Scholar Publishing Group
Frontiers in Ocean Engineering
https://doi.org/10.38007/FOE.2021.020202
ISSN 2790-2706 Vol. 2, Issue 2: 9-16

C4.5 Algorithm in the Analysis of Marine Engineering Scores

Warm Katrin*

LJMU, Dept Elect & Elect Engn Comp Sci, Liverpool L3 3AF, Merseyside, England

**corresponding author*

Keywords: Decision Tree Algorithm, C4.5 Algorithm, Ocean Engineering Major, Performance Analysis

Abstract: In the age of network information, science and technology advance by leaps and bounds, and people have put forward higher requirements for a large amount of data and information in their lives. Data mining can extract effective and valuable information that can guide people to make decisions from a large number of existing data by using data mining techniques and algorithms. This paper mainly studies the application of C4.5 algorithm in the performance analysis of ocean engineering major. This paper collects data based on online learning platform, and conducts data processing and correlation analysis of learning behavior data. In this paper, by analyzing the influence of each special on the final grade, the C4.5 algorithm was introduced to improve. Based on C4.5 algorithm, this paper constructs a performance analysis model, which can effectively analyze teachers' performance in each stage of course learning.

1. Introduction

With the popularization of the Internet and the progress of information technology, teaching and sharing knowledge through information digitization has been accepted by more and more groups [1]. In early 2020, the COVID-19 pandemic pushed online teaching further into the track of mainstream teaching model. Online teaching is no longer only the main way for primary and secondary school students to catch up on their lessons, but also has become the inevitable choice for universities to change the teaching mode under the epidemic. At present, there are many network teaching systems available for teachers to teach courses through the network. However, network teaching has the disadvantage of not intuitive classroom teaching, which leads to the teacher's judgment of students' learning situation is not as accurate as classroom teaching. Therefore, how to effectively analyze the effect of network teaching has always been a hot issue in teaching research [2]. Many researches on online teaching show that the data of students' learning paths and learning performance can be used

as an important basis for teachers to evaluate students' learning outcomes and diagnose students' learning difficulties. These data are of valuable research value. Education big data has caused extensive concern of the society from all walks of life, the use of education method combined with the characteristics of data in data mining to analyze data, and explore the important information hidden in the data and then find the teaching problems, to assist teachers in teaching and promoting education reform, improve the quality of teaching has important meaning [3]. In education data, student's result the location data can not only determine student achievement, also can adjust the teaching for teachers and teaching management training plan provides feedback information, so with the help of education of the data mining method, the course of " modern education technology application " performance data for data mining, to explore the question of " what factors affect the teaching " . It plays a positive role in optimizing teaching, improving academic ability of ocean engineering and promoting the development of educational informatization [4-5].

The empirical analysis of predicting student performance at home and abroad has also been developed from initial artificial statistics to common artificial intelligence statistics [6]. Statistical forecasting methods. The key point of this method is the prediction of grades. The prediction model is established by statistical method, and the results calculated by statistical model are used to predict grades. Specific methods include fuzzy clustering, Logistics regression model, PRIDIT principal component analysis, multivariate neural network, etc. [7]. Artificial intelligence grade prediction technology. This method is improved and realized on the basis of statistical prediction methods, and the prediction effect can be improved through the joint use of statistical regression methods [8]. Education as one of the most commonly used methods in data mining, domestic the DT(DT) algorithm is mainly used in the field of education student performance analysis, analysis of students' learning behavior, has been applied to the proportion of college students graduate and prediction and the employment situation, the students' learning experience, to help students' course selection etc. [9]. Compared with China, foreign DT algorithms are more widely applied in education, not only in performance analysis, but also in education donation, calculation of income and expenditure of education funds, student management and other aspects [10].

Therefore, how to effectively evaluate the academic achievements of students specializing in marine engineering in the context of the epidemic situation is an urgent problem that needs to be studied. The present study proposes an integrated analysis algorithm, through the processing of these heterogeneous large-scale learning files, we analyze these learning files from multiple perspectives, identify students' learning behaviors and predict their possible learning outcomes according to their current learning conditions.

