, Experimental\ \ operation\ \ ability\right)=0.289$$

$$Gain\left(\mathrm{S}, literature\ \ retrieval\ \ ability\right)=0.204$$

Comparing these values of information gain, the attribute "engineering drawing ability" has the highest value. So this attribute is selected as the root node to complete first splitting to the sample set. For four values generate four branches and generate four sub-nodes namely internal nodes, as shown in Figure 3:

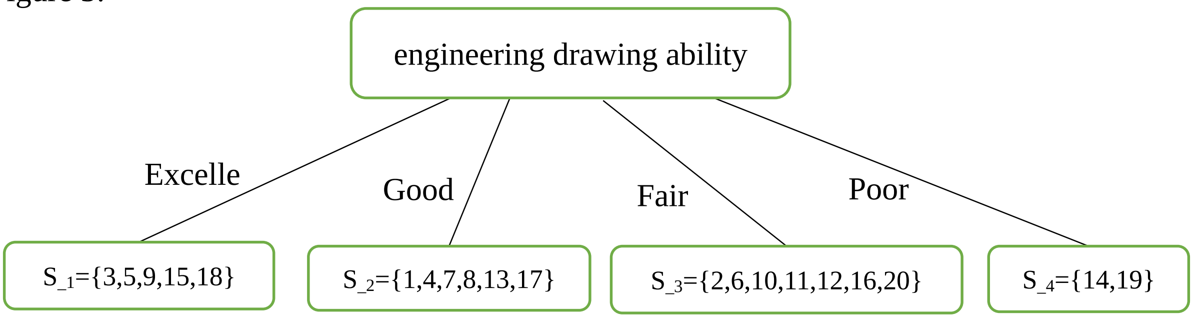

Figure 3: Complete first splitting to the sample set.

4.3. Create Decision tree learning model

Recursively calls on each subset that correspond to each possible value of the splitting feature. At each recursive call, the function selects the feature with the highest information gain as the splitting feature, and creates a new node with this feature.

Taking the sub-node ("engineering drawing ability = excellent"), Continue recursively to calculate the information gain based on the available attributes for subset S_1 in Figure 3. As follow:

$$Gain\left(\mathrm{S}_{\_1}, basic\ theoretical\ knowledge\right) = 0.722$$

$$Gain\left(\mathrm{S}_{\_1}, innovation\ and\ design\ ability\right) = 0.171$$

$$Gain\left(\mathrm{S}_{\_1}, practical\ ability\right) = 0.171$$

$$Gain\left(\mathrm{S}_{\_1}, literature\ review\ ability\right) = 0.322$$

Comparing these values of information gain for each attributes, the attribute "basic theoretical knowledge" has the highest information gain, so this attribute is chosen to split subset S_1. Similarly, recursively calls on each other subset to build the decision tree model. All students who have taken this course can be accurately categorized so that different design tasks can be assigned to meet the learning needs of students at different levels based on the categorization results.

5. Conclusions

The application of decision tree algorithms in educational data mining has emerged as a powerful method for predictive analysis of students' learning station. The algorithm enables the identification of key factors that are crucial in predicting students' performance and learning outcomes. As a result, teachers can make informed decisions based on the insights provided by the algorithm to improve the learning experience and academic performance of students. Decision tree algorithms also provide personalized guidance for students, which is beneficial in ensuring their success and ultimately improving overall educational outcomes. Future research could explore how decision trees can be combined with other machine learning algorithms to improve accuracy and performance.

References

[1] Romero C., Ventura S. (2010). Educational data mining: A review of the state-of-the-art. IEEE transactions on systems, man, and cybernetics, part C (Applications and Reviews), 40(6), 601-618.

[2] Li T., Fu G. S. (2010) An overall view of the educational data mining domain. Modern Educational Technology, 20(10):21-25.

[3] Jeff C. F., Tony C. Y. (2020) Measuring Students' Academic Performance through Educational Data Mining. International Journal of Information and Education Technology, 10 (11), 797-804

[4] Wang F., Zhao C. (2018). Application of decision tree algorithms in student performance prediction analysis. Journal of Big Data, 5(1), 1-12.

[5] Alaee S., Silberglitt R. (2020). Predicting academic performance using decision tree and logistic regression algorithms. Journal of Big Data, 7(1), 1-13.

[6] Patakamuri R. D., & George B. C. (2018) Application of decision tree algorithm in educational data mining for student performance prediction. Journal of Computer Science and Engineering Education, 8(2), 24-33.

[7] A. El Khalfi, A. Aqqal. (2017) Student Classification Based on Decision Tree Algorithm: A Case Study. International Journal of Innovative Research in Computer and Communication Engineering, 5 (8).

[8] Yang J.Y., Guan L. X., Yu L. (2016) College Student Classification in Swimming Teaching and Training [P]. Proceedings of the 2016 International Seminar on Education Innovation and Economic Management (SEIEM 2016)

[9] Hussain A., Khan M. (2022) Student's performance prediction model and affecting factors using classification techniques. Education and Information Technologies. Volume 27, Issue 6, 8841-8858

[10] Subitha S. Siva Kumar V.(2016)Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree. Indian Journal of Science and Technology, 9(4).