Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Application of Data Mining in Education

Qihao Chen *

College of Education, Zhejiang University of Technology, Zhejiang, China

* Corresponding Author Email: 202205720103@zjut.edu.cn

Abstract. With the rapid growth of educational data, how to extract valuable information from multi-source heterogeneous data has become an important challenge in educational research. The existing technology applications are scattered and lack systematic integration, which makes it difficult for educational practitioners to comprehensively evaluate the applicability of different methods. This paper systematically reviews the research progress in student behavior analysis, academic prediction and educational intervention from four perspectives: cluster analysis, prediction model construction, association rule mining and data visualization. By integrating domestic and foreign empirical cases, it is found that: clustering algorithms can effectively divide student groups, but the initial center selection and mixed data type processing still need to be optimized; deep learning models perform well in dynamic prediction, but the lack of interpretability limits the direct application of educational decision-making; association rule mining reveals the complex interaction between cognitive style and learning behavior; data distillation bridges the gap between algorithm output and educational interpretability through visualization technology. This paper proposes an innovative path for technical collaboration and emphasizes the importance of interdisciplinary cooperation in solving ethical challenges such as data integrity and algorithm fairness.

Keywords: Cluster analysis; Prediction model; Association rule mining; Data visualization; Educational data mining.

1. Introduction

The explosive growth of educational data stems from the widespread popularity of online education platforms, intelligent learning tools and campus information systems, forming a multimodal data source covering learning behavior logs, course interaction records, academic assessment results and social emotional feedback. According to statistics from the International Society for Technology in Education (ISTE), about 78% of higher education institutions around the world have deployed learning management systems, generating TB-level structured and unstructured data every day, such as video click streams, forum texts and physiological signals collected by sensors [1]. However, the complexity of data dimensions also brings severe challenges: on the one hand, traditional analysis methods (such as linear regression) are difficult to capture nonlinear associations and dynamic interaction patterns, resulting in the potential educational value not being fully explored; on the other hand, the "island effect" of technology application is significant - although clustering algorithms can divide student groups, it is difficult to coordinate with causal reasoning models to optimize intervention strategies, and the high-precision predictions of deep learning often hinder educational interpretability due to the "black box" characteristics [2] [3]. This paper reviews the following four dimensions: cluster analysis, prediction model construction, association rule mining, and data visualization and annotation. The significance of this review is: first, by comparing the applicable scenarios of different technologies horizontally, it provides a basis for educational researchers to choose methodologies; second, it combines empirical cases to refine innovative paths for technology collaboration, such as "clustering + association rules" to optimize clustering strategies; finally, it critically reflects on the limitations of technology and points out directions for future research, including multimodal data fusion. Through systematic integration, this paper aims to promote the leap from theoretical exploration to practical implementation of educational data analysis.

2. Methods and Process

2.1. Cluster Analysis (K-Prototype)

Cluster analysis is an unsupervised learning method. Its core goal is to divide data objects into several clusters according to their similarity, so that the samples in the same cluster are highly similar and the differences between different clusters are large. Commonly used clustering methods include K-Prototype, K-Means, etc [4] [5]. As a key technology in data analysis, clustering is widely used in various data-driven scenarios and is considered an indispensable task in machine learning [2]. Among them, K-Prototype combines the advantages of K-Means (processing numerical data) and K-Modes (processing categorical data). By defining a mixed distance metric (Euclidean distance measures numerical attributes, and Hamming distance measures categorical attributes), it effectively solves the limitations of traditional clustering algorithms in mixed data scenarios. K-means achieve clustering by dividing the data set into K clusters. The algorithm first randomly selects K initial center points, then assigns each data point to the nearest center point, and then updates the center point of each cluster. This process is repeated until the clustering results converge [6]. K-means is simple and efficient, but it is sensitive to the selection of the initial center point and requires the number of clusters K to be specified in advance. It is suitable for large-scale data and situations with obvious clustering patterns.