2. Achievement Analysis of Ocean Engineering Major Based on C4.5 Algorithm

2.1. Optimize DT C4.5 Algorithm

C4.5 algorithm optimizes the DT generation process on the basis of ID3 algorithm [11]. The resulting DT is an ordinary tree. Firstly, the information gain ratio of attributes is selected as the basis for selecting branch attribute nodes. Secondly, the pruning operation is carried out at the same time when the DT is constructed to avoid the over-fitting phenomenon of the whole tree [12-13]. Before constructing the DT, the data is preprocessed and the transformation from continuous attribute data to discrete attribute data is considered in advance. The following is the pseudocode of the C4.5 algorithm

Input: training sample set D={(x1,y1), (x2,y2),... ... (xn,yn)}

Attribute set A={A1, A2,... ... , an}

Output: A DT

Example Create a root node

If both D's belong to the same class C, then node is returned as a leaf node labeled as class C

If A is empty or the number of samples left in D is less than A given value

Node is returned as a leaf node, marking node as the class with the most occurrence in D

For each attribute in A

Calculate the information GainRatio

Select the attribute with the largest information gain rate from A as the best partition attribute

After inputting the training set and attribute set of samples, a DT can be constructed according to the C4.5 algorithm that generates the DT.

ID3 algorithm generates DTs based on the information gain value of attributes, while C4.5 algorithm proposes the information gain ratio of attributes [14]. Let's use PlayFootball as an example to create a DT. Assuming that the Day attribute is selected as the training attribute, since all values of Day are discrete and do not have the same value, it will be divided into 14 different classes, and a DT with 14 leaves and depth of 2 will be generated. Equations 1, 2,3 are used to calculate:

$$H(day) = -\frac{5}{14}\log\frac{5}{14} - \frac{9}{14}\log\frac{9}{14} \tag{1}$$

$$E(day) = \sum_{j=1}^{n} \frac{a_{1j} + a_{2j} + ... + a_{nj}}{a} H(a_{1j} + a_{2j} + ... + a_{nj}) = 0 \tag{2}$$

$$Gain(Day) + H(Day) - E(Day) \tag{3}$$

In addition to Day, such attributes are the student number, teacher number, identification number, bank card number, and so on. In conclusion, the information gain ratio of attribute Day is still the largest, so attribute Day is selected as the splitting node, and so on, the DT is created to obtain the classification rules.

In the DT constructed according to C4.5 algorithm, each attribute node in the DT is selected and determined according to the information gain ratio of the attribute, and the value of each node mainly depends on the maximum value of information entropy reduction [15]. Based on this principle, the constructed DT structure is the most intuitive, but not the simplest. It can be seen that C4.5 algorithm not only has the advantages of ID3 algorithm, but also obtains the maximum gain effect by combining with the theory of information gain when attribute value is determined. C4.5 not only retains the advantages of ID3 algorithm, but also has its own advantages.

(1) Advantages

When building DT, the information gain rate of attributes is used to select branch nodes, which effectively overcomes the multi-value bias problem caused by ID3 algorithm in selecting branch nodes. The discretization of continuous data is realized reasonably and effectively through the techniques of equal width method and equal frequency method. In the case of missing attribute values, the data sample set is effectively processed. Pruning is carried out in the process of DT construction, which effectively overfits the whole DT [16-17].

(2) Disadvantages

During the DT construction process, the set of data samples must be scanned and sorted several times, resulting in the low efficiency of C4.5 [18] algorithm. In addition, the algorithm is applied only to sets of data samples that can reside in memory. When the set of training samples is too large for memory, the program cannot be executed.

2.2. The Construction of DT Algorithm Model

(1) Improve C4.5 algorithm

According to the weight of each feature in the final grade, the information gain formula of feature av after introducing the weight is as follows:

$$Gain(D,a_v)=\max_{t\in T_a} Gain(D,a_v,t)=\max_{t\in T_a} Info(D)-\sum_{\lambda\in\{-,+\}}(\frac{|D_t^{\lambda}|}{D}+v)Info(D_t^{\lambda}) \tag{4}$$

According to the weight of each feature in the final grade, the split information entropy formula of feature av after introducing the weight is as follows:

$$SplitInfo_{av}(D)=-\sum_{j=1}^{s}(\frac{|D_j|}{D}+v)\log_2\frac{|D_j|}{D} \tag{5}$$

Where, S is the number of values of feature av, and Dj is the subset of the JTH value of feature A in dataset D.

The pseudo-code of the improved C4.5 algorithm is as follows:

The function createImprovedC4_5Tree(D,A,v)

Create the root node N.

