Journal of Cases on Information Technology
Volume 24 • Issue 5

Online Learning Behavior Feature Mining Method Based on Decision Tree

Juxin Shao, Basic Teaching Department, Yantai Institute of Technology, China*
Qian Gao, Basic Teaching Department, Yantai Institute of Technology, China
Hui Wang, Yantai Branch of China Broadcasting Network Corporation Ltd., China

ABSTRACT

This research mainly discusses the design of an online learning behavior feature mining method based on decision tree. Data collection is the real-time collection of online learning behavior data from distance learning websites. OWC (office web component) technology is used to draw real-time charts on the page. Online learning students are selected as the research object, and the student's system log data and questionnaire data are selected. When combining the pre-pruning method and the post-pruning method to make decisions after the tree is pruned, the same source data is used to adjust, test, and evaluate the decision tree model. The evaluation process to generate a complete decision tree is completed by the c4.5tree algorithm in C4.5, which can be named with a suffix of .names. The type definition file is used to record the type of each attribute item or the range of possible values. In the study, the prediction accuracy rate of predicting learning effect based on "online learning behavior" reached more than 66%.

KEYWORDS:

decision tree, feature mining, learning behavior, learning planning, online learning

1. INTRODUCTION

With the development of information technology, many industries (institutions) have accumulated a lot of data. These data often contain a lot of useful information. It is difficult to obtain this effective information only through statistical analysis or database retrieval. In recent years, data mining technology has penetrated into various fields, digging out hidden knowledge or patterns from data in different fields. At present, it has been successfully applied in the fields of finance, biology and so on.

The school opens its own teaching resources, teacher strength, education time, etc. to students. Through distance education, students can arrange their own learning time according to their actual conditions. At the same time, learning is a process of continuous updating and open learning Resources are conducive to students' access to the latest knowledge and information. Online education generally does not set thresholds. Anyone who has learning needs can obtain the knowledge they need through online education, which provides conditions for the realization of lifelong learning for the whole people.

Fraudulent consumption of large amounts of electricity may balance the gap between supply and demand to a large extent. Therefore, it is necessary to develop a solution that can accurately detect these thefts in complex power networks. Therefore, in response to these problems, Jindal A proposed a comprehensive top-down scheme based on decision trees (DT) and support vector machines

DOI: 10.4018/JCIT.295244 *Corresponding Author

(SVM). Different from existing solutions, this solution has sufficient capabilities to accurately detect and locate real-time power theft at each level of power transmission and distribution (T&D). The scheme he proposed is based on the combination of DT and SVM classifiers, which can perform rigorous analysis on the collected electricity consumption data. Although his research can regard the proposed scheme as a two-level data processing and analysis method because the data processed by DT is provided as input to the SVM classifier, the research process does not have too much data interpretation (Jindal et al., 2016). Chang T C proposed a multi-standard group decision model based on subjectivity, imprecision and ambiguity in the evaluation of university teachers. First of all, selecting appropriate evaluation indicators according to the evaluation objectives, showing a clear structural relationship between the evaluation indicators and the objectives, and establishing an appropriate evaluation system are all essential and basic tasks. Then, collect expert evaluation data and process data, establish a training set to build a decision tree, extract evaluation rules, simplify the evaluation process, and reduce the evaluation cost in practical applications. Third, establish an interval cloud evaluation matrix through the decision cloud, transform the evaluation value through the cloud model, determine the order of importance of the decision-making process, and make a decision. Although he proposed a method to solve the problem of language decision-making for the evaluation of university teachers to illustrate the effectiveness of the model, the sample data in the research process was too few (Chang & Wang, 2016). Hen Da Lianpour A believes that the decision tree (DT) model is the most commonly used tool for investigating CRM and providing appropriate support for the implementation of the CRM system. He calculated three decision models for SMEs and analyzed different decision tree methods (C&RT, C4.5 and ID3). These methods are then used to calculate the ME and VoE of the model, and then they are used to calculate the mean error (ME) and error variance (VoE) estimates to study the predictive power of these methods. His decision tree method is used to analyze small and medium enterprise (SME) data sets. Although he put forward strong technical support to better guide market trends and mining in CRM, no sample comparison was made during the research process (Hen Da Lianpour et al., 2016). Lu L believes that due to the lack of effective technology for extracting PML for large geographic areas, little is known about the temporal and spatial distribution of PML. He proposed a decision tree classifier to extract transparent PML information from Landsat-5 TM images. The classifier is constructed based on the rules obtained by analyzing the spectral characteristics of the transparent PML on the Landsat-5 TM image, and covers the research area of Xinjiang, the largest cotton-growing province based on PML in China. Then, the classifier was applied to the study area in 1998, 2007 and 2011. The results show that the classifier successfully extracted PML from Landsat-5 TM images with total accuracy of 97.82%, 85.27%, and 95.00%, and the Kappa coefficients were 0.9782, 0.80, and 0.93 in 2011, 2007 and 1998, respectively. Although the decision tree classifier in his research is stable over time, it does not mean that it can be applied in different years (Lu et al., 2017).

