Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2021.020403
ISSN 2790-0983 Vol. 2, Issue 4: 18-25

Application and Optimization of Decision Tree Algorithm in Classification Prediction

Jiwei Zhang*

Gansu Industry Polytechnic College, Gansu, China

635479027@qq.com

**corresponding author*

Keywords: Decision Tree Algorithm, Classification Prediction, ID3 Algorithm, Learning Achievement

Abstract: In online teaching situations, a decision tree (DT) analysis algorithm is proposed in order to enable teachers to track students' learning behaviors and learning status in a timely manner. With the help of this algorithm, teachers can use the existing learning record files to analyze students' learning behaviors through multiple observation dimensions. In this paper, by constructing a learning achievement classification prediction (CP) model based on ID3 decision tree algorithm (DTA), the optimized ID3 DTA is used to analyze students' learning achievement and predict students who may not reach the expected teaching goals, so that teachers can give appropriate teaching assistance in time to achieve teaching goals. Through experiments, it is proved that with the increase of data volume, the accuracy of the model for CP of learning achievement is stable above 80%, which realizes the function of ID3 algorithm achievement prediction.

1. Introduction

Using classical data mining (DM) algorithms to mine existing student databases, it is possible to obtain correlation rules between grades in various subjects, as well as to predict students' graduation assessment scores based on their entrance grades. This is of great importance to the progress of students in school, to the teachers' targeted teaching arrangements, and to the school administration's enrollment planning.

There are many scholars and experts in China who have conducted research on the diagnosis and prediction of student performance (SP), and there are many references in the literature. For example, some scholars have used two DM methods, GA and DT, to build models and analyze the potential relationship between behavior and learning, and their preliminary classification using decision trees [1]. Some scholars used SVM, logistic regression, and DT C5.0, to predict SP, and found that the

Copyright: © 2021 by the authors. This is an Open Access article distributed under the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (https://creativecommons.org/licenses/by/4.0/).

best prediction method (PM) was SVM [2]. Some scholars used various DM methods and optimized the DM algorithm using GA based on previous experience to come up with an optimal SP prediction model, where the best PM is class neural network [3].. Some researchers have compared and analyzed the results of GBDT, Xgboost, and random forest models for student achievement prediction, and the Xgboost model has a stronger learning ability compared to the random forest and GBDT models for the extracted student behavior features, and the final trained prediction model has a relatively higher accuracy and precision, and therefore the reliability of the prediction is relatively higher [4]. Thus, there are various algorithms for predicting SP, and all of them have good prediction results.

In this paper, we first introduce the DT classification algorithm, then propose a tool for constructing ID3 DT classification model, then propose ID3 DT model with learning achievement CP, elaborate the principle of ID3 algorithm optimization, and compare the efficiency and accuracy of ID3 algorithm and optimization algorithm, and finally analyze the accuracy of DTA for learning achievement CP through experiments.

2. DT Related Algorithms and Construction Tools

2.1. DT Classification Algorithm

DT classification has been very popular since its inception, mainly because the knowledge obtained by DT classification is intuitive and easily understood in a tree representation, whereas many other classification techniques, such as ANN classification and SVM classification, are much less understandable in terms of the final result [5]. DT classification rule acquisition involves two operators, one is the data mining engineer and the other is the DT classification algorithm. DT classification rule acquisition mainly involves training data import, algorithm operation, and result display and analysis. While DT CP mainly involves import of data to be predicted, import of decision rules, and display and analysis of prediction results [6].

DTAs are usually expressed in terms of information gain, which is the difference of information entropy, as in Equation (1).

$$Gain(F,A) = Entropy(F) - \sum_{v \in V(A)} \frac{|Fv|}{|F|} Entropy(Fv) \tag{1}$$

v values constitute the value domain of attribute A, |Fv| is the number of Fv, |F| is the number of F; Entropy(F) is the information entropy of sample F, calculated by Equation (2).

$$Entorpy(F) = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{2}$$

Entropy (Fv) is the information entropy of sample Fv, and Pi is the probability of occurrence of category i.