If all the samples in D belong to the same category C then

Mark N as a leaf node of class C; return

end if

If A = ∅ then

N is labeled as a leaf node, and its category is labeled as the class with the largest number of samples in D. return

end if

Perform chooseImprovedBestSplitWay (D, A, v) function, from A to obtain the optimal partition and the characteristics of ai

Good partition point t

According to the best partition point t, the dataset is divided into two parts. Dt- means that the value of D for the feature AI is less than and equal

For the sample subset of t, Dt + denotes the sample subset of D whose value of feature AI is greater than t.

Recursively run createImprovedC4_5Tree(Dt-,A/{ai},v)

Recursively run createImprovedC4_5Tree(Dt+,A/{ai},v)

(2) Model construction

In this paper, Python language and Spyder compiler platform are used as data mining tools to mine and analyze the online learning behavior data and academic performance of students. As the most popular language at present, Python's biggest advantage is its simplicity and ease of use, intuitive syntax and many class libraries related to data mining [16]. Scikit-learn is a Python algorithm library for machine learning, providing common machine learning algorithms. Spyder Library and Scikit-Learn Library, two Python libraries that provide the basis for data mining using Python.

Import the Scikit-Learn library DT algorithm and other related libraries.

The.read_csv() function in Spyder library is called to read student behavior data and academic performance data, and the two types of data are stored in variables X and y respectively.

The initial data set is divided into training set and data set. The classification model is installed with the training set, and the quality of the classification model is tested with the test set to ensure

the accuracy of the model and make the model have a good effect on the new data.

All model objects on Scikit-Learn provide a fit() interface, which is used to build models. For the DT model, FIT (x, y) can be used to train and build the model.

Once the DT model is built, the Predict () interface is called to test the test set. Then the results of the training set and the test set are scored respectively. The interface returns a value of x[I], the closer the value is to 1, the better the performance of the model.

Using Spyder library, the DT model can be output, namely FIG_tree picture.

Randomly select a sample from students' scores for prediction, and output the prediction results to the result table.

3. C4.5 Algorithm Application

3.1. Experimental Environment

The experimental environment of algorithm implementation and model evaluation in this paper is shown in Table 1.

Table 1. Experimental environment for algorithm implementation and model evaluation

Environment type	Attribute
CPU	Intel i5
Memory	16GB
Operating system	Window 10
Algorithm implementation language	Python

3.2. The Data Collection

The research data of this paper come from the educational administration system of our school, and the research object is the achievement data of students majoring in ocean engineering in our school.

4. Analysis of Experimental Results

4.1. Engineering Graphics Course

For the course of "Engineering Graphics", according to the above rules, [81-96] will be marked as "excellent", [71-80] as "good", and [60-70] as "qualified". The prediction performance of this course under the two algorithms is shown in Table 2.

Table 2. Algorithm performance comparison table

	Time(s)	Accuracy rate
C4.5 algorithm	1.207	72.9%
Improved C4.5 algorithm	1.03	81.5%

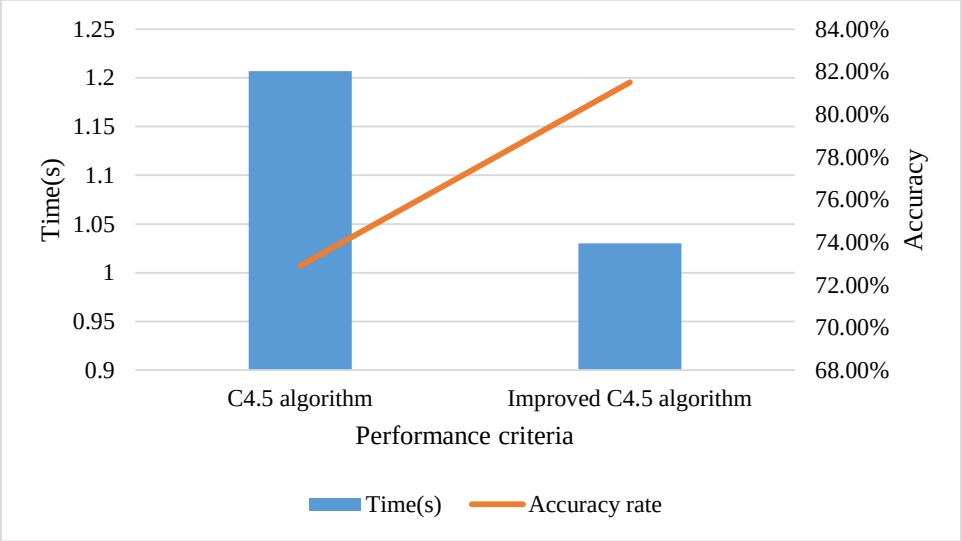

Figure 1. Engineering graphics course algorithm application

As shown in Figure 1, the improved C4.5 algorithm is significantly better than the original C4.5 algorithm in terms of speed and accuracy.