Research proposed a method for analyzing and predicting college student behavior based on multi-source data mining. The method first integrated heterogeneous data from multiple university functional departments such as academic affairs, scholarships, and employment, and constructed a comprehensive education data set covering three dimensions: basic information, training process, and graduation behavior. The sample covered approximately 18,000 undergraduates. In view of the problem that traditional clustering methods make it difficult to handle categorical variables, research [6] used the K-Prototype algorithm to perform cluster analysis of student behavior characteristics and successfully divided students into three groups: high, medium, and low academic performance. Further analysis showed that the correlation between college grades and college entrance examination scores and family economic status was low, while students' academic performance in the freshman year had a significant continuity in subsequent grades. In terms of employment prediction, the study used Bayesian optimization combined with the XGBoost algorithm to build a prediction model with an F1 value of 0.872, which is better than the comparison models such as random forest and SVM. The SHAP interpretability analysis method was introduced to reveal the key driving factors behind employment choices: for example, students with higher admission scores are more likely to find employment, students with more scholarships are more inclined to study in China, and the proportion of students from rural backgrounds or with financial difficulties studying abroad is relatively low. This study explored clustering modeling, interpretable prediction, and other aspects, and put forward specific management suggestions based on data mining results, providing scientific and feasible decision-making basis for colleges and universities in student development support, employment guidance, risk warning, etc.

Research proposed an improved K-means algorithm for the problem of student spatial prediction. By optimizing the initialization steps and cluster number selection in the clustering process, it is more suitable for student spatial ability analysis. When processing the operation behavior of students in the three-dimensional design course, the algorithm can accurately identify student groups with different spatial ability characteristics. Compared with the traditional algorithm, the improved K-means algorithm performs better in the clustering effect of students' spatial ability and can better capture the spatial ability characteristics of students under different operations and command behaviors.

2.2. Deep Learning

Deep learning is a method of machine learning that simulates the way the human brain processes information through multi-layer neural networks, automatically extracting features from large amounts of data and making decisions. It can self-learn features without manual design. Deep learning

usually requires a large amount of data and computing resources. Common models include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs). A convolutional neural network is a feedforward neural network, which evolved from multi-sensor, and has the structural characteristics of local connection, weight sharing and down sampling. The network uses the convolution kernel in the neural network to perform convolution operations on information such as images and share weights. Each convolution kernel focuses on a core feature, and these features are combined as the global feature information of the overall input [7]. In addition to the K-Means method, the researchers in the study [3] proposed a two-layer network model combining a convolutional neural network (CNN) and a long short-term memory network (LSTM) to solve the temporal impact of students' operation behavior on spatial ability changes in the three-dimensional design course (Figure 1). CNN is responsible for extracting features from students' operation data, while LSTM processes time series data to capture the long-term dependence of operation sequence on students' spatial ability changes. The CNN-LSTM model successfully transforms the spatial ability prediction problem into a time series classification task, and through experimental verification, it achieved a prediction accuracy of 0.89 on the validation set. Compared with traditional machine learning methods and single neural network models, the CNN-LSTM model performs better in all indicators, demonstrating the powerful ability of deep learning in dynamic spatial ability prediction.

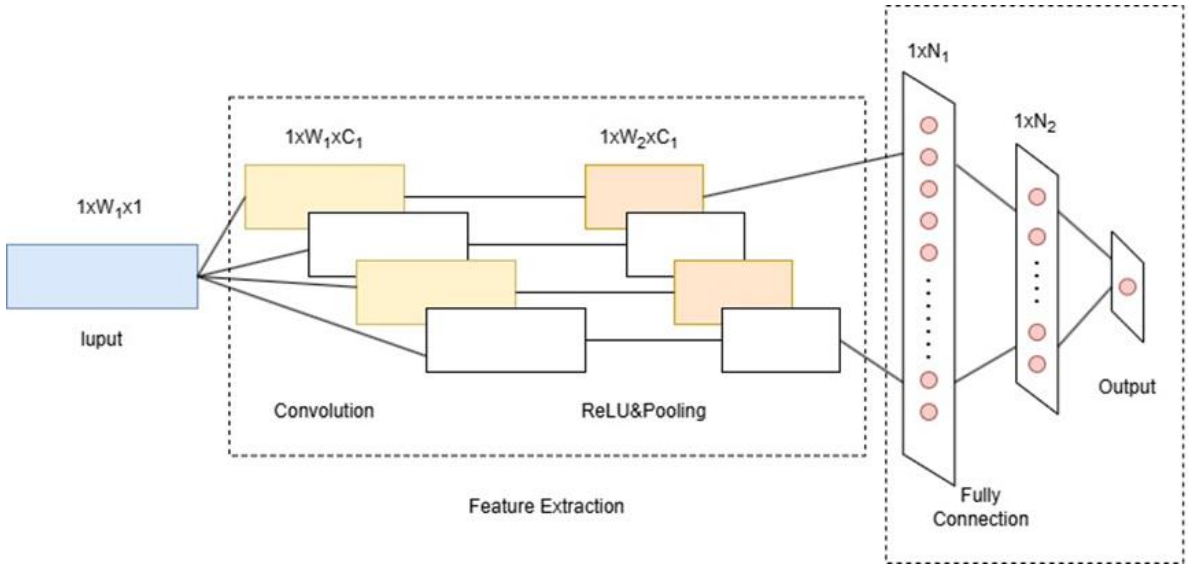

Figure 1. CNN-LSTM Structure [2]