2.2. Tools for Building the ID3 DT Classification Model

The core element of DM is modeling, that is, to determine what algorithm to choose to build the DT. In this paper, Matlab tools are used.Matlab algorithms are easier to learn and program, and the system is very extensive in terms of simulation applications; the pages developed in Matlab are user-friendly and non-professionals can get started quickly [7-8].

3. ID3 DT Based Learning Achievement CP

3.1. Research Architecture

In this study, the data is analyzed by using the basic data of the students from the database of study record files that have been pre-processed, and the various attributes of the students' study records on the website as analysis variables, which include nominal scales (gender, education, etc.) and isometric scales (number of classes, number of online discussions, etc.) [9-10]. First, the various implications of these analytic variables for different courses, students with different learning outcomes, and the potential relationships between each analytic variable were identified by means of categorical analysis in data mining. Then, DT software is used to further identify decision rules(DR) between various attributes of students' learning behaviors from learning records, and to verify whether these DR can be applied across semesters using historical data and existing data. Finally, the instructor verifies whether these decision rules do reflect the instructor's instructional style for the course to assess the usefulness of these decision rules as supporting information for the instructor's instructional strategies and course design when teaching online [11-12].

3.2. Data Preparation

The main purpose of this phase is to create a database of learning records with different observation dimensions such as courses and students so that the immediate learning of students in each course can be quickly probed during the data analysis phase, and also different dimensions can be used to analyze student learning behaviors and types of learning, and then the relationship between learning behaviors and learning performance can be analyzed using DTs [13]. In the processing procedure, it is first necessary to identify the data source and the learning behavior variables to be pre-analyzed to ensure that all the data needed for data processing can be obtained correctly. The data sources studied combined three components: basic student data and grade files obtained from the information management system, files of all learning action records from the website records, and data files with course-related information [14]. The data processing flow of the ID3 algorithm is shown in Figure 1.

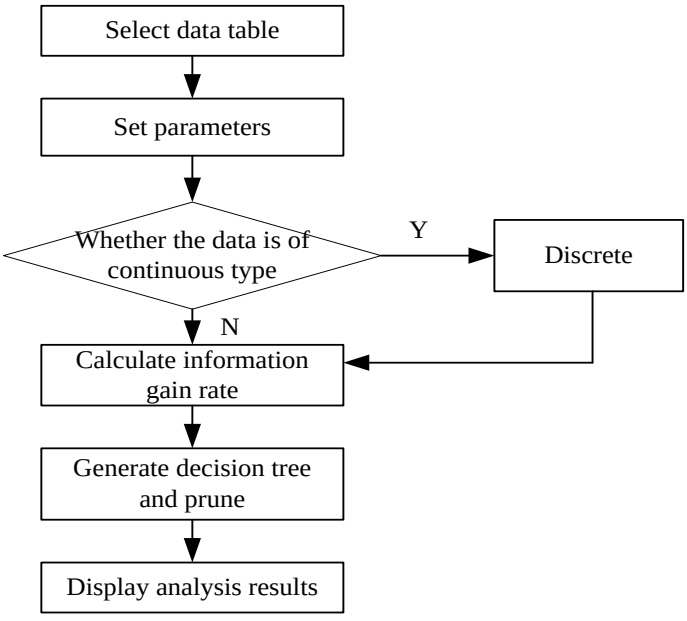

Figure 1. Data processing of ID3 algorithm

3.3. Generation, Analysis and Collation of DT Classification

Although most of the attribute values of the studied analysis variables are not pre-classified and may present difficulties in how to effectively classify them, this problem has been solved by a considerable amount of research, for example, by using currently available DT analysis software, which can directly handle continuous data to solve the problem of unclassified attribute values [15]. On the other hand, since students' academic performance is not pre-classified and there are no studies or algorithms to improve the classification of students' academic performance levels, this study uses different ways of classifying performance in order to find the most suitable level classification.