Data collection is the real-time collection of online learning behavior data from distance learning websites. When a student logs into the remote teaching website, a Session will be started, the basic information of the login will be recorded, and the Session ID will be generated. In order to make the statistical results intuitively reflect the learning status of students, OWC (Office Web Component) technology is used to draw real-time charts on the page. The learning behavior evaluation module mainly conducts data mining on the historical records of students' learning process, constructs the relationship between learning behavior and learning effect, evaluates students' academic performance according to the relationship model and gives a learning plan. In the study, online learning students are selected as the research object, and the student's system log data and questionnaire data are selected. This data set is used as the input training sample set of the decision tree generation algorithm. When combining the pre-pruning method and the post-pruning method to make decisions after the tree is pruned, the same source data is used to adjust, test and evaluate the decision tree model.

2. ONLINE LEARNING BEHAVIOR CHARACTERISTICS

2.1 Online Learning

Online learning refers to a kind of learning activity carried out through a computer network, allowing learners to freely choose. Learning content, learning location and learning time (Bilal et al., 2016; Tsangaratos & Ilia, 2016). Compared with traditional learning activities, online learning has the following three characteristics: First, rich and shared networked learning resources (Tanha et al., 2017). The second is the main form of independent learning and collaborative learning of individuals. The third is to break through the time and space limitations of traditional learning. Therefore, online learning is the process and method of interaction between learners and learning resources, teachers and other basic elements in the network environment. It is a process from conformity to identification and then to internalization. It is behavioral learning, value learning and norms. Unity of learning. The advantage of online learning is that it can happen at any time and any place, has great flexibility, diversified teaching and learning methods, and can provide learners with abundant resources, but the process and management of online learning teaching and learning is difficult control, learners are easy to get lost due to lack of autonomy. Because of the particularity and complexity of online learning activities and behaviors, the process of collecting and analyzing them is very difficult to complete. It needs to be analyzed and processed from the following aspects (Berthon et al., 2016; Goodman et al., 2016). Average number of readings:

$$\mathrm{AvgRead} = \frac{\mathrm{ReadCount}}{\mathrm{BlogCount}} \tag{1}$$

Average number of updates:

$$\mathrm{AvgUpdate} = \frac{\mathrm{UpdateCount}}{\mathrm{BlogCount}} \tag{2}$$

The study time span is $\mathrm{TimeSpan}$ (Mistry et al., 2016).

$$\mathrm{TimeSpan} = \frac{\mathrm{EndTime}_{(\max)}\text{-}\mathrm{BeginTime}_{(\min)}}{\mathrm{SysTime}} \tag{3}$$

The learner's normal log-in and exit rate $\mathrm{QuitRate}$ is the ratio of the total number of normal log-outs after the learner logs into the system to the total number of learners log-in to the system for learning (Pham et al., 2017; Wang et al., 2017).

$$\mathrm{QuitRate} = \frac{\mathrm{Quit}}{\mathrm{Login}} \times 100\% \tag{4}$$

The total number of learner's notes Blog :

$$\mathrm{Blog} = 0.7 \times \frac{\mathrm{ReadCount}}{\mathrm{BlogCount}} + 0.3 \times \frac{\mathrm{UpdateCount}}{\mathrm{BlogCount}} \tag{5}$$

2.2 Online Learning Mechanism

The online learning mechanism refers to the correlation between various learning elements in the online learning system and subsystems and the operation mode, so the research on the online learning mechanism is at the height of the system (Hameed et al., 2016). Structuralism believes that: any structure is composed of elements and their relationships. In the traditional environment, the elements of learning mainly come from the following four aspects: learners, teachers, content and learning environment, teachers and learners belong to online learning human factor. Similarly, in the online learning environment, the elements of online learning also include these four aspects, learners, teachers, online courses, and online learning environment. The online learning mechanism is to study the four elements and the relationship between them in the online environment. But compared with the traditional environment, the mode of action of each factor is very different (Hen Da Lianpour et al., 2016; Sanz et al., 2017). The learning mechanism in the network environment mainly studies the learning of students. It is particularly important to study the learning behavior of learners, that is, to study the characteristics of learners, the activities and behaviors of learners to conduct online learning, and to establish a guarantee on the basis of research. Learning methods and strategies to guide the effective development of teaching and learning under the network environment (Andrew, 2018).

The network provides a powerful and integrated information medium, and the collection and acquisition of information resources is an important form of online learning activities. In the network environment, learners not only learn the content of online courses, but also learn auxiliary resources and resources on the Internet, web browsing and information retrieval and other behaviors (Di & Xu, 2019; Shalabi & Hadjisophocleous, 2020). Information processing is the process of understanding and absorbing information, including perception, attention, memory, comprehension and other processes. It can be completed with the help of information downloading, classified storage, and blog study notes, so that learners can use information creatively and effectively. The release of information includes actions such as asking questions in the BBS discussion area, answering questions, turning in homework, and showing study notes (Petrich, 2020). Information exchange includes the communication between the learner and the teacher and the exchange between the learner and the learner. The information dissemination and sharing are completed through BBS, Blog study notes, message, telephone, email and other methods. Information application is to use the learned information to solve practical problems, such as completing informatization teaching design and making multimedia courseware. The number of times each student submitted R correctly (Roy et al., 2019).