2.3. Association Rule Mining

Association rule mining is a data mining technique that aims to discover correlations between data items from a large number of data sets. Its core goal is to reveal patterns of frequent co-occurrence in a data set, that is, when certain data items appear at the same time, other data items may also appear. In the field of education, association rule mining technology is mainly used to analyze students' learning behavior and academic performance to discover important factors that affect learning outcomes [3].

The use of association rule mining was well applied in the study of Charoula Angeli et al. [8]

The first study analyzed the interaction differences between field-dependent (FD) and field-independent (FI) students when using transparent simulation tools through association rule mining. The study found that FI students can systematically control all independent variables (such as IV1-IV5) and form high-order combination rules (Table 2), while FD students' rules only involve limited variables (such as IV1 and IV2) (Table 1). This shows that association rules can reveal the impact of cognitive style on the efficiency of technology use and provide data support for optimizing learning design.

Table 1. Sequential rules for the FD learners

Antecedent ==>Consequent
1.(B),(B)==>(T)
2.(B),(B)==>(M)
3.(B),(B)==>(P)
4.(B),(B),(T)==>(M)
5.(B),(B),(T).(M)==>(S)
6.(B),(B),(T)==>(P)
7.(B),(B),(T),(P)==>(S)
8.(B),(B)==>(T),(S)
9.(B).(B)==>(M),(P)
10.(B),(B),(M),(P)==>(S)
11.(B)==>(T),(P)
12.(B),(T),(M)==>(P),(IV2)
13.(B),(T),(M)==>(P),(IV1)

Note: B: BUILD: T: TEST: M: METER: P: PLAY: S: STOP: IV1 = Country A-Number of births: lV2=Country B-Movement of businesses

Table 2. Sequential rules for the FI learners

Antecedent ==>Consequent
1.(B),(T).(M).(P)==>(IV1)
2.(B),(T).(M).(P)==>(IV2)
3.(B),(T).(M).(P)==>(IV3)
4.(B),(T).(M).(P)==>(IV4)
5.(B),(T).(M).(P)==>(IV5)
6.(B),(T),(M).(P)==>(IV1),(IV2)
7.(B),(T).(M).(P),(IV1)==>(IV5)
8.(B),(T).(M),(P)==>(IV1),(IV3)
9.(B).(T),(M).(P),(IV1)==>(IV2).(IV5)
10.(B),(T).(M)(P)==>(IV1),(IV4)
11.(B).(T),(M),(P)==>(IV1).(IV5)
12.(B).(T).(M),(P)==>(IV2),(IV4)
13.(B),(T),(M),(P),(IV1)==>(IV2),(IV3)
14.(B),(T),(M),(P),(IV2)==>(IV3)
15.(B).(T),(M).(P).(IV1)==>(IV2).(IV4)
16.(B),(T).(M),(P),(IV1).(IV3)==>(IV4)
17.(B),(T),(M),(P),(IV1)==>(IV2).(IV5)
18.(B),(T),(M),(P).(IV1),(IV3)==>(IV5)
19.(B),(T),(M),(P),(IV1)==>(IV4).(IV5)
20.(B),(T),(M),(P),(IV4)==>(IV5)
21.(B),(T).(M),(P),(IV2)==>(IV5)
22.(B),(T),(M),(P),(IV3)==>(IV4)
23.(B),(T),(M),(P).(IV3)==>(IV5)

Note: B: BUILD: T: TEST: M: METER: P: PLAY: S: STOP: IV1 = Country A-Number of births: lV2=Country B-Movement of businesses

The second study combined fuzzy representation and association rule mining to analyze the relationship between Australian school questionnaire data and NAPLAN scores. The results showed that students' positive or negative attitudes towards ICT were significantly correlated with computer performance (such as "no knowledge" or "low performance") (Figure 2, Figure 3). For example, in the ICT-positive group, low computer performance was associated with medium literacy/calculation scores, while in the ICT-negative group, such rules were more complex. This shows that association

rules can capture the "complex system characteristics" of multi-factor interactions and assist schools in developing targeted intervention strategies.

