Expert Systems With Applications 237 (2024) 121555

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualization

Guiyun Feng, Muwei Fan [*]

School of Management, Guizhou University, Guiyang, 550025, China

ARTICLE INFO

Keywords:
Educational data mining
Learning behavior patterns
Evaluation methodologies
Classification algorithms

ABSTRACT

The rapid growth of educational data creates the requirement to mine useful information from learning behavior patterns. The development of data mining technology makes educational data mining possible. The paper intends to use a public educational data set to study learning behavior patterns from the perspective of educational data mining, so as to promote the innovation of educational management. Firstly, in order to reduce the dimension of data analysis that facilitates the improvement in efficiency, principal component analysis is carried out to reduce the number of attributes in the data set. The significant attributes in the rotating principal component matrix rather than principal components which are not closely related to learning behavior patterns are extracted as the research variables. Then, a pseudo statistic is proposed to determine the number of clusters and the preprocessed data set is clustered according to the extracted attributes. The clustering results are applied to add class labels to the data, which is convenient for the later data training. Finally, six classification algorithms J48, K-Nearest Neighbor, Bayes Net, Random Forest, Support Vector Machine and Logit Boost are used to train the data with labels and build prediction models. At the same time, the performance and applicable conditions of six classifiers in terms of accuracy, efficiency, error, and so on are discussed and compared. It is found that the performance of the integrated algorithm is better than that of a single classifier. In the integrated algorithm, compared with Random Forest, the running time of Logit Boost is shorter.

1. Introduction

Data mining is capable of finding the useful information hidden behind simple data and processing large data sets. Educational data mining (EDM) is the application of data mining in the field of education, which was proposed by Romero and Ventura (2007) . EDM bridges the gap between two disciplines: education on the one hand, computing sciences on the other, where both data mining and machine learning as subfields of computing sciences are the focus (Bakhshinategh, Zaiane, ElAtia, & Ipperciel, 2018). The goal of EDM is to better understand how students learn and identify the settings in which they learn to improve educational outcomes and to gain insights into and explain educational phenomena (Romero & Ventura, 2013). It is concerned with developing methods for exploring the unique types of data that come from educational environments (Bakhshinategh et al., 2018). In short, EDM aims to utilize data mining technology to find the information hidden behind massive educational data, and use relevant information to promote the progress and development of education.

There are many reasons to study students' learning behavior patterns. For example,

- if students are ranked according to the traditional ranking system, it is very likely that many students are concentrated in one rank, resulting in the phenomenon that there is no one in other ranks. This is not an objective method of evaluating student performance. The reason for this phenomenon is that a certain course is difficult. Many students' scores are concentrated between 60 and 80, and no one gets 90, so no one gets "excellent". Clustering, as an important method in EDM, can solve this problem well by learning the law from real data and ranking them objectively.
- if we can find students with academic warning risk at the initial stage of the course, we could improve their performance opportunely (Riestra-Gonzalez, Paule-Ruiz, & Ortin, 2021). If we want to find students at risk and help them in time, we can use historical data to predict students' future performance. The classification algorithm in data mining just has such a function which can not only predict future performance, but also provide prediction metrics such as accuracy.
- a large amount of data related to students' behavior is useless in front of anyone without being processed. By using data mining to process data and present the results to stakeholders in a visual

[*] Corresponding author.
E-mail addresses: 1051184826@qq.com (G. Feng), mwfan@gzu.edu.cn (M. Fan).

https://doi.org/10.1016/j.eswa.2023.121555
Received 30 March 2023; Received in revised form 28 August 2023; Accepted 9 September 2023
Available online 15 September 2023

form, the data can be transformed into useful information which can be made full use of to improve teaching and learning.

Thus, it is very appropriate to apply EDM to study learning behavior patterns in terms of technology and research content which corresponds to the title of the paper. By using the methods which are closely related to EDM to study the learning behavior patterns of students, we can analyze, evaluate and predict the learning behavior. Therefore, this paper intends to study the learning behavior patterns that can best reflect the connotation of educational data from the perspective of educational data mining so as to provide reference for stakeholders. What is more, we can also innovate educational management by using visual research results. With the purpose of achieving research objectives, three research questions of this paper are proposed.

1. The rapid growth of educational data makes traditional processing methods unsuitable for studying learning behavior patterns, so data mining comes into play. Data mining technology has been applied in the field of education, and formed an important research field – education data mining. Therefore, how to study learning behavior patterns from the perspective of educational data mining is an urgent problem to be solved.
2. Learning behavior pattern is a relatively abstract phrase. To implement it into specific data mining technology, learning behavior pattern must be concrete. Thus, what aspects of learning behavior patterns can be studied?
3. The ultimate purpose of our research is to promote the development of education and innovate the management of education. How to promote the development of education and innovate the management of education by studying the learning behavior pattern from the perspective of educational data mining?

The remaining sections of this paper are organized as follows. Section 2 introduces the related work of the research topic in this paper, including analysis, evaluation and prediction of learning behavior. Section 3 analyzes the research methods used in this paper and describes the research results briefly. Section 4 answers the research questions according to the process and results of the research. The last section concludes the full text and makes an outlook to the research directions in the future.

2. Literature review

2.1. Analysis and evaluation of learning behavior

Many researchers use or improve the existing data mining technology to analyze and evaluate learning behavior. Manoharan, Ganesh, Felciah, and Banu (2014) introduce a deterministic model based on the clustering algorithm to analyze and monitor students' performance. Busalim, Masrom, and Wan (2019) study and analyze the effects of social software addiction and self-esteem on academic performance. Crivei, Czibula, Ciubotariu, and Dindelegan (2020) explore the usefulness of unsupervised machine learning methods, principal component analysis and association rule mining in analyzing students' academic achievement data in order to develop a supervised learning model for students' achievement prediction. Delgado, Morán, José, and Burgos (2021) adopt a new unsupervised clustering technology based on self-organizing mapping (SOM) artificial neural network (ANN) model to analyze online learning records. Mai, Bezbradica, and Crane (2022) propose a new method to deal with the problems of noise and trend effect in data to analyze students' learning behavior, and achieve success in detecting students with similar learning behavior and results.

Dynamic evaluation method has been proved to be a tool to find students' learning potential. In a learning environment where learning is currently mediated through technology, dynamic evaluation methods have a significant impact on students' academic performance (Zhang, Lai, Cheng, & Chen, 2017). Varela, Montero, Vásquez, Guiliany, Mercado, et al. (2019) use clustering technology as a useful management strategy tool to divide the population into homogeneous groups according to students' characteristics and skills to evaluate learning behavior. The evaluation of students' academic performance can be regarded as a clustering problem, and the hybrid clustering method is applied to evaluate academic performance in the educational environment (Yadav, 2020). Karthikeyan, Thangaraj, and Karthik (2020) propose a hybrid educational data mining (HEDM) model to analyze students' academic performance, and combine Naive Bayes and J48 classifiers to classify students' performance. Kumar, Balamurugan, and Sasikala (2021) establish a multi-tier student performance evaluation model (MTSPEM) using a single classifier and an integrated classifier to evaluate student records, and conduct a comparative evaluation to prove the effectiveness of the proposed model. The analysis and evaluation of students' performance will not only help colleges and universities improve the quality of education, but also help enhance the overall performance and identify students at risk, which aim to optimize the management of educational resources (Mallik, Roy, Maheshwari, Pandey, & Rautray, 2019).