$$R = \frac{A}{S} \times 100\% \tag{6}$$

Among them, A is the number of correct submissions. The average value of the time interval between the first submission time of the job and the start of the job T_{m} (Nourani et al., 2019).

$$T_{\mathrm{m}} = \frac{\sum_{i}^{m} yF_{s}}{\mathrm{m}} \tag{7}$$

2.3 WEB Data Mining

There are many definitions of data mining, and there are also many discussions about what data mining is and not (Kad et al., 2019). The broad definition of data mining period includes traditional statistical methods; the narrow definition emphasizes automatic and heuristic methods. All data mining processes are targeted at specific applications in a certain business field. Therefore, before starting data

mining, First, determine the application purpose of data mining. The factors that need to be considered include the following: First, the functional and performance problems currently encountered in this field; Second, which tasks in this field can be done by computers instead of manual work: Third is the function and performance indicators of the system; Fourth is the data mining the simplicity, accuracy and comprehensibility of the model. The loss function is used to measure the degree of inconsistency between the predicted value of the model and the actual value (Subbotin, 2020).

$$S\left(\mathrm{x}, f(x)\right) = \left(\mathrm{x} - f(x)\right)^2 \tag{8}$$

The modeled function is:

$$D = \sum_{\mathrm{i}=1}^{n} l\left(\mathrm{x} - f(x)\right)^2 + \lambda f(x) \tag{9}$$

Among them, l represents the loss function. Information entropy (information entropy) is one of the most commonly used indicators to detect the purity of a sample set (Sheng, 2019).

$$\mathrm{Ent(d)} = \sum_{\mathrm{k}=1}^{\mathrm{m}} \mathrm{plnP_m} \tag{10}$$

P_{m} is the proportion of the m-th sample in the sample set d (Niu & Chen, 2019). The information gain is:

$$G\left(D, A\right) = \mathrm{Ent}\left(D\right) - \sum_{b=1}^{b} \mathrm{Ent}\left(D\right)^b \tag{11}$$

2.4 Decision Tree

Decision tree algorithm is a process of recursive realization. The recursive termination condition is that the program traverses the attributes of all divided data sets or all instances under each branch have the same classification. One disadvantage of decision trees is that they are prone to overfitting. The resulting performance problem is that the training set can be classified correctly, but the performance of the test set is very poor, which is reflected in the algorithm that has too many branches. In order to overcome this shortcoming, the decision tree introduced the concept of pruning. Pruning is divided into pre-pruning and post-pruning to see whether dividing nodes can improve the generalization performance of the algorithm. If not, the node is a leaf node, and the post-pruning is a spanning tree, check each node. If the node is removed, the generalization performance of the algorithm can be improved, then the node is set as a leaf node (Kg et al., 2019). The C4.5 algorithm proposes an evaluation index for optimal attribute selection, namely the gain rate $\mathrm{GainR(D,a)}$.

$$\mathrm{GainR(D,a)} = \frac{\mathrm{Gain(D,a)}}{\mathrm{V(D,a)}} \tag{12}$$

For the C4.5 decision tree, the "Gini index" is used to select attributes (Gudelis et al., 2019). The purity of data set D can be measured by the Gini value:

$$GINI = \sum_{K=1}^{Y} \sum_{K=1}^{Y_2} P_1 P_2 \tag{13}$$

The smaller the Gini value, the higher the purity of the data set D (Jin et al., 2019). The Gini index is defined as follows:

$$GINI\_\text{index} = \sum_{i=1}^{n} K \frac{D_2}{D_1} GINI + \sum_{i=1}^{n} \text{p} \frac{D_2}{D_1} GINI \tag{14}$$

Decision tree algorithm has many advantages, such as high accuracy, strong interpretability, not enough sensitivity to missing values, outliers, etc. It also has its own shortcomings. For example, for continuous variables, discretization is needed, which is prone to overfitting. It is these advantages and disadvantages that led to the development of ensemble learning methods based on decision trees. Ensemble learning is very effective in solving the phenomenon of over-fitting, and it also improves the accuracy of the algorithm (Kumar & Ghosh, 2019).

3. ONLINE LEARNING BEHAVIOR FEATURE MINING EXPERIMENT

3.1 Data Acquisition Module

The data collection module is responsible for real-time collection and quantification of student online learning behavior data from the remote teaching website, and store it in the database. When a student logs into the remote teaching website, a Session will be started, the basic information of the login will be recorded, and the Session ID will be generated. In the process of choosing video course learning, text data learning, and real-time Q&A learning, a record is generated in the data table S_study, and the session ID, student ID, course ID, learning column code and learning start of the visit are generated. The time is stored in the corresponding field of the record. When the student finishes the study of this column, record the time when the study ends and calculate the duration of this study, and store it in the corresponding field of the record.