After pre-processing the data and processing them in a continuous and temporal manner, the DT software was used to classify the students in each course and generate DTs and decision rules [16]. The generated decision rules describe the association and if then relationship between the overall learning behavior and learning outcomes of the students taking the course, and for the teachers, they can know which learning patterns the overall students have for the course, and further infer which learning patterns the students in the high or low scoring groups may have, and this supporting decision information can be used as a reference for the teachers to implement future teaching The decision information can be used as a reference for teachers to implement teaching strategies and teaching aids in the future [17-18].

4. Research on DTA Optimization and CP Applications

4.1. DTA Optimization Research

(1) Optimization principle

DT classification algorithms have high requirements for data sets, and most algorithms require accurate samples without missing to ensure algorithm performance, but missing data are common in practical applications, and improper handling can accumulate a large number of errors, thus requiring data analysts to use reasonable methods to handle these vacant data to reduce the algorithm's computing time and complexity [19]. Since the ID3 algorithm tends to select attributes that take more values as test attributes, it cannot reasonably distinguish sample information, resulting in inefficient algorithms and low accuracy rates. This paper adopts a new method to process each attribute, converting all non-binary attributes into binary attributes to avoid the problem of bias caused by attribute values, and using information gain to select attributes to classify the data set on this basis, which avoids the uneven attribute taking values, and the final generated DT after processing is a standard binary tree, and the formed classification rules are concise and easy to easy to understand [20].

(2) Analysis of the algorithm before and after optimization

Efficiency analysis: the improved algorithm needs to process the attribute values first in the process of constructing the DT model, and thus needs to scan and sort the data set several times. Meanwhile, most of the ID3 algorithm attribute values are floating-point (16 bytes), integer (8-byte number), string (2 bytes) and other data types, which take up more storage space, while the type of attribute values after the improved algorithm is Boolean (1 byte), which has less space overhead in comparison, and thus the efficiency of the improved algorithm is higher.