2.2. Prediction of learning behavior and dropout rate

In order to improve the accuracy of learning behavior prediction, some researchers utilize a variety of algorithms or models for comparison. Huang and Ning (2013) construct four types of mathematical models to predict students' academic performance, including multiple linear regression, multi-layer perceptron network, radial basis function network and support vector machine. Agrawal, Nigam, and Sahu (2018) apply two clustering algorithms to predict students' academic execution. In order to provide reliable admission standards for colleges and universities, Mengash (2020) uses ANN, decision tree, support vector machine and Naive Bayes to predict students' behavior. The results demonstrate that the accuracy of ANN is the highest. Turabieh et al. (2021) propose an improved Harris Hawks optimization algorithm to search the most valuable features in the student achievement prediction problem, and evaluate the whole prediction system using k-nearest neighbor, multilayer recurrent neural network, Naive Bayes and ANN. The results indicate that the combination accuracy of the optimization algorithm and multilayer recurrent neural network is the highest. Lee and Recker (2022) examine how student and instructor participation in online discussions impacts students' course performance. Multilevel modeling results show that online listening behaviors significantly predict students' course performance.

Some researchers also link other aspects related to learning to predict learning behavior. Przepiorka, Blachnio, Cudo, and Kot (2021) analyze the relationship between social anxiety and social skills by studying the use of smart phones to predict physical symptoms and academic performance. Zaffar, Hashmani, Habib, Quraishi, Irfan, et al. (2022) design a hybrid feature selection framework to identify important features and relevant features to predict students' performance. In order to reveal the problem of the internal relationship between questions and skills, Gao, Zhao, Li, Zhao, and Zeng (2022) propose a deep cognitive diagnosis framework, which enhances the traditional cognitive diagnosis methods to predict learning behavior through deep learning.

Some scholars have studied the prediction of dropout rate alone. Heredia, Amaya, and Barrientos (2015) apply C4.5 and ID3 to predict the possibility of dropping out, and compare the results of the two algorithms. In order to solve the problem that it is difficult to ensure the accuracy of manually extracted features, Lin, Liu, and Yi (2018) propose an integrated framework with feature selection to predict the dropout rate in massive online open courses (MOOCs), including feature generation, feature selection and dropout rate prediction. Two aspects of students' performance have been concerned, which are predicting students' academic performance and combining typical progress with prediction results. By focusing on a few courses with particularly good or poor performance, teachers can provide timely warning and support for students with poor performance, and provide advice and opportunities for students with good performance (Asif, Merceron, Ali, & Haider, 2017)

Most of the existing studies focus on the analysis and prediction of students' academic performance, and researchers usually predict learning behavior from the aspects which are relevant to academic performance. The purpose is to monitor students' academic performance in advance and find the space for students' progress, aiming to provide decision-making reference for the development of colleges and universities and to provide reference for innovating educational management. Existing studies either utilize clustering to analyze and evaluate academic performance, or apply a variety of classification algorithms to predict academic performance. On the basis of previous studies, this paper intends to comprehensively employ supervised and unsupervised learning technology to evaluate and predict learning behavior patterns. And different from the methods of extracting attributes previously, we select the significant attributes in the rotated component matrix as research variables by using principal component analysis in order to reduce dimension. We also compare the prediction accuracy, error and operation efficiency of different algorithms. At the same time, we use the results of data visualization to discuss and compare the application conditions of each classification algorithm and prediction model in the process of research.

3. Material and methods

3.1. Data

3.1.1. Data source

This study chooses a public educational data set from UCI (https://archive.ics.uci.edu/ml/datasets/Higher+Education+Students+Performance+Evaluation+Dataset). The data set is suitable for both the size of the data and the matching degree of the research content, and a public data set can be used for the readers to reproduce the experiment. There are 145 instances and 33 attributes in the data set. Except for the student number and course number, the remaining 31 attributes are divided into three parts: personal information, family information and information related to learning. Personal information has 10 attributes, family information has 6 attributes, and information related to learning has 15 attributes.

3.1.2. Data standardization

The student attribute matrix (Fan & Frederick, 2018) is used to quantify the student attribute for further analysis. There are no missing values in the data set, and the attribute characteristic is integer. Different attributes have different scales, so the data are preprocessed by dimensionless standardization.

3.2. Experiment and results

3.2.1. Attribute extraction

Feature selection (FS), also known as attribute extraction, is the most commonly used method in dimension reduction. It selects the proper subset of the feature from all the features of the original data in order to reduce the redundant information, noise, and the dimension of data analysis. It applies the proper subset of the feature to train the model constructed by learning algorithm, which aims to improve the learning performance of the algorithm. Attribute selection measures include information gain (e.g., ID3, C4.5), Gini index (e.g., SLIQ, SPRINT), and G-statistics. Considering the nonlinearity of data and the number of attributes and instances, the attributes are extracted with the help of principal component analysis (PCA). Before feature selection using PCA, we need to perform KMO and Bartlett's test in order to judge the data set whether to be suitable to use PCA. The result of the test is demonstrated in Table 1. The KMO and significance in the Bartlett test of sphericity (KMO > 0.5 and $P < 0.05$) indicate that the variables are highly correlated, which is enough to provide a reasonable theoretical basis for PCA.

Table 1
KMO and Bartlett's test.

Kaiser–Meyer–Olkin measure of sampling adequacy		0.568
Bartlett's test of sphericity	Approx. Chi-Square	1002.700
	df	465
	Sig.	0.000

According to the previous study, it is generally believed that the principal components with eigenvalues greater than 1 or cumulative contribution rate greater than 85% are representative. In the paper, we calculate both cumulative contribution rate and eigenvalue. We find that if we choose the principal components in which a cumulative contribution rate is greater than 85%, the final result almost covers all the principal components. In this situation, we cannot realize the goal of dimension reduction. Therefore, we select the principal components with eigenvalues greater than 1. Moreover, corresponding cumulative contribution rate reaches 63.660% that is acceptable to some extent. Table 2 shows 11 principal components with eigenvalues greater than 1 and cumulative contribution rate obtained by PCA.

The principal components are not the remaining variables after the original variables have been filtered, but the "comprehensive variables" after recombining the original variables. In the previous step, we reduce dimension by selecting the representative principal components calculated by PCA. These components are comprehensive variables which cannot replace original attributes that are direct illustration of learning behavior and will be used for the analysis of learning behavior pattern. Therefore, we make full use of rotated component matrix which is calculated from the previous step. We select the attributes with large positive load or small negative load in the rotation component matrix (Table 3) as the variables for subsequent research, which is functioned as extracting the attributes. The specific implementation condition is that if the absolute value of positive load or negative load in the row of each attribute is greater than 0.7, the attribute is extracted as the research variable. Table 3 shows the rotation component matrix composed of 11 principal components and 31 attributes. According to the definition of the matrix, the greater the absolute value of positive or negative load, the greater the correlation between the attributes of the row where the load resides and the principal component. In order to accurately study the learning behavior pattern and dimension reduction, we need to extract the highly correlated attributes. And considering the dimension of data analysis, the absolute value of the load is chosen as 0.7. Therefore, Partner, Study_hours, Read_frequency1, Read_frequency2, Preparation1, Preparation2, Take_notes, Listen, GPA and Expected_GPA are extracted as research variables. From the results of attribute extraction, nine attributes are attributes relevant to learning, and only one attribute is personal information. The use of personal statistics has no significant impact on the prediction accuracy (Tomasevic, Gvozdenovic, & Vranes, 2020), so the research variables extracted in this study are reasonable to certain extent.