The learning start time $svalue and the learning end time $evalue are recorded in the database in the form of UNIX timestamps. The length of the learning time $time is represented by minutes. For the collection of $svalue, when you enter a learning section page of a certain course, call the time() function to get the time stamp at that time, assign it to $svalue and record it in the database; at the same time, call the "time processing" page. This page learns and stays time to collect. It is checked every 60 seconds. If the dwell time exceeds 30 seconds, $sTime is increased by 1 (that is, the learning time is increased by 1 minute), and 30 seconds is the minimum error point.

3.2 Data Statistics Module

The main function of the data statistics module is to provide a graphical interface in the form of a Web page to realize real-time query, statistics and analysis of student online learning behavior data. In order to make the statistical results intuitively reflect the learning status of students, OWC (Office Web Component) technology is used to draw real-time charts on the page.

Course visit statistics are divided into overall statistics and detailed statistics. The overall statistics list the basic information of all courses visited, including the type of course, the number of people studying, the total study time, etc. Detailed statistics are specific to a specific course, recording which students have studied the course, which columns have been studied, how much time each has studied, and the status of participating in BBS discussions and online tests.

According to the data obtained from course access statistics, teachers provide learning resources that meet the needs of students according to the rules and characteristics of students' online course learning, and design reasonable teaching content and activities. The teaching administrator can also evaluate the teaching quality of the teacher in charge of the course based on this. The statistics of course visits are shown in Table 1.

Table 1. Course access statistics

Course number	Course Title	Course category	Number of people studying online	Total study time (minutes)
2934	Photoshop application	Elective	5	140
479	Financial Management (A)	Limited selection	200	235
5402	College Chinese	Elective	700	6100
1088	Fundamentals of Economic Mathematics	Limited selection	100	7600

3.3 Learning Behavior Evaluation Module

The learning behavior evaluation module mainly conducts data mining on the historical records of students' learning process, constructs the relationship between learning behavior and learning effect, evaluates students' academic performance according to the relationship model and gives a learning plan.

(1) Data preprocessing

The data used in the mining is taken from the online learning behavior data of students collected in the first semester and the next semester of 2019 on the distance learning website. In the last semester and the next semester, 15,000 students were randomly selected for the online learning behavior data of a certain course, and a total of 30,000 samples were preprocessed. Through the collection module of students' online learning behavior data, the following records can be obtained:

①The first type of records: video course learning, written materials learning, online Q&A behavior records

②The second type of record: the behavior record of BBS discussion

③The third type of record: behavior record of online test

The main work of the data preprocessing stage is to integrate, transform and reduce the above three types of learning behavior records in the database to finally form a data set to be mined.

What is obtained from the first type of record is the time-related data in student learning. Combining the characteristics of learning in distance education, and according to the advice of teaching experts, we divide the study of each course in a semester (20 weeks in total) into 5 learning units in a unit of 4 weeks, that is, taking 4 weeks as the time cutting point discretize time data. After merging the records, it is possible to obtain the time spent on text material learning, video course learning, and real-time Q&A respectively when each student learns a certain course in 5 learning units.

According to the objective laws of learning, students usually study hard and earnestly, and accordingly they will get better final grades. Therefore, when we construct the "learning behavior effect" model, we use the final grades obtained by the students as the learning effect to participate in the mining. The learning effect is divided into 4 levels: A, B, C, D, and the corresponding final scores are 80 points or more, 70-79 points, 60-69 points and 60 points or less.

(2) Generate decision tree

Generating a complete decision tree is done by the c4.5tree algorithm in C4.5. It needs to know the structure information of the data table item, and record the type or the range of the value of each attribute item through a type definition file with a suffix of .names.

The algorithmic process of constructing a decision tree: After the data structure is constructed, the decision tree is constructed in a circular manner, and each attribute is analyzed, and the classification of categories is determined by counting the sample number of the relationship between the attribute and the learning effect. The program goes through the cycle In this way, the entropy gain rate of each attribute is calculated first, and then the value of each attribute entropy gain rate is compared, and the attribute with the largest value is selected for classification, until a complete decision tree is finally generated.

It mainly includes calculating the expected information needed to classify a given sample. In this system, student evaluation results are divided into four grades: excellent, medium, passing, and failing; calculating the information entropy of each attribute; calculating the information gain rate of the attribute ; Inductive decision tree.

(3) Decision tree pruning and evaluation

When generating a decision tree, this study uses the depth of the limit tree and the number of elements contained in each node to perform the pre-pruning of the decision tree, and combines the post-pruning method of cross-checking to perform 10-fold cross-validation and cross-validation on the results of these tests. Random sequence distribution analysis, after the results of such verification, the rules are screened out, and then the confidence of these rules and variable support attributes are analyzed, and finally a better decision tree is selected as the predictive model to extract the rules.

This study selects online learning students as the research object, selects the student's system log data and questionnaire data to process the above steps, uses this data set as the input training sample set of the decision tree generation algorithm, when combined with the pre-pruning method and after the decision tree is pruned by the post-pruning method, the same source data is used to adjust, test and evaluate the decision tree model to verify the prediction accuracy of the model and perform final rule screening and extraction based on the evaluation results. Decision models with strong applicability, these two models are the basic basis for verifying and predicting the effect of online learning.