4.2. Introduction to Marine Technology

In order to prove the rationality and repeatability of the algorithm, the experiment was carried out again for the course " Introduction to Ocean Technology " . The prediction performance of the course under the two algorithms is shown in Table 3.

Table 3. Comparison of different algorithms

	Time(s)	Accuracy rate
C4.5 algorithm	1.032	72.3%
Improved C4.5 algorithm	0.915	80.6%

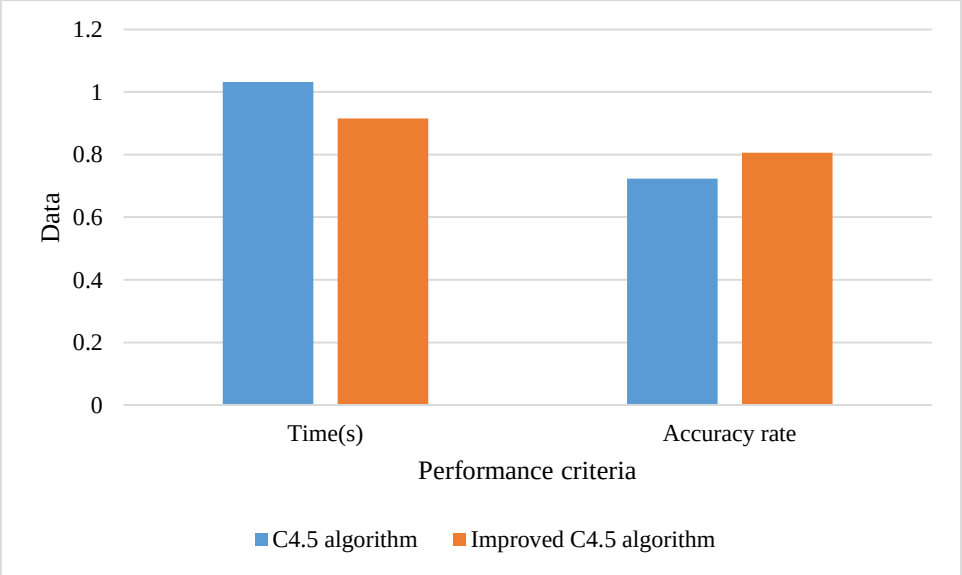

Figure 2. Analysis results of an Introduction to Marine Technology course

As shown in Figure 2, for the course " Introduction to Ocean Technology " , adding balance degree system to the calculation formula can also effectively improve the prediction accuracy of each category.

5. Conclusion

This paper studies the performance analysis model was constructed based on C4.5 algorithm, and the model is applied to implement the course grade of early warning system, can provide real time for teachers to understand the class students learning situation, the course grade prediction results and the function of the crisis warning notice students, thus can guide in advance through the crisis students improve class teaching quality. It can provide students with the function of real-time understanding of their personal learning situation, prediction results of course grades and checking the warning reasons and study suggestions, so that they can avoid failing by strengthening their learning. However, there are still some areas to be improved in the research, among which are as follows: the accuracy of the phased grade prediction model in predicting the final grade interval is not good, so the construction of the phased grade prediction model needs to be optimized in the following work.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Irawan Y . Penerapan Algoritma DT C4.5 Untuk Memprediksi Kelayakan Calon Pendonor Melakukan Donor Darah Dengan Klasifikasi Data Mining. JTIM Jurnal Teknologi Informasi dan Multimedia, 2021, 2(4):181-189. https://doi.org/10.35746/jtim.v2i4.75

[2] Permana B , Ahmad R , Bahtiar H , et al. Classification of diabetes disease using DT algorithm (C4.5). Journal of Physics: Conference Series, 2021, 1869(1):012082 (8pp). https://doi.org/10.1088/1742-6596/1869/1/012082

[3] Sulaiman H . Perbandingan Algoritma DT C4.5 Dan Naive Bayes pada Analisa Tekstur Gray Level Co-Occurrence Matrix Menggunakan Citra Wajah. SISTEMASI, 2021, 10(2):470. https://doi.org/10.32520/stmsi.v10i2.1305

[4] Kustiyahningsih Y , Khotimah B K , Anamisa D R , et al. DT C 4.5 Algorithm for Classification of Poor Family Scholarship Recipients. IOP Conference Series: Materials Science and Engineering, 2021, 1125(1):012048 (7pp). https://doi.org/10.1088/1757-899X/1125/1/012048

[5] Benkercha R , Moulahoum S . Fault detection and diagnosis based on C4.5 DT algorithm for grid connected PV system. Solar Energy, 2018, 173(OCT.):610-634. https://doi.org/10.1016/j.solener.2018.07.089

[6] Murty N , Babu M . A Critical Study of Classification Algorithms for LungCancer Disease

Detection and Diagnosis. International Journal of Computational Intelligence Research, 2017, 13(5(3)):1041-1048.

[7] Sadiq M H , Ahmed N S . Classifying and Predicting Students' Performance using Improved DT C4.5 in Higher Education Institutes. Journal of computer sciences, 2019, 15(9):1291-1306. https://doi.org/10.3844/jcssp.2019.1291.1306

[8] Santoso P , Setiawan R . Penerapan Metode Klasifikasi DT dan Algoritma C4.5 dalam Memprediksi Kriteria Nasabah Kredit Mega Auto Finance. JURIKOM (Jurnal Riset Komputer), 2020, 7(2):200. https://doi.org/10.30865/jurikom.v7i2.1762

[9] Pujianto U , Setiawan A L , Rosyid H A , et al. Comparison of Nave Bayes Algorithm and DT C4.5 for Hospital Readmission Diabetes Patients using HbA1c Measurement. Knowledge Engineering and Data Science, 2019, 2(2):58. https://doi.org/10.17977/um018v2i22019p58-71

[10] Mardiansyah H , Zarlis M , Sitompul O S . Analysis of C4.5 Algorithm of Water Quality Dataset. Journal of Physics: Conference Series, 2021, 1898(1):012002. https://doi.org/10.1088/1742-6596/1898/1/012002

[11] Elacio A A , Lacatan L L , Vinluan A A , et al. Machine Learning Integration of Herzberg's Theory using C4.5 Algorithm. International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(1.1):57-63. https://doi.org/10.30534/ijatcse/2020/1191.12020

[12] Thohari A H , Anita W S . Smart dunning to improve collection ratio in internet service provider using C4.5 algorithm. Journal of Physics: Conference Series, 2020, 1450(1):012062 (5pp). https://doi.org/10.1088/1742-6596/1450/1/012062

[13] Primartha R , Tama B A , Arliansyah A , et al. DT combined with PSO-based feature selection for sentiment analysis. Journal of Physics: Conference Series, 2019, 1196(1):012018 (6pp). https://doi.org/10.1088/1742-6596/1196/1/012018

[14] Sugiatna E , Ibrahim A M , Hadi I A . Implementasi Algoritma Klasifikasi C4.5 Untuk Memprediksi Kelayakan Pembelian Kendaraan. JTIM Jurnal Teknologi Informasi dan Multimedia, 2019, 1(2):124-132. https://doi.org/10.35746/jtim.v1i2.26

[15] Wajhillah R , Yulianti I . Penerapan Algoritma C4.5 Untuk Prediksi Penggunaan Jenis Kontrasepsi Berbasis Web. Klik - Kumpulan Jurnal Ilmu Komputer, 2017, 4(2):160. https://doi.org/10.20527/klik.v4i2.98

[16] Wahyudi M , Andriani A . Application of C4.5 and Nave Bayes Algorithm for Detection of Potential Increased Case Fatality Rate Diarrhea. Journal of Physics: Conference Series, 2021, 1830(1):012016 (12pp). https://doi.org/10.1088/1742-6596/1830/1/012016

[17] Rathinasamy R , Raj L . Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data. International Journal of Innovative Research in Computer and Communication Engineering, 2019, 5(1):2017.

[18] Tozzi A E , Gesualdo F , Rizzo C , et al. A data driven clinical algorithm for differential diagnosis of pertussis and other respiratory infections in infants. PLoS ONE, 2020, 15(7):e0236041. https://doi.org/10.1371/journal.pone.0236041