3.2.2. Using a pseudo F statistic to determine the number of clusters

Clustering belongs to unsupervised learning. Clustering is not to categorize the data according to the existing rules, but to categorize the data by learning the law in the data. The result of clustering is closer to reality, which makes this method widely applied in various areas. In terms of the problem we solve, fast clustering, as a typical prototype clustering algorithm, is simple and fast, and has fast convergence of the objective function. These make the algorithm highly efficient in processing data sets. One of the disadvantages of fast clustering, also known as classic K-means clustering, is that the determination of cluster number is manually subjective. Thus, we add a step where a pseudo F

Table 2
Eigenvalues and contribution rate of principal components.

Principal components	Extract sums of squares loading		
	Eigenvalues	Variance contribution rate/%	Cumulative contribution rate/%
1	3.083	9.945	9.945
2	2.559	8.254	18.199
3	2.337	7.538	25.737
4	2.090	6.742	32.479
5	1.854	5.982	38.461
6	1.569	5.060	43.521
7	1.398	4.510	48.031
8	1.328	4.282	52.314
9	1.256	4.052	56.366
10	1.182	3.812	60.178
11	1.079	3.482	63.660

Table 3
Rotated component matrix.

	1	2	3	4	5	6	7	8	9	10	11
Age	−0.131	0.268	0.548	0.041	−0.020	−0.007	−0.217	0.328	−0.267	−0.299	0.192
Sex	−0.026	0.279	0.553	−0.048	−0.103	−0.002	0.079	0.114	0.307	0.420	−0.104
High_school	−0.076	0.043	0.189	0.004	−0.111	−0.078	0.004	0.690	0.209	−0.182	0.066
Scholarship	−0.022	0.371	−0.691	0.084	−0.125	−0.002	0.092	−0.065	0.085	0.012	0.093
Work	0.068	0.026	0.000	−0.179	−0.357	−0.206	0.019	−0.247	0.159	0.558	0.136
Activity	−0.256	−0.006	0.239	−0.548	−0.094	−0.113	0.061	0.200	−0.185	0.295	−0.113
Partner	−0.139	−0.038	−0.170	0.107	0.018	−0.277	−0.070	−0.047	−0.094	0.102	0.749
Salary	0.031	0.028	0.151	−0.086	0.007	0.501	−0.201	−0.147	−0.442	−0.025	0.043
Transportation	0.626	0.095	−0.018	−0.204	0.138	−0.311	−0.154	0.176	−0.084	0.026	0.126
Accommodate	0.637	−0.063	−0.112	−0.121	0.117	−0.018	0.071	−0.090	0.125	−0.407	−0.152
M_education	0.575	−0.068	0.195	0.156	−0.117	0.275	0.000	−0.255	0.039	0.139	0.120
F_education	0.447	−0.282	0.406	0.083	0.011	0.021	0.102	−0.342	0.114	0.126	0.231
Number_sb	−0.677	0.085	0.115	−0.063	0.202	−0.059	0.072	−0.116	0.024	−0.163	−0.061
Parent_status	−0.071	0.265	0.085	0.218	0.406	0.189	−0.466	0.002	−0.080	0.041	−0.183
M_occupation	0.676	0.050	−0.118	−0.043	−0.044	0.179	0.101	−0.009	−0.045	−0.005	−0.215
F_occupation	−0.084	−0.028	−0.086	0.162	0.052	−0.232	−0.146	−0.205	−0.025	0.079	−0.684
Study_hours	−0.068	−0.088	−0.073	0.242	0.716	−0.301	0.105	−0.143	0.031	0.096	−0.060
R_frequency1	−0.064	−0.014	0.123	0.769	0.111	−0.124	−0.005	−0.042	0.091	0.041	−0.007
R_frequency2	−0.072	0.070	−0.063	0.724	0.047	0.082	0.076	0.229	−0.117	0.029	−0.071
Attendance	0.152	−0.146	−0.552	0.024	0.189	−0.018	0.308	0.215	−0.212	0.008	0.032
Impact	0.096	−0.141	−0.178	0.113	−0.002	0.213	0.037	0.526	−0.170	0.102	0.112
Attend_class	−0.088	−0.129	0.108	0.103	−0.261	−0.080	−0.030	−0.112	−0.601	0.106	0.060
Preparation1	0.158	−0.102	−0.082	0.023	0.077	0.761	−0.048	0.119	0.147	−0.025	−0.065
Preparation2	−0.084	0.199	0.004	0.002	0.722	0.400	0.063	−0.026	0.103	0.005	0.090
Take_notes	0.093	0.082	−0.094	0.071	0.135	−0.146	0.760	0.140	0.118	0.037	−0.008
Listen	−0.073	0.018	0.031	−0.048	−0.200	−0.043	−0.002	0.035	0.109	−0.749	0.027
Discussion	−0.144	0.299	−0.060	−0.023	0.027	0.180	0.550	−0.310	0.037	0.006	0.052
Flip_class	−0.178	0.177	0.507	0.140	−0.020	−0.029	0.455	−0.021	−0.151	−0.023	−0.033
GPA	−0.048	0.824	0.073	−0.096	0.031	−0.034	0.095	−0.074	0.178	−0.033	0.041
Expect_GPA	0.013	0.860	−0.031	0.140	0.090	−0.058	0.077	0.037	0.101	0.054	−0.059
Grade	−0.086	0.244	0.178	0.167	−0.153	0.061	0.035	−0.150	0.680	0.041	0.022

statistic is proposed to determine the cluster number objectively in the algorithm, and the calculation equation is shown in (1).

$$F = \frac{\left(T - P_k\right) / (k-1)}{P_k / (n-k)} \tag{1}$$

T represents the sum of squares of total deviations. P_k is the sum of squares of intra class deviations when data are clustered into k classes. n represents the size of the sample. The pseudo F statistic is used to evaluate the effect of clustering into k classes. If clustering effect is good, the sum of squares of deviations between classes is larger than the sum of squares of deviations within classes, so the clustering level with large pseudo F statistics and small number of clusters should be taken.

Taking 10 research variables as inputs for fast clustering, the output intra-class distance and the number of known instances can be utilized to calculate the pseudo statistic corresponding to each cluster number. The number of clusters k starts from 2 as input. The general principle is $k_{max} \le \sqrt{n}$ (Ramze Rezaee, Lelieveldt, & Reiber, 1988), and we take $k_{max} = \sqrt{n}$. n is equal to 145, so, the maximum value k is 12. However, there are only 10 research variables, and it is generally considered that the number of clusters does not exceed the number of variables. Therefore, the range of cluster number k in this study is $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. We calculate the pseudo F statistics for the number of clusters within this range, and the results are shown in Fig. 1. Based on the principle that the clustering level with large pseudo F statistics and small number of clusters should be taken, it can be found from Fig. 1 that when the number of clusters is 5, the number of clusters is small and the pseudo F statistic is large. Therefore, it is more reasonable for the data to be clustered into 5 classes. At this time, the number of iterations is 7 and the sum of squares of intra-class deviations is 84.76. 145 instances are clustered into five classes. The detailed clustering results show that there are 56 instances in the first class, 26 instances in the second class, 10 instances in the third class, 14 instances in the 4th class and 39 instances in the 5th class.