4. RESULTS AND DISCUSSION

Calculate the average of the academic performance as a form of expression of the learning effect. In order to avoid the decision-making errors caused by the score alone as the evaluation criterion, this study also examined the relationship between the attribute data of the learner's submission of homework and whether the homework was completed and the academic performance, as a supplement to the judgment of the learning effect. The statistical description of the learning performance data set is shown in Table 2.

Because this set of data is in the form of decimals with a relatively concentrated distribution, the method of finding quarter points is adopted to convert it into the form of four grades of "good, medium and bad". Use the method of the statistical formula QUARTILE to find the quarter points, and convert the data between each equal point (0.00/30.00/62.50/78.50/99.00) into the corresponding grade (good/medium/poor), and the converted data distribution percentages of the data in the four levels are flat, indicating that the division results are evenly distributed and well structured. The online learning effect level distribution is shown in Figure 1.

Table 2. Statistical description of the academic performance data set

Average	Standard error	Median	Mode	Standard deviation	Variance kurtosis
50	1.1	60	1	30	1000

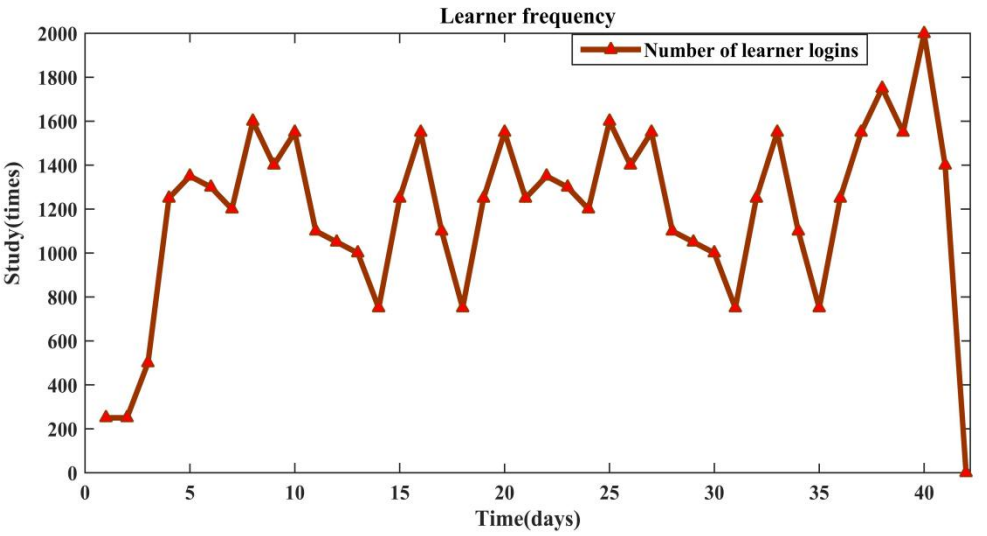
Figure 1. Online learning effect level distribution

The number of student online learning logins is shown in Figure 2. The number of logins will reflect the learning effect. The more logins, the higher the learner's enthusiasm and autonomy for online learning. Learners should persist in learning more than twice a week. Counting the cumulative number of learners logging in to the system every day, most learners can log in to the system twice a week or more, and the number of logins in the later period is significantly more than that in the early period. The learning frequency will also affect the learning effect. The learning effect is best when the learning frequency is moderate. The learning frequency is too large, indicating that the learner frequently logs in and exits the system, and the duration of each learning is too short, which is not conducive to the knowledge points of the same chapter continuous learning is not conducive to memory and connection. If the learning frequency is too small, it will take too long to log in to the system each time, and it will easily cause learning fatigue and cognitive load, and it is not conducive to the mastery and consolidation of knowledge.

In this study, the attribute of performance ranking is divided into four levels: "top 5%", "top 5%-10%", "top 10%-30%" and "other", representing top students and excellent students. Students, excellent students, and other four categories are identified by the application of 1 to 4; for the persistence of learning, four options for independent study time per week are "occasionally, uncertain", "1-2 days", " "3-5 days" and "6-7 days" correspond to the four levels of "poor, moderate, good, and excellent" for learning persistence, which are marked by 4 to 1; for the other dimensions, they are all very consistent. The options that do not meet the five grade distributions are represented by 1 to 5 (that is, small numbers are used to represent better performance of non-intellectual factors).

The correlation between different factors is shown in Table 3. It can be seen from Table 3 that the descending order of the correlation with the attribute of "score ranking" is "self-learning ability", "knowledge mastery", "persistence in learning", and "curriculum knowledge" in order.

Figure 2. Number of student logins in online learning



"Recognition"and "learning interest", and these attributes are positively correlated with "ranking results".

Table 3. Correlation between different factors

Parameter	Persistence	Self-learning	Grasp knowledge	Recognition	Interest
Persistence	0.3034	-	-	-	-
Self-learning	0.1946	0.6161	-	-	-
Grasp knowledge	0.2107	0.4205	0.4886	-	-
Recognition	0.1812	0.3440	0.4411	0.6837	-
Interest	0.0829	0.1842	0.1426	0.0314	0.0183

The top students' recognition of autonomous learning ability is quite prominent, while the proportion of good students who choose "neutral" has increased sharply. For the students of the four ranking levels, they all agree with the importance of autonomous learning ability. Only the students in the fourth level group disagree with the autonomous learning ability at all. Figure 3 shows the statistics of the importance of autonomous learning ability.