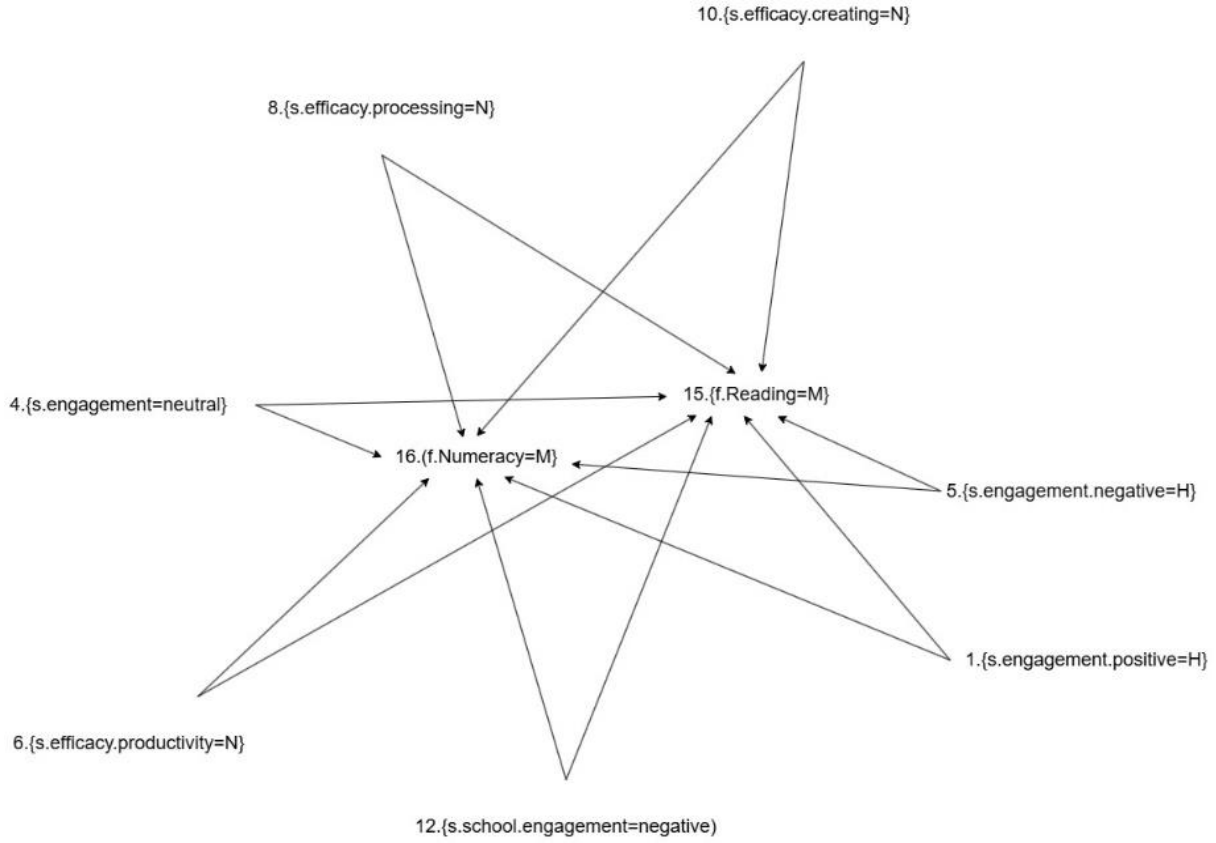

Figure 2. Positive ICT engagement [8]

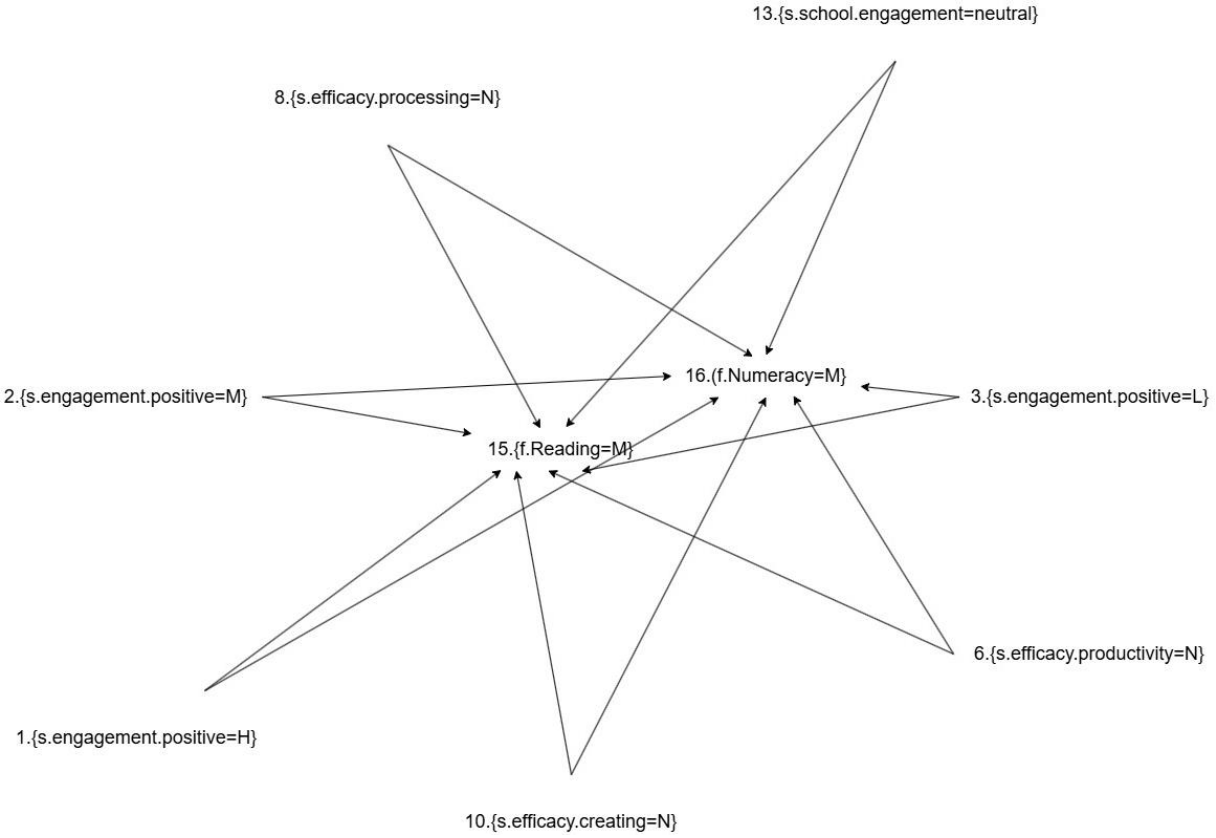

Figure 3. Negative ICT engagement [8]

2.4. Human-computer Interaction Data Distillation

Data distillation refers to the use of information visualization technology to transform complex data sets into a form that is easy for humans to understand and judge, thereby assisting researchers in identifying potential patterns or completing manual annotation to support subsequent modeling and analysis [1]. In educational data mining, this method is often used in two scenarios: first, to reveal implicit rules in learning behaviors, such as the mastery trajectory of knowledge components, through visualization tools (e.g., learning curves) [9]; The second is to speed up the manual annotation process by simplifying the data presentation format and improve the data annotation rate [10], providing high-quality labeled data for building predictive models.

Baker defines data distillation as a key method to support educational decision-making by transforming complex educational data into a form that can be intuitively analyzed by humans through visualization and structured display technology. Its core goal is to combine human cognitive advantages with automated analysis capabilities to solve pattern recognition and decision support problems that are difficult for traditional algorithms to directly handle. Data distillation includes two core tasks: pattern recognition and classification labeling, emphasizing that the uniqueness of data distillation lies in demonstrating multi-level information integration capabilities and dynamic interactive support through cross-level association analysis and real-time parameter adjustment.

Baker's literature describes in detail the application of data distillation in the following scenarios: student behavior pattern recognition, that is, through the visualization of time-series behavior sequences (such as clickstream analysis), such as identifying two types of learning modes: "cramming" and "continuous learning"; model verification and improvement, that is, through the heat map to display the frequency of erroneous co-occurrence between knowledge points, verify the rationality of the division of knowledge components. For example, he found that the "calculus chain rule" and the "integral substitution method" are often confused by students, and then correct the boundaries of knowledge components.

3. Current Limitations

First, in the K-means clustering process, the number of clusters k needs to be determined. However, when faced with unfamiliar data sets, researchers often find it difficult to accurately determine the appropriate k value, and can usually only rely on experience or heuristic methods to make selections. Due to different judgment criteria of different people, the final clustering results may be biased, and even affect the understanding of the data pattern. In addition, K-means needs to randomly select the initial clustering center, and different initial points may lead to different clustering results. If the position of the initial center deviates from the actual distribution of the data, it may cause the convergence speed to slow down or even fall into a local optimum, which greatly reduces the clustering effect. Therefore, in practical applications, it is often necessary to run K-means multiple times and use methods such as K-means++ to optimize the selection of the initial center to improve the stability and accuracy of clustering.

The second is algorithm interpretability. In educational data mining, many studies use deep learning and complex machine learning models, such as neural networks and random forests. Although these methods can provide high-precision predictions, they have low interpretability. For example, although research based on association rule mining can reveal the relationship between variables, it lacks an intuitive explanation of the rationality of the rules and is difficult to directly apply by educational decision makers. In addition, the "black box" nature of deep learning models makes it difficult to interpret the results of tasks such as student performance prediction and learning behavior analysis, affecting the transparency and acceptability of decision-making.

The third is the problem of data completeness. Currently, many research data come from online platforms. Although they have been anonymized, there are still some omissions and ambiguities, and the data integrity is poor. When conducting cluster analysis based on the input dimension, there are

certain problems with the data on learner input. Learner input is a complex multidimensional concept that cannot be measured simply by data. In addition to learners' behavioral and social dimensions, it also involves learners' cognitive and emotional levels, and this information cannot be obtained in the current data set [6].

4. Conclusion

This paper systematically reviews the research progress and practical significance of educational data mining from four technical perspectives: cluster analysis, deep learning, association rule mining, and data distillation. In cluster analysis, K-means and its improved algorithm reveal the stability differences in academic performance through student clustering, but the initial center selection and mixed data processing are still key challenges; the CNN-LSTM hybrid model of deep learning significantly improves the accuracy of student ability prediction through spatiotemporal feature fusion, but its lack of interpretability limits its direct application in educational decision-making; association rule mining starts from the correlation between cognitive style and learning behavior, reveals the complex laws of multivariate interaction, and provides a basis for personalized learning design; data distillation bridges the gap between algorithm output and educational interpretability through visualization and manual annotation.

The significance of this review is that, on the one hand, it provides a methodological reference for educational researchers by comparing the advantages and limitations of the technology horizontally; on the other hand, it proposes innovative paths for technology integration (such as "clustering + association rules" to optimize student clustering strategies), and calls for strengthening interdisciplinary cooperation to address ethical challenges such as data integrity and algorithm fairness. Future research needs to further explore the integration of lightweight models, dynamic annotation tools and multimodal data, and promote educational data mining from theoretical exploration to large-scale application.

References

[1] Baker, R. S. J. d. Data Mining for Education. International Encyclopedia of Education (3rd ed.), Elsevier, in press.

[2] Han Xiaotian. Research on Student Spatial Ability Analysis and Prediction Based on Educational Data Mining. Shanghai Ocean University, 2024.

[3] Wang Yuan. Research on Behavior Analysis and Prediction of College Students Based on Multi-Source Data Mining. Beijing University of Chemical Technology, 2023. DOI: 10.26939/d.cnki.gbhgu.2023.000049.

[4] Casolla G, Cuomo S, Di Cola V S, et al. Exploring Unsupervised Learning Techniques for the Internet of Things. IEEE Transactions on Industrial Informatics, 2019, 16 (4): 2621 - 2628.

[5] Liang W, Li K, Long J, Kui X, Zomaya A Y. An Industrial Network Intrusion Detection Algorithm Based on Multifeature Data Clustering Optimization Model. IEEE Transactions on Industrial Informatics, 2019, 16 (3): 2063 - 2071.

[6] Levkivskyi V. Research of Algorithms of Data Mining. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2020, 16: 417 - 428.

[7] Li Yan. Research on Academic Prediction and Intervention Based on Online Learning Data. Yunnan Normal University, 2024. DOI: 10.27459/d.cnki.gynfc.2024.000635.

[8] Angeli C, Howard S K, Ma J, et al. Data Mining in Educational Technology Classroom Research: Can It Make a Contribution? Computers & Education, 2017, 113: 226 - 242.

[9] Corbett A T, Anderson J R. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction, 1995, 4 (4): 253 - 278.

[10] Baker R S J d, Corbett A T, Koedinger K R, et al. Developing a Generalizable Detector of When Students Game the System. User Modeling and User-Adapted Interaction, 2008, 18 (3): 287 - 314.