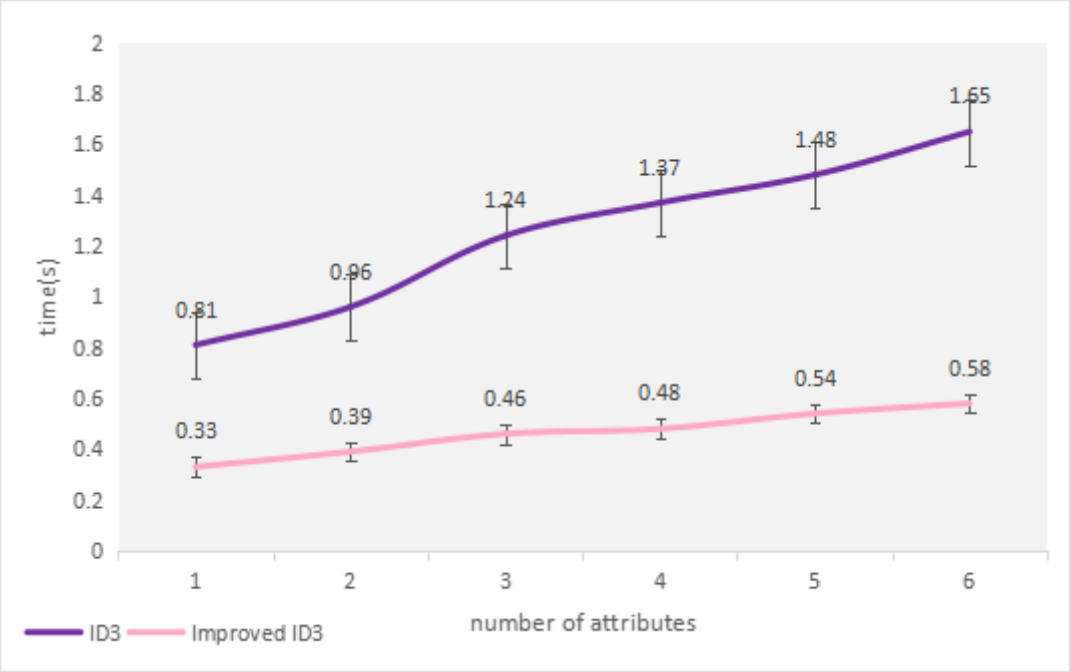

Figure 2. Experimental results of algorithm efficiency

Figure 2 shows the efficiency of the algorithm with increasing number of attributes, from which it can be seen that the execution time of both algorithms increases as the number of(TNO) attributes increases. When TNO attributes is 6, the execution time of ID3 algorithm is 1.65s and the execution time of improved ID3 algorithm is 0.58s. It shows that the execution time of improved ID3 algorithm is less and shows better robustness in terms of time efficiency of the algorithm.

(2) Accuracy analysis

Accuracy definition: Different data are selected as the test set(TS) and training set, the tested attributes are set, and the tested attribute column in the test set is set to empty, and the algorithm is called to output the results. The accuracy rate of the algorithm is expressed as R. Then its specific calculation formula is :

$$R = Q / S \tag{3}$$

In the above equation, S denotes the total number of(TTNO) data items(DI) in the TS, and Q denotes TTNO DI with the same values before and after the test attributes in the test set.

So as to illustrate the advantages and disadvantages of the two algorithms in terms of test data accuracy before and after improvement, this paper compares the accuracy of the two algorithms on three test data sets, A, B and C, each with equal amount of data, and yields the results shown in Table 1.

Table 1. Accuracy rates of algorithms before and after improvement

	A	B	C
ID3	73%	78%	77%
Optimization algorithm	84%	89%	92%

Comparing the accuracy rates of testing A, B and C data before and after the algorithm improvement in Table 1, it shows that the improved ID3 algorithm has a higher accuracy rate.

4.2. CP of Student Performance by DTA

CP accuracy test: This experiment analyzed the data on the classification accuracy of the model with increasing amount of data. This was done by (1) randomly sampling and controlling the amount of sample data, which were input to the learning achievement CP model, respectively; (2) performing statistics based on classification attributes; and (3) comparing with the original data samples and calculating the accuracy by category, as detailed in Table 2.

Table 2. Classification accuracy of the achievement prediction model

Data Tuple	Accuracy	Category Ⅰ	Category Ⅱ	Category Ⅰ Ⅲ
500	42%	45%	48%	55%
1000	49%	48%	53%	62%
1500	63%	59%	67%	78%
2500	74%	72%	78%	81%
3000	81%	80%	82%	80%
4000	83%	82%	81%	84%
5000	85%	85%	84%	86%

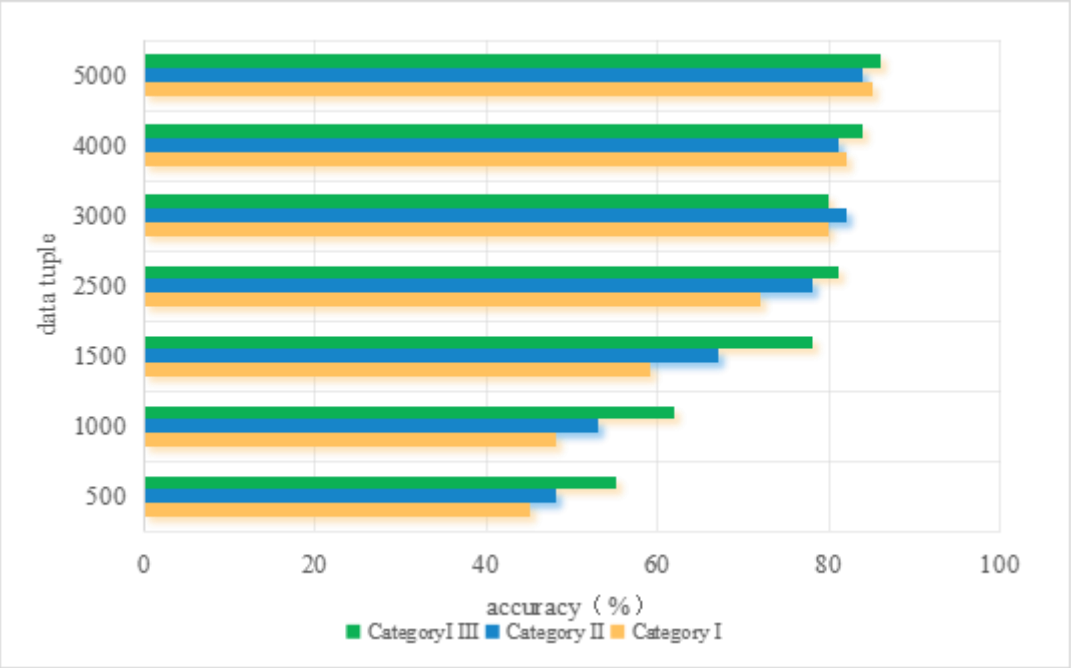

Figure 3. Accuracy rates on different categories

Figure 3 shows that the prediction agreement rate gradually leveled off with the increase of data volume, and the CP accuracy of the model was stable above 80% on categories I, II, and III, which indicates that the CP accuracy of the model is high and relatively stable.

5. Conclusion

To improve the effect of intelligent teaching and increase the effectiveness of student management, with various existing management systems in colleges, how to integrate the resources of each teaching system, study the intrinsic important connections existing between various resource data, analyze the important information within the data using data mining technology, and

continuously figure out the necessity of discovering the information construction of colleges and universities to improve the teaching quality becomes an urgent need for colleges and universities. This paper takes student achievement as the research object, by integrating student information base and achievement data in teaching system as resources, and using improved ID3 algorithm to mine achievement data to achieve CP of student achievement, and also verifies the applicability of DTA in CP of learning achievement.

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] Alejandro Barredo Arrieta, Sergio Gil-Lopez, lbai Lata, Miren Nekane Bilbao, Javier Del Ser: On the post-hoc explainability of deep echo state networks for time series forecasting, image and video classification. Neural Comput. Appl. 34(13):10257-10277 (2021). https://doi.org/10.1007/s00521-021-06359-y

[2] Everton Castelao Tetila, Bruno Brandoli Machado, Jose F. Rorigues Jr, Diego A. Zanoni, Nicolas Alessandro de Souza Belete, Thayliny Zardo, Michel Constantino, Hemerson Pistori: Associative classification model for forecasting stock market trends. Int. J. Bus. Intell. Data Min. 19(1): 97-112 (2021). https://doi.org/10.1504/IJBIDM.2021.115968

[3] Mohammad Masdari, Afsane Khoshnevis: A survey and classification of the workload forecasting methods in cloud computing. Clust. Comput. 23(4): 2399-2424 (2020). https://doi.org/10.1007/s10586-019-03010-3

[4] Joy Dhar, Asoke Kumar Jodder: An Effective Recommendation System to Forecast the Best Educational Program Using Machine Learning Classification Algorithms. Ingenierie des Systemes d Inf. 25(5):559-568 (2020). https://doi.org/10.18280/isi.250502

[5] Asmaa H. Rabie, Shereen H. Ali, Ahmed 1. Saleh, Hesham A. Ali: A fog based load forecasting strategy based on multi-ensemble classification for smart grids. J. Ambient Intell. Humaniz. Comput. 11(1): 209-236 (2020). https://doi.org/10.1007/s12652-019-01299-x

[6] Ranjeeta Bisoi, P. K. Dash, Pragyan P. Das: Short-term electricity price forecasting and classification in smart grids using optimized multikernel extreme learning machine. Neural Comput. Appl. 32(5): 1457-1480 (2020). https://doi.org/10.1007/s00521-018-3652-5

[7] Mohamed Amgad, Lamees A. Atteya, Hagar Hussein, Kareem Hosny Mohammed, Ehab Hafiz , Maha A. T. EIsebaie, Pooya Mobadersany, David Manthey, David A. Gutman, Habiba Elfandy, Lee A. D. Cooper: Explainable nucleus classification using DT Approximation of Learned Embeddings. Bioinform. 38(2): 513-519 (2021). https://doi.org/10.1093/bioinformatics/btab670

[8] Pravin S. Game, Vinod Vaze, Emmanuel M: Optimized DT rules using divergence based grey wolf optimization for big data classification in health care. Evol. Intell. 15(2): 971-987 (2021).

[9] Chandrashekhar Azad, Bharat Bhushan, Rohit Sharma, Achyut Shankar, Krishna Kant Singh, Aditya Khamparia: Prediction model using SMOTE, genetic algorithm and DT (PMSGD) for classification of diabetes mellitus. Multim. Syst. 28(4): 1289-1307 (2021). https://doi.org/10.1007/s00530-021-00817-2

[10] Mohebbanaaz, L. V. Rajani Kumari, Y. Padma Sai: Classification of ECG beats using optimized DT and adaptive boosted optimized DT. Signal Image Video Process. 16(3): 695-703 (2021). https://doi.org/10.1007/s11760-021-02009-x

[11] Pietro Ducange, Francesco Marcelloni, Riccardo Pecori: Fuzzy Hoeffding DT for Data Stream Classification. Int. J. Comput. Intell. Syst. 14(1): 946-964 (2021). https://doi.org/10.2991/ijcis.d.210212.001

[12] Hong Nguyen Thi Khanh: Classification of Concepts Using DTs for Inconsistent Knowledge Systems Based on Bisimulation. J. Inf. Hiding Multim. Signal Process. 12(1): 22-30 (2021).

[13] Vinay Arora, Rohan Singh Leekha, Inderveer Chana: An Efficacy of Spectral Features with Boosted DTA for Automatic Heart Sound Classification. J. Medical Imaging Health Informatics 11(2): 513-528 (2021). https://doi.org/10.1166/jmihi.2021.3287

[14] Kholoud Maswadi, Norjihan Abdul Ghani, Suraya Hamid, Muhammad Babar Rasheed: Human activity classification using DT and Naive Bayes classifiers. Multim. Tools Appl. 80(14): 21709-21726 (2020). https://doi.org/10.1007/s11042-020-10447-x

[15] Yousef Ali Khan, Q. S. Shan, Qianning Liu, Syed Zaheer Abbas: A nonparametric copula-based DT for two random variables using MIC as a classification index. Soft Comput.25 (15): 9677-9692 (2020). https://doi.org/10.1007/s00500-020-05399-1

[16] Yashesh D. Dhebar, Kalyanmoy Deb: Interpretable Rule Discovery through Bilevel Optimization of Split-Rules of Nonlinear DTs for Classification Problems. IEEE Trans. Cybern. 51(11): 5573-5584 (2020). https://doi.org/10.1109/TCYB.2020.3033003

[17] Jugal Patel, Jeffrey Katan, Liliana Perez, Raja Sengupta: Transferring decision boundaries onto a geographic space: Agent rules extracted from movement data using classification trees. Trans. GIS 25(3): 1176-1192 (2021). https://doi.org/10.1111/tgis.12770

[18] B. Durgalakshmi, V. Vijayakumar: Feature selection and classification using support vector machine and DT.Comput. Intell. 36(4): 1480-1492 (2020). https://doi.org/10.1111/coin.12280

[19] Mahdi Abbasi, Aazad Shokrollahi: Enhancing the performance of DT-based packet classification algorithms using CPU cluster. Clust. Comput. 23(4): 3203-3219 (2020). https://doi.org/10.1007/s10586-020-03081-7

[20] Stathis Malakis, Panagiotis Psaros, Tom Kontogiannis, Christina Malaki: Classification of air traffic control scenarios using DTs: insights from a field study in terminal approach radar environment. Cogn. Technol. Work. 22(1): 159-179 (2019). https://doi.org/10.1007/s10111-019-00562-7