As shown in Figure 4, the number of people who agree with the importance of the curriculum basically decreases in the order of rank, while the number of people who hold a neutral attitude towards the recognition of the curriculum increases approximately in order. From the overall distribution, the proportion of people who agree with the importance of the course is the highest.

The corresponding relationship between "interest in continuing learning" and "ranking scores" is shown in Figure 5. According to the four levels of performance ranking from high to bottom, the number of people who fully agree and the number of people who agree more closely form two peaks with opposite peaks. Normal distribution curve. From the results of the correlation coefficient analysis, we can see that the element of interest has the lowest correlation with the performance

Figure 3. The importance of autonomous learning ability

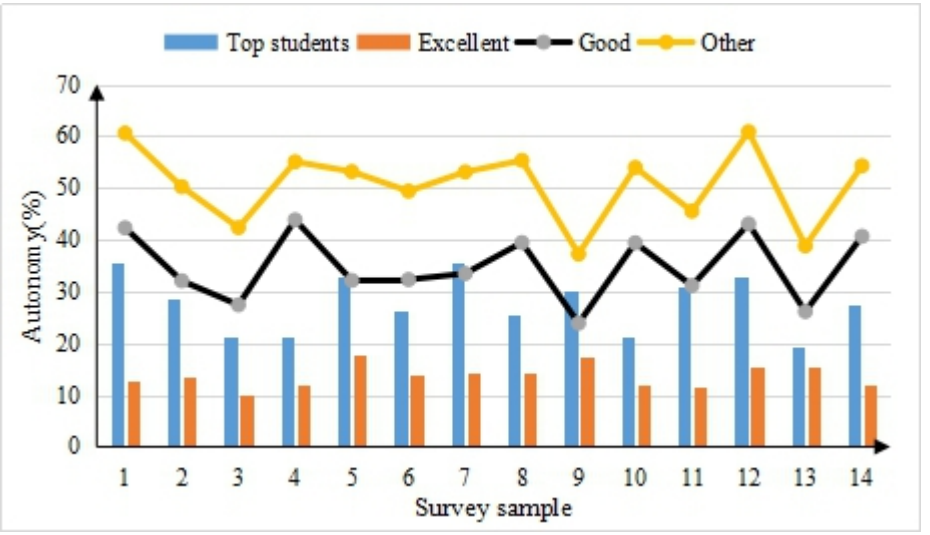
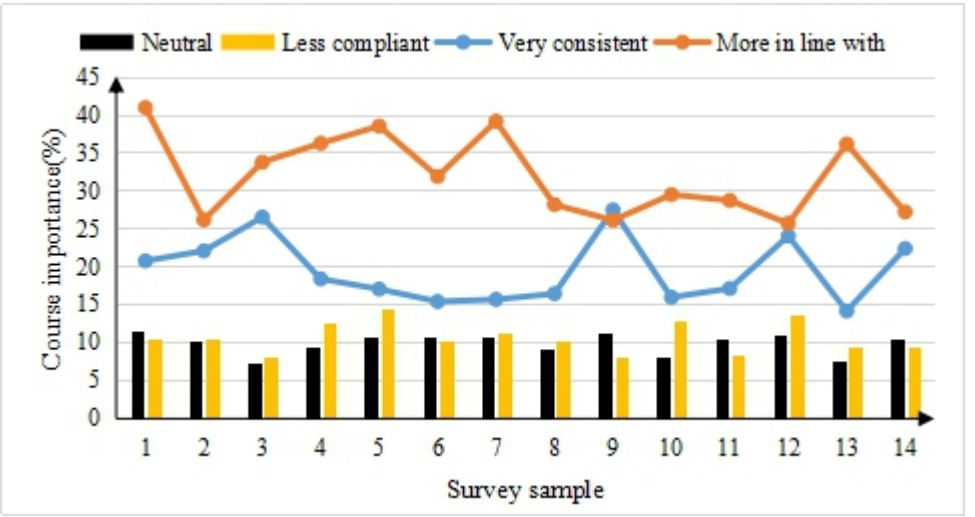


Figure 4. Recognize the importance of the curriculum



ranking, and it can also be seen from Figure 8 that the distribution of the number of people at each level is basically irregular.

Self-learning persistence is shown in Figure 6. Only the top student group thinks that their learning persistence is excellent, while the number of people who think their learning persistence belongs to good, medium, and poor has a smooth distribution curve, which is approximately horizontal. On the whole, the number of people holding a neutral attitude has the highest rate at all levels.

The effect of online learning predictive learning is shown in Figure 7. It can be seen from Figure 7 that the prediction accuracy rate of predicting learning effects based on "online learning behavior"

Figure 5. Correspondence between interest in continuing learning and ranking results

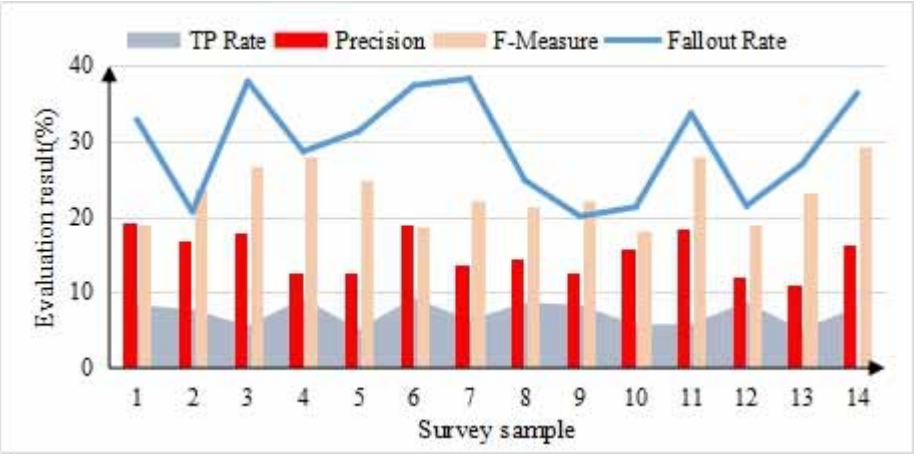
Figure 6. Self-learning persistence

has reached more than 66%, indicating that its confidence level is relatively reliable, showing that the level of learning performance is distributed in "good, medium, and poor" the prediction accuracy rate of the students of all is higher than 66%, and the prediction results support the effectiveness of the prediction mechanism to be obtained in this research.

5. CONCLUSION

Classification mining methods are mostly used in commercial systems, including reputation verification, medical diagnosis, customer classification and so on. In the existing data mining applications, we rarely see the mining of educational information, and we rarely see the mining of classification rules directly on the data reflecting students' academic conditions. This article aims to conduct exploratory research on data mining in the field of education. In the study, online learning students are selected as the research object, and the student's system log data and questionnaire data

Figure 7. Online learning predicts the learning effect



are selected. This data set is used as the input training sample set of the decision tree generation algorithm. When combining the pre-pruning method and the post-pruning method to make decisions after the tree is pruned, the same source data is used to adjust, test and evaluate the decision tree model. In future research, data such as student course performance data and basic student information can be mined. Use the discovered rules and laws to provide meaningful information for educational management decision-making, help educational management allocate educational resources, adjust educational plans, reform teaching models and curriculum settings, and strengthen and promote related disciplines.

ACKNOWLEDGMENT

Topic: Introduction to the Basic Principles of Marxism online and offline hybrid teaching mode reform and practice, the teaching reform project of Yantai institute of Technology, project number: 2020JYA09.

REFERENCES

Andrew, R. (2018). Decision Tree for Pretreatments for Winter Maintenance. *Transportation Research Record: Journal of the Transportation Research Board*, *2055*(1), 106–115.

Berthon, B., Marshall, C., Evans, M., & Spezi, E. (2016). ATLAAS: An automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography. *Physics in Medicine and Biology*, *61*(13), 4855–4869. doi:10.1088/0031-9155/61/13/4855 PMID:27273293

Bilal, M., Israr, H., & Shahid, M. (2016). Sentiment classification of Roman-Urdu opinions using Nave Bayesian, Decision Tree and KNN classification techniques. *Journal of King Saud University - Computer and Information Sciences, 28*(3), 330-344.

Chang, T. C., & Wang, H. (2016). A Multi Criteria Group Decision-making Model for Teacher Evaluation in Higher Education Based on Cloud Model and Decision Tree. *Eurasia Journal of Mathematics, Science and Technology Education*, *12*(5), 1243–1262. doi:10.12973/eurasia.2016.1510a

Di, J., & Xu, Y. (2019). Decision tree improvement algorithm and its application. *International Core Journal of Engineering*, *5*(9), 151–158.

Goodman, K. E., Lessler, J., & Cosgrove, S. E. (2016). A Clinical Decision Tree to Predict Whether a Bacteremic Patient Is Infected With an Extended-Spectrum β-Lactamase–Producing Organism. *Acta Crystallographica*, *67*(7), 363–367. PMID:27358356

Gudelis, M., Garcia, J., & Cabello, J. (2019). Diagnosis of Pain in the Right Iliac Fossa. A New Diagnostic Score Based on Decision-Tree and Artificial Neural Network Methods. *Cirugia Espanola*, *97*(6), 329–335. doi:10.1016/j.ciresp.2019.02.006 PMID:31005266

Hameed, A., Dai, R., & Balas, B. (2016). A Decision-Tree-Based Perceptual Video Quality Prediction Model and Its Application in FEC for Wireless Multimedia Communications. *IEEE Transactions on Multimedia*, *18*(4), 764–774. doi:10.1109/TMM.2016.2525862

Hen Da Lianpour, A., Razmi, J., & Sarvestani, A. R. (2016). Applying decision tree models to SMEs: A statistics-based model for customer relationship management. *Management Science Letters*, *6*(7), 509–520. doi:10.5267/j.msl.2016.5.002

Jin, W. T., Rivers, C. S., & Fallah, N. (2019). Decision tree analysis to better control treatment effects in spinal cord injury clinical research. *Journal of Neurosurgery. Spine*, *31*(4), 1–9. PMID:30933908

Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., & Mishra, S. (2016). Decision Tree and SVM-Based Data Analytics for Theft Detection in Smart Grid. *IEEE Transactions on Industrial Informatics*, *12*(3), 1005–1016. doi:10.1109/TII.2016.2543145

Kad, A., Bp, B., & Bus, C. (2019). Decision tree for modeling survival data with competing risks. *Biocybernetics and Biomedical Engineering*, *39*(3), 697–708. doi:10.1016/j.bbe.2019.05.001

Kg, A., Mtb, C., & Sk, D. (2019). An assessment of the risk factors for vitamin D deficiency using a decision tree model. *Diabetes & Metabolic Syndrome*, *13*(3), 1773–1777. doi:10.1016/j.dsx.2019.03.020 PMID:31235093

Kumar, A., & Ghosh, A. K. (2019). Decision Tree- and Random Forest- Based Novel Unsteady Aerodynamics Modeling Using Flight Data. *Journal of Aircraft*, *56*(1), 403–409. doi:10.2514/1.C035034

Lu, L., Di, L., & Ye, Y. (2017). A Decision-Tree Classifier for Extracting Transparent Plastic-Mulched Landcover from Landsat-5 TM Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(11), 4548–4558. doi:10.1109/JSTARS.2014.2327226

Mistry, P., Neagu, D., Trundle, P. R., & Vessey, J. D. (2016). Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology. *Soft Computing*, *20*(8), 2967–2979. doi:10.1007/s00500-015-1925-9

Niu, F., & Chen, L. (2019). Forecasting of Landslide Stability Based on Gradient Boosting Decision Tree Model. *International Core Journal of Engineering*, *5*(11), 42–48.

Nourani, V., Tajbakhsh, A. D., & Molajou, A. (2019). Data mining based on wavelet and decision tree for rainfall-runoff simulation. *Nordic Hydrology*, *50*(1-2), 75–84. doi:10.2166/nh.2018.049

Petrich, J. (2020). Development and application of a decision tree to determine appropriate handling of gene therapies. *Cytotherapy*, *22*(5), S157–S158. doi:10.1016/j.jcyt.2020.03.327

Pham, B. T., Khosravi, K., & Prakash, I. (2017). Application and Comparison of Decision Tree-Based Machine Learning Methods in Landside Susceptibility Assessment at Pauri Garhwal Area, Uttarakhand, India. *Environmental Processes*, *4*(3), 711–730. doi:10.1007/s40710-017-0248-5

Roy, S., Mondal, S., Ekbal, A., & Desarkar, M. S. (2019). Dispersion Ratio based Decision Tree Model for Classification. *Expert Systems with Applications*, *116*(FEB), 1–9. doi:10.1016/j.eswa.2018.08.039

Sanz, J., Fernandez, J., & Sola, H. B. (2017). A decision tree based approach with sampling techniques to predict the survival status of poly-trauma patients. *International Journal of Computational Intelligence Systems*, *10*(1), 440–455. doi:10.2991/ijcis.2017.10.1.30

Shalabi, H., & Hadjisophocleous, G. (2020). Decision tree for FSSA rooms in CANDU. *Nuclear Engineering and Design, 361*(May), 110568.1-110568.27.

Sheng, Z. (2019). Application of Logistic regression and decision tree analysis in prediction of acute myocardial infarction events. *Zhejiang da xue xue bao. Yi xue ban = Journal of Zhejiang University. Medical Science*, *48*(6), 594–602.

Subbotin, S. (2020). Radial-Basis Function Neural Network Synthesis on the Basis of Decision Tree. *Optical Memory and Neural Networks (Information Optics)*, *29*(1), 7–18. doi:10.3103/S1060992X20010051

Tanha, J., Van Someren, M., & Afsarmanesh, H. (2017). Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics*, *8*(1), 355–370. doi:10.1007/s13042-015-0328-7

Tsangaratos, P., & Ilia, I. (2016). Landslide susceptibility mapping using a modified decision tree classifier in the Xanthi Perfection, Greece. *Landslides*, *13*(2), 305–320. doi:10.1007/s10346-015-0565-6

Wang, Y., Xia, S. T., & Wu, J. (2017). A less-greedy two-term Tsallis Entropy Information Metric approach for decision tree classification. *Knowledge-Based Systems, 120*(15), 34-42.

Juxin Shao, female, born in June of 1973 in Zhaoyuan, Shandong, is lecturer, postgraduate, master of philosophy, Marxist philosophy, and social development.

Qian Gao, female, born in April of 1978 in Yantai, Shandong, is lecturer, postgraduate, master of science, operations research and cybernetics, optimal control.

Hui Wang, male, born in June of 1976 in Yantai, Shandong, is senior engineer, undergraduate, bachelor of engineering, computer network and technology, artificial intelligence.