3.2.3. *Classifiers*

Classification belongs to supervised learning. Supervised learning trains labeled data, realizes mapping from input to output, and then applies this mapping relationship to unknown data to achieve classification and prediction. The biggest difference between supervised learning and unsupervised learning is whether the data is labeled or not. Therefore, if the supervised learning algorithm is to be trained

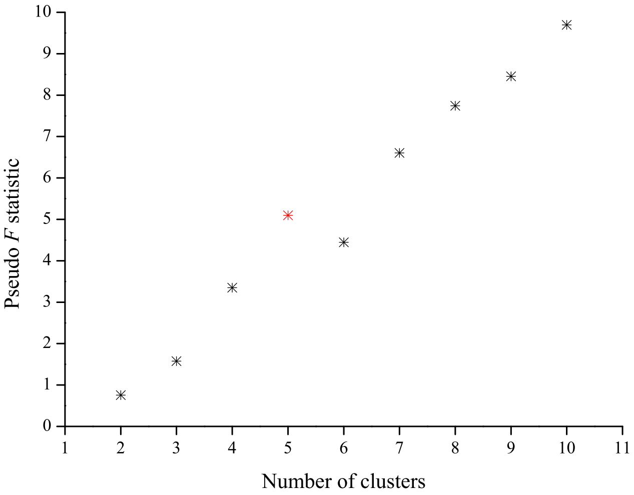


Fig. 1. Pseudo F statistics corresponding to different cluster numbers.

the model, the original data must be labeled or labeled in some way. Supervised learning falls into two main categories: Classification and regression. For discrete data, classification is suitable; For continuous data, regression is suitable. The data in this paper is discrete, so only the classification algorithm is applied. Due to the need for a labeled data set to interpret the analyzed data samples, supervised machine learning technology is considered to be more suitable for prediction tasks in educational data mining than unsupervised machine learning technology (Tomasevic et al., 2020).

Classifier, as the name suggests, is the implementation of classification tasks. There is a simple example to illustrate how to achieve prediction task by applying a classifier. Students who pass are classified as the first category, while students who fail are classified as the second category. In order to focus on those who fail, it is necessary to classify students based on historical performance data. The classification criteria of the classifier are that a student whose score is greater than or equal to 60 (in a percentage system) is the first category, and a student whose score is less than 60 is the second category. The result of classification is a prediction of students' future academic performance. If one is divided into the second category, the student needs to be paid more attention to. At this time, the classifier plays a role of early warning.

There is no class label on the metadata. The previous clustering results are applied to add labels to the preprocessed data (Feng, Fan and Chen, 2022), so that the data can be trained with the classification algorithm. The classification models are constructed based on six algorithms: J48, K-nearest neighbor (KNN), Bayes Net, Random Forest, Support vector machine (SVM) and Logit Boost. Among them, Random Forest and Logit Boost are integrated algorithms. WEKA integrates a large number of machine learning algorithms that can undertake data mining tasks. The experiments are based on WEKA, and the parameters of the six classification algorithms are default. Several classification algorithms will be briefly introduced below.

- **J48**: J48 is an improved algorithm of ID3. Improvements are made in the following four aspects: (1) The information gain rate is used to select attributes. (2) Pruning is performed during tree construction. (3) The discrete processing of continuous attributes can be completed. (4) The incomplete data can be processed. When constructing a tree, the algorithm requires multiple sequential scanning and sorting of the data set, resulting in low efficiency. However, due to its easy to understand classification rules, the accuracy is high.
- **KNN**: Cover and Hart (1953) proposed a KNN classification algorithm based on distance. KNN algorithm, also described as reference sample plot method, determines the distance between the samples to be classified and each training sample, and then chooses the k samples which are closest to the samples to be classified as the k nearest neighbors of the samples to be classified. If most of individuals among k similar samples in the feature space belong to one class, the sample also belongs to this class (Goldberger, Roweis, Hinton, & Salakhutdinov, 2005). As a typical non-parametric method, calculations can be performed by simply searching for similar units. Even if the system is linearly indivisible, this method can still be applied.
- **Bayes Net**: When the nodes, the states of nodes and the connections between nodes increase, the calculation of simple Bayes Net is very complex and the calculation of probability propagation becomes very heavy, which limit the application of Bayes Net in practice. Until (Pearl, 1986) proposed the message passing algorithm (polytree algorithm), and Lauritzen and Spiegelhalter (1988) further proposed the junction tree algorithm using the concept of message passing, it provided an effective algorithm for the probability propagation of Bayes Net and laid the foundation for practical application. The drawbacks of this classification method are how to effectively calculate probability when the model becomes complex, and how to handle continuous variables, which are research directions in recent years.
- **Random Forest**: Random Forest integrates the characteristics of Bagging series algorithms and random selection, and introduces random attribute selection for prediction in the training process of decision tree, which was proposed by Breiman (2001) . The algorithm adopts the put-back sampling strategy to extract samples from the original data set, and uses the non-put-back sampling strategy to extract different features as input variables. It constructs a decision tree on each new data set, and synthesizes the prediction results of multiple decision trees as the prediction results of the whole Random Forest. Random Forest has a good tolerance for outlier and noise, and is not easy to overfit. It is widely used in medicine, bio-informatics, management and other fields.
- **SVM**: Cortes and Vapnik (1995) formally proposed support vector machine. SVM is a supervised binary classifier based on VC (Vapnik–Chervonenkis) dimension theory of statistics and the principle of structural risk minimization. SVM can automatically find those support vectors that have better discrimination ability for classification by training. The classifier constructed from above theory can maximize the interval between classes, so it has better adaptability and higher discrimination ability. SVM has a deep theoretical foundation, which can ensure that the extremal solution is the global optimal solution rather than the local optimal solution. Thus, SVM has good generalization ability for unknown samples.
- **Logit Boost**: Boosting was first proposed by Schapire (1989) . Because it required priori knowledge of the performance of weak learnability, its application was limited. Freund and Schapire proposed an improved method: adaptive boosting, called AdaBoost, which did not require prior knowledge of weak learnability (Freund & Schapire, 1997). Friedman, Hastie, and Tibshirani (2000) lately improved AdaBoost: logit adaptive boosting, called Logit Boost. AdaBoost algorithm adopts exponential loss function, while Logit Boost adopts negative log likelihood loss function. Logit Boost constructs a basic weak classifier on the existing sample set, repeatedly calls the weak classifiers, and gives more weights to the samples with wrong classification each time. These weak classifiers are weighted and synthesized to obtain a strong classifier. The superposition model is fitted through maximum likelihood estimation in order to obtain a classification model with higher precision in a strong classifier. The main idea is that

Table 4
The confusion matrix.

Actual category	Prediction category	
	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

the strong classifier is made up of an army of weak classifiers (Demirer, Pierdzioch, & Zhang, 2017; Goessling, 2017; Kanamori & Takenouchi, 2013).

3.2.4. Evaluation metrics of classifying

In order to effectively evaluate the classification results, Kappa coefficient, accuracy, precision, recall, F-Measure, Youden index, receiver operator characteristic (ROC) curve and precision–recall (P–R) curve are used to evaluate the results from the performance of the classifier.

10-fold cross validation is applied in the training process with the purpose of avoiding the influence of over-fitting on the results and check the classification effect. The advantage of 10-fold cross validation is that it allows all data to participate in the training and testing process, fully reflecting the characteristics of “crossover”. The basic idea of 10-fold cross validation is to randomly divide the data into 10 parts and conduct 10 experiments. Each time, 9 parts are used as training data, and the remaining part is used as test data. The experiment will repeat 10 times until all the data is used as the training set for validation. The average value of each metric in the 10 experiments is used as the final result. After training data, the classification results are compared with the class label. The evaluation metrics of the six classifiers are calculated, and corresponding metrics are compared. The confusion matrix based on the binary classification problem is applied to calculate each metric. The confusion matrix is shown in Table 4. True positive (TP) represents the number of positive samples that are judged as positive samples. True negative (TN) means the number of negative samples that are judged as negative samples. False positive (FP) denotes the number of negative samples that are judged to be positive samples. False negative (FN) indicates the number of positive samples that are judged to be negative samples.

In classification problems, the most common evaluation metric is accuracy, which can directly reflect the correct proportion. However, in practice, the sample size of each category is often unbalanced. Without adjustments on such unbalanced data sets, the model is prone to bias toward larger categories and abandon smaller ones. At this point, a metric that penalizes the model’s “bias” is necessary. Kappa coefficient is utilized to evaluate the difference between the classification results of the classifier and the random classification. The value range is [−1,1]. Kappa value is positively correlated with the accuracy of the classifier. The closer the value is to 1, the more accurate the algorithm is. The calculation equation of Kappa coefficient based on confusion matrix is shown in (2). p is calculated as shown in Eq. (3) which denotes that the numerator is the sum of the diagonal elements, and the denominator is the sum of all the elements (in fact, p is the accuracy). q is calculated as shown in Eq. (4) which means that the numerator represents the sum of elements in column r multiplied by the sum of elements in row r and then all products are summed up, and the denominator represents the square of the sum of all elements, which is the sum of the “product of the actual quantity and the predicted quantity” corresponding to all the categories respectively, and divided by the “square of the total number of samples”. According to the calculation of Kappa, the more unbalanced the confusion matrix, the higher the q, the lower the Kappa, which can achieve the goal of punishing the model with strong “bias”.

$$Kappa = \frac{p-q}{1-q} \tag{2}$$

$$p = \frac{\sum_{i=j} f_{ij}}{\sum_i \sum_j f_{ij}} = \frac{TP+TN}{TP+FN+FP+TN} \tag{3}$$

$$q = \frac{\sum_r \left(\sum_i f_{ir} \cdot \sum_j f_{rj}\right)}{\left(\sum_i \sum_j f_{ij}\right)^2} = \frac{(TP+FN)(TP+FP)+(FP+TN)(FN+TN)}{(TP+FN+FP+TN)^2} \tag{4}$$

The accuracy, precision, recall (also called sensitivity), and F-Measure range from 0 to 1, and the calculation equations are shown in (5), (6), (7) and (8), respectively. With regard to F-Measure in the paper, α is taken as 1. Accuracy is a direct evaluation metric, which means the ratio of correctly classified samples to all samples. This is an easy to understand metric, but there is an obvious problem in case of imbalanced samples. For example, when people predict whether an earthquake will occur in a certain area on a certain day, the occurrence of an earthquake is classified as 0 and the non occurrence is classified as 1. If a classifier classifies all test cases into 1, the accuracy is high. But once an earthquake occurs, the loss caused by the classification results will be difficult to estimate. This indicates that accuracy is not a comprehensive and scientific metric. Compared with the calculation equations of precision and recall, it can be seen that in some situations these two metrics contradict each other. The F-Measure, as a combination of these two metrics, is the harmonic average of precision and recall when α is taken as 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} = Sensitivity \tag{7}$$

$$F-Measure = \frac{\left(\alpha^2+1\right)\times Recall \times Precision}{\alpha^2(Recall+Precision)} \tag{8}$$

True positive rate (TPR) is the ratio of positive samples determined as positive samples by the classifier, also known as sensitivity. In effect, it is the recall of the ‘positive category’. True negative rate (TNR) is the ratio of negative samples determined as negative samples by the classifier, also known as specificity. Actually, it is the recall of the ‘negative category’. False negative rate (FNR) is the ratio of positive samples determined as negative samples by the classifier. Some researchers also call it false reject rate (FNR). False positive rate (FPR) is the ratio of negative samples determined as positive samples by the classifier. Some researchers also call it false acceptance rate (FAR). It was originally negative, but it is recognized as positive. The calculation equations of TPR, TNR, FNR and FPR are shown in (9), (10), (11) and (12), respectively.

$$TPR = \frac{TP}{TP+FN} = Sensitivity \tag{9}$$

$$TNR = \frac{TN}{FP+TN} = Specificity \tag{10}$$

$$FNR = \frac{FN}{TP+FN} = 1 - Sensitivity \tag{11}$$

$$FPR = \frac{FP}{FP+TN} = 1 - Specificity \tag{12}$$

TPR and FPR which are two indicators of ROC curve do not depend on the specific distribution of the class. The curve takes the positive samples and negative samples into account, and will not change significantly with the change of the proportion of positive and negative samples. It is an effective tool for more balanced research on classification problems. TPR and FPR can be represented by sensitivity and (1-specificity), so two coordinate axes of the curve can also be composed of sensitivity and (1-specificity). The closer the ROC curve is to the upper left, the better the performance of the classifier is. When two ROC curves intersect, it is impossible for us to directly see which has better performance. At this time, it is necessary to compare the

Table 5
The specific values of seven metrics.

	Kappa	Accuracy	Precision	Recall	F-Measure	AUC	Youden index
J48	0.7291	0.800	0.813	0.800	0.804	0.910	0.737
KNN	0.7723	0.834	0.842	0.834	0.834	0.887	0.775
Bayes Net	0.8475	0.890	0.889	0.890	0.887	0.983	0.850
Random Forest	0.8566	0.897	0.898	0.897	0.894	0.990	0.854
SVM	0.8653	0.903	0.908	0.903	0.895	0.968	0.865
Logit Boost	0.8655	0.903	0.903	0.903	0.901	0.986	0.864

size of area under curve (AUC). The calculation equation of AUC is shown in (13). AUC is the area enclosed by the coordinate axis under the ROC curve. The larger the AUC, the better the performance of the classifier. When AUC = 0.5, the classifier is the same as random prediction, which is similar to coin tossing and the probability of both sides is 50%. When 0 < AUC < 0.5, the use of the classifier is worse than random prediction. In this situation, only reverse prediction is better than random prediction. When AUC>0.5, the classifier is better than random prediction. When AUC = 1, the classifier at this time is a perfect classifier. Generally, such a classifier cannot exist.

$$AUC = \frac{TPR}{FPR} = \frac{Sensitivity}{1 - Specificity} \tag{13}$$

The two indicators related to the Youden index are sensitivity and specificity. The sensitivity and specificity can be represented by TPR and FPR, so the Youden index can also be calculated by TPR and FPR. The range of the index is [−1,1]. The larger the index value, the better the performance of the classifier in predicting positive examples. When the index value is negative, it has no application value. The Youden index can be used in conjunction with the ROC curve. The geometric representation of the Youden index maximizes the vertical distance from the point on the ROC curve to the x-axis to ensure that the TPR is large while the FPR is as small as possible. The equation for calculating the Youden index is shown in (14). Although it is technically possible to obtain a value less than 0 from this equation, a value less than 0 only implies that the positive and negative labels have been switched.

$$\text{Youden index} = Sensitivity + Specificity - 1 = TPR - FPR \tag{14}$$

Kappa, accuracy, precision, recall, F-Measure, AUC and Youden index of the classifiers constructed by the six algorithms are shown in Table 5. From the Kappa value alone, except for J48 and KNN, other classifiers perform well, among which Logit Boost performs best. In order to clearly compare the performance of each classifier in terms of each metric, seven metrics are shown in Fig. 2. The polygonal line reflects the performance of the six classifiers in predicting positive examples, and value corresponding to the point on the polygonal line is the Youden index of each classifier. Histograms reveal six metrics of six classifiers. The larger the seven evaluation metrics in Fig. 2, the better the performance of the classifier. J48 and KNN perform poorly in terms of Kappa, accuracy, precision, recall and F-Measure. In terms of AUC, J48, Bayes Net, Random Forest, SVM and Logit Boost exceed 0.9. Bayes Net, Random Forest, SVM and Logit Boost exceed 0.95, among which Random Forest performs best. The Youden index of the six classifiers is between 0.7 and 0.9, and the Youden index of SVM is the largest. Although the AUC of SVM is smaller than that of Bayes Net, Random Forest and Logit Boost, it has the best performance in predicting positive examples. Therefore, if we pay more attention to the rate of the correction in the problem of predicting academic risk, we should not choose J48 and KNN.

The ROC curves of the classifiers constructed by the six algorithms under five classes are shown in Fig. 3. Compared with (a), (b) and (e), the ROC curves of Bayes Net and Logit Boost are relatively smooth and the change range of each class is small in (c) and (f). The area enclosed by the coordinate axis under the ROC curves of Random Forest and Logit Boost in (d) and (f) are larger, indicating that the performance of these two classifiers is better. Therefore, when we consider the positive samples and negative samples at the same time, we had better utilize Random Forest and Logit Boost to construct classifiers to predict results.

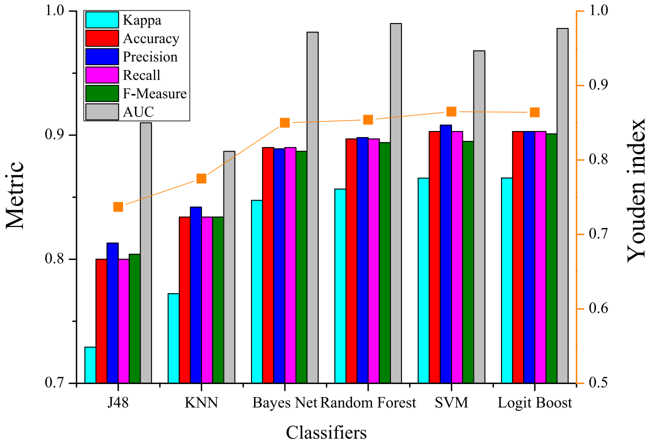
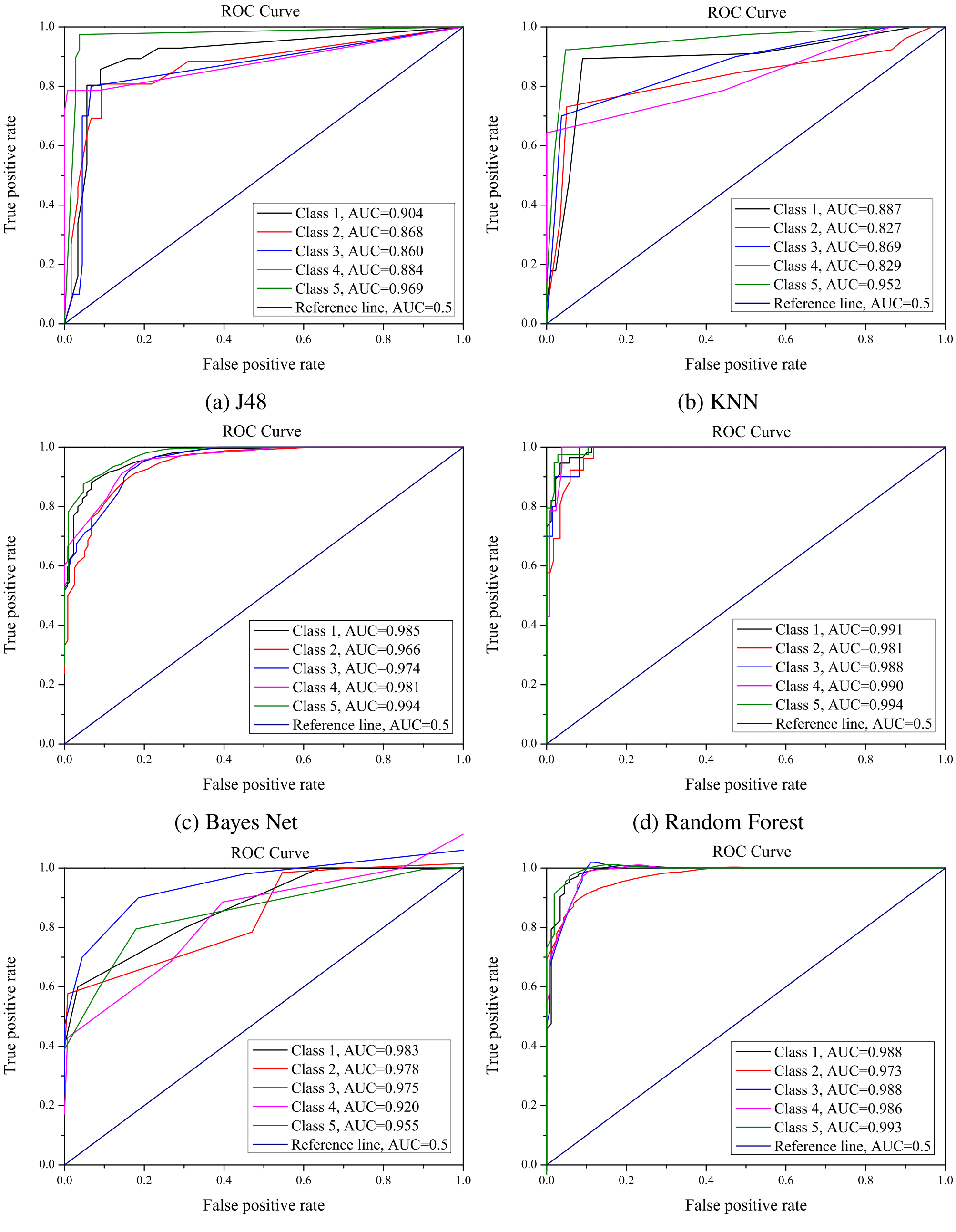


Fig. 2. Seven evaluation metrics of six classifiers.

When the classification result is unbalanced, due to the strong robustness of the ROC curve, the ROC curve may not change significantly, but it has a great impact on the model that attaches importance to accuracy. Precision and recall, which are two indicators of P–R curve, focus on positive samples. When positive samples are more important in class imbalance, P–R curve is better than ROC curve. If the P–R curve of a classifier is more convex to the upper right, it means that the performance of the classifier is better. If the P–R curve of one classifier is completely covered by the P–R curve of another classifier, the latter has better performance. If there is an intersection between the two curves, the area under the curve and the coordinate axis or the F-Measure can be used for comparison. The larger the area or the larger the F-measure value, the better the performance of the classifier. In addition to comparing the performance through calculation, one of the most important advantages of P–R curve is that the performance of classifier can be seen intuitively. Break-even point (BEP) is the value when precision is equal to recall. The larger the value corresponding to the intersection of P–R curve and the line that precision is equal to recall, the better the performance of the classifier.

The P–R curves formed by the six classifiers in the training process are shown in Fig. 4. The intersection values of the six curves and the line composed of BEP from small to large are J48, KNN, Bayes Net, SVM, Random Forest and Logit Boost respectively. The intersection values of Random Forest and Logit Boost coincide. The classifiers constructed by Random Forest and Logit Boost algorithm have the best performance. Therefore, if positive samples are more important in imbalanced problems, we could apply Random Forest and Logit Boost in the problem of predicting academic risk.

Integrated algorithms are divided into two categories according to whether there are dependencies between base classifiers: Bagging series algorithms without dependencies between base classifiers and Boosting series algorithms with dependencies between base classifiers. Random



(a) J48

(b) KNN

(c) Bayes Net

(d) Random Forest

(e) SVM

(f) Logit Boost

Fig. 3. ROC curves of six classifiers under five classes.

Forest is one of Bagging series algorithms, and Logit Boost is one of Boosting series algorithms. From the above results, it can be concluded that the performance of two integrated algorithms is better than that of a single classifier, but in terms of error and running time, the Logit Boost algorithm with dependency between base classifiers performs better. We also summarize the applicable conditions of each classifier in the actual prediction of academic performance according to the different expression forms of metrics.

4. Discussion

4.1. How to study learning behavior patterns from the perspective of educational data mining?

EDM is dominated by data mining, machine learning, statistics and other methods. With the development of large open online data sets and data mining technology, the field of EDM has attracted more and more attention. In this paper, we choose a public educational data set from UCI which is famous for data mining. In order to extract the features with great significance, we utilize PCA, and eigenvalues and contribution rate of principal components are shown in Table 2. However, we extract research variables from rotated component matrix in Table 3 instead of directly adopting the principal components as research variables. A pseudo statistic is proposed to determine cluster number objectively in Fig. 1 that is applied to add labels to metadata. The classification algorithm is applied to construct the prediction model with the extracted research variables. Seven evaluation metrics have a sharp contrast in Fig. 2. The intersections between BEP line and P–R curves which pay more attention to positive samples are shown in Fig. 4. According to different metrics, we summarize the application conditions which are convenient for the education manager to select the appropriate classifier to mine and analyze the educational data set

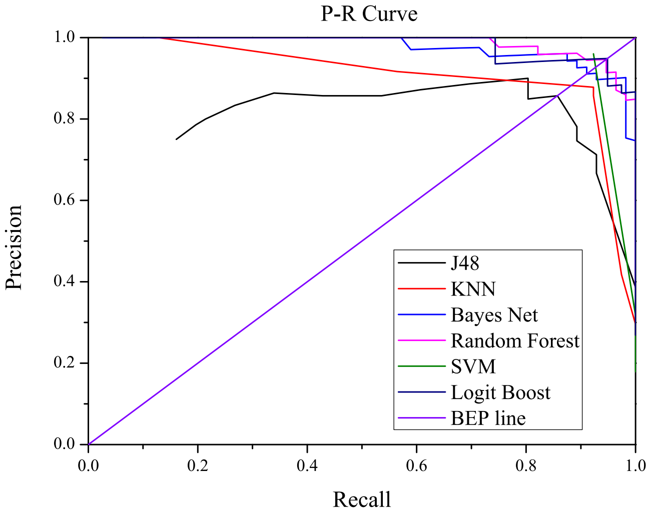


Fig. 4. P–R curves of six classifiers.

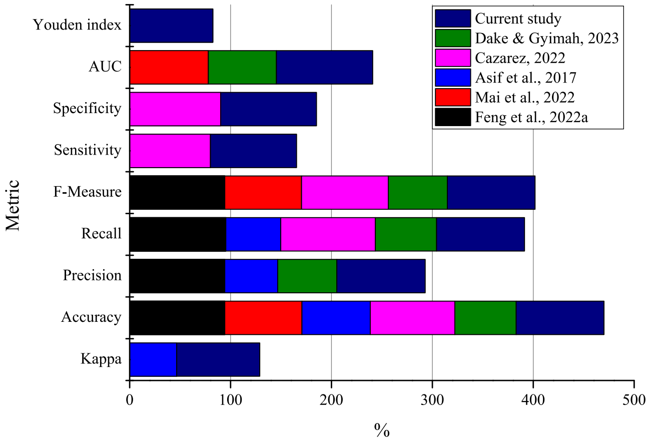


Fig. 5. Comparison between 5 previous works and current study.

in terms of the size of the educational data set, time and required accuracy in the actual situation. We compare the results of five works with the results of this paper, as shown in Fig. 5. Dake and Gyimah (2023), Feng, Fan and Ao (2022) and current study all employed accuracy, precision, recall, and F-Measure. It is found that accuracy, precision, recall, and F-Measure are used more frequently than Kappa, sensitivity, specificity, and AUC. Youden index is only used in this paper. In terms of Kappa, current study is obviously superior to the result from Asif et al. (2017) . With regard to sensitivity and specificity, current study is slightly better than (Cazarez, 2022). Compared with Dake and Gyimah (2023) and Mai et al. (2022), the performance of current study has a great advantage about AUC. Although not every metric is optimal, current study outperforms others in most metrics.

4.2. What aspects are included in the research on learning behavior patterns?

The public data set selected in this study contains not only information involved in learning, but also students' personal information and family information. Although personal information and family information are not directly related to learning behavior patterns, they will affect learning behavior patterns to some degree. In the results of attribute extraction in Table 3, only one attribute related to personal information is significant. The above is the preparation for studying the learning behavior patterns.

- As the name of the data set indicates, the 1st aspect of the research on learning behavior patterns is to evaluate learning behavior patterns. By visualizing the results of cluster number in Fig. 1, we can objectively determine students' learning groups in the data set. We can give students corresponding ranks according to the actual examination situation, instead of getting "excellent" only when they score more than 90 in the hundred mark system. We can also discover students' learning group preferences by analyzing the data in the same cluster.
- The 2nd aspect of research on learning behavior patterns in this paper is to predict the academic performance in the future through the analysis of the existing learning behavior. As can be seen from the title of the paper, one of the purposes of our research is to predict the students' academic performance based on their historical performance. 'Prediction' can be realized by classifying. In fact, PCA, clustering, and cross validation are all served for classifying. Firstly, PCA is applied to reduce dimension to extract more relevant attributes. Secondly, these extracted attributes as research variables are input aiming at clustering. Clustering results are used as the label of data. Then, labeled data can be classified in order to realize the goal of predicting. Finally, cross validation is utilized to obtain some metrics with the purpose of evaluating the effect of classifying. By using machine learning algorithms to build prediction classifiers, the evaluation metrics of different classifiers are displayed in Figs. 2, 3, and 4. The applicable conditions of different classifiers are compared and concluded according to the above evaluation metrics.
- The 3rd aspect of the research runs through the whole research process. We can instantly see the changing trend of the number of clusters and a pseudo-statistic proposed in this paper in Fig. 1, which aims to determine the number of clusters. Figs. 2, 3 and 4 clearly show the comparison of various performance metrics of each classifier. The teaching manager can choose the corresponding algorithm for prediction according to a certain aspect that should be taken into account when actually predicting academic performance.

4.3. How to innovate educational management by studying learning behavior patterns?

Simple statistical analysis methods are no longer suitable for analyzing the growing educational data sets. Therefore, in order to optimize educational management, we must innovate the analysis method radically. Data mining can discover the information hidden behind massive data, so educational data mining, which is a growing field, came into being. The research on learning behavior patterns is the most direct embodiment of educational data mining. Table 3 illustrates that personal information does not significantly affect the academic performance of students, so stakeholders had better attach more importance to attributes which are relevant to learning. According to quantitative standards in Eq. (1) rather than subjective score segmentation criteria, we determine the cluster number. When teachers pay more attention to the evaluation metrics of classifiers performance, they should avoid using J48 and KNN in Fig. 2. From Fig. 3, we can find that AUC of Random Forest and Logit Boost is larger, which indicates that the models constructed by these two classifiers perform better. If the classification result is unbalanced and positive samples are more important in class imbalance, we recommend to use Random Forest and Logit Boost, as can be seen in Fig. 4. By using data mining methods to study learning behavior patterns, we can extract the indirect information behind educational data sets. It is not only helpful for teaching managers to evaluate students' learning behavior patterns objectively, but also conducive to timely helping students who may be at risk in the future and encouraging the progress and development of excellent students.

5. Summary and prospect

With the help of data mining technology, this paper studies the learning behavior patterns from three aspects: evaluation, prediction and visualization. Now we summarize four places different from previous studies in the paper. First, in order to add labels to the data, clustering is performed and a pseudo statistic is proposed to determine the number of clusters objectively which avoids the arbitrariness of subjectivity. Second, although PCA is applied to extract attributes, the final selection of research variables is not the comprehensive variables extracted by PCA, but the attributes with greater significance in the rotating component matrix, which reduces the dimension of data analysis to certain extent and can improve the efficiency of later prediction because of the reduction of the number of attributes. Third, when using different classification algorithms to construct prediction classifiers, it is found that the accuracy, efficiency and error of the integrated classifier with dependent base classifier are better than that of a single classifier. The visualization of the analysis results makes the "calm" data set show intuitive and clear information, which is not only beneficial to the innovation of educational management, but also promote the development of students. Fourth, in the application of classification algorithms, we summarize the applicable conditions of different algorithms. When calculating the evaluation metrics of the classifier, we also summarize the applicable situation according to different visual forms.

The existing research and this paper all analyze, evaluate and predict learning behavior with different functional modules. In the future, we can consider designing an integrated framework to integrate data preprocessing technology, unsupervised learning technology, supervised learning technology and data visualization technology into EDM. As long as education managers input data, they can get the output of each functional module. Using the output of different functional modules can realize the idea of EDM proposed at the beginning, which is to serve education with technology and to develop education with innovation.

CRediT authorship contribution statement

Guiyun Feng: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Muwei Fan:** Writing – review & editing, Supervision, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Agrawal, S., Nigam, S., & Sahu, K. (2018). Prediction of students academic execution using K-Means and K-Medoids clustering technique. In *2018 2nd international conference on trends in electronics and informatics (ICOEI)* (pp. 1308–1315). USA: IEEE, http://dx.doi.org/10.1109/ICOEI.2018.8553747.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177–194. http://dx.doi.org/10.1016/j.compedu.2017.05.007.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, *23*(1), 537–553. http://dx.doi.org/10.1007/s10639-017-9616-z.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Busalim, A. H., Masrom, M., & Wan, N. (2019). The impact of facebook addiction and self-esteem on students' academic performance: A multi-group analysis. *Computers & Education*, *142*, Article 103651. http://dx.doi.org/10.1016/j.compedu.2019.103651.

Cazarez, R. L. U. (2022). Accuracy comparison between statistical and computational classifiers applied for predicting student performance in online higher education. *Education and Information Technologies*, *27*(8), 11565–11590. http://dx.doi.org/10.1007/s10639-022-11106-4.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. http://dx.doi.org/10.1007/BF00994018.

Cover, T. M., & Hart, P. E. (1953). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. http://dx.doi.org/10.1109/TIT.1967.1053964.

Crivei, L. M., Czibula, G., Ciubotariu, G., & Dindelegan, M. (2020). Unsupervised learning based mining of academic data sets for students' performance analysis. In *2020 IEEE 14th international symposium on applied computational intelligence and informatics (SACI)* (pp. 11–16). USA: IEEE, http://dx.doi.org/10.1109/SACI49304.2020.9118835.

Dake, D. K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. *Education and Information Technologies*, *28*(4), 4629–4647. http://dx.doi.org/10.1007/s10639-022-11349-1.

Delgado, S., Morán, F., José, J. C. S., & Burgos, D. (2021). Analysis of students' behavior through user clustering in online learning settings, based on self organizing maps neural networks. *IEEE Access*, *9*, 132592–132608. http://dx.doi.org/10.1109/ACCESS.2021.3115024.

Demirer, R., Pierdzioch, C., & Zhang, H. (2017). On the short-term predictability of stock returns: A quantile boosting approach. *Finance Research Letters*, *22*(3), 35–41. http://dx.doi.org/10.1016/j.frl.2016.12.032.

Fan, Y., & Frederick, W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, *123*, 97–108. http://dx.doi.org/10.1016/j.compedu.2018.04.006.

Feng, G., Fan, M., & Ao, C. (2022). Exploration and visualization of learning behavior patterns from the perspective of educational process mining. *IEEE Access*, *10*, 65271–65283. http://dx.doi.org/10.1109/ACCESS.2022.3184111.

Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, 19558–19571. http://dx.doi.org/10.1109/ACCESS.2022.3151652.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. http://dx.doi.org/10.1006/jcss.1997.1504.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, *38*(2), 337–374.

Gao, L., Zhao, Z., Li, C., Zhao, J., & Zeng, Q. (2022). Deep cognitive diagnosis model for predicting students' performance. *Future Generation Computer Systems*, *126*, 252–262. http://dx.doi.org/10.1016/j.future.2021.08.019.

Goessling, M. (2017). LogitBoost autoregressive networks. *Computational Statistics & Data Analysis*, *112*(3), 88–98. http://dx.doi.org/10.1016/j.csda.2017.03.010.

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2005). Neighbourhood components analysis. *Advances in Neural Information Processing Systems*, *17*, 513–520.

Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student dropout predictive model using data mining techniques. *IEEE Latin America Transactions*, *13*(9), 3127–3134. http://dx.doi.org/10.1109/TLA.2015.7350068.

Huang, S., & Ning, F. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers & Education*, *61*, 133–145. http://dx.doi.org/10.1016/j.compedu.2012.08.015.

Kanamori, T., & Takenouchi, T. (2013). Improving Logitboost with prior knowledge. *Information Fusion*, *14*(2), 208–219. http://dx.doi.org/10.1016/j.inffus.2011.11.004.

Karthikeyan, V. G., Thangaraj, P., & Karthik, S. (2020). Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. *Soft Computing*, *24*(24), 18477–18487. http://dx.doi.org/10.1007/s00500-020-05075-4.

Kumar, E., Balamurugan, S., & Sasikala, S. (2021). Multi-tier student performance evaluation model (MTSPEM) with integrated classification techniques for educational decision making. *International Journal of Computational Intelligence Systems*, *14*(1), 1796–1808. http://dx.doi.org/10.2991/ijcis.d.210609.001.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *50*(2), 157–224.

Lee, J.-E., & Recker, M. (2022). Predicting student performance by modeling participation in asynchronous discussions in university online introductory mathematical courses. *Educational Technology Research and Development*, *70*(6), 1993–2015. http://dx.doi.org/10.1007/s11423-022-10153-5.

Lin, Q., Liu, Y., & Yi, L. (2018). An integrated framework with feature selection for dropout prediction in Massive Open Online Courses. *IEEE Access*, *6*, 71414–71484. http://dx.doi.org/10.1109/ACCESS.2018.2881275.

Mai, T. T., Bezbradica, M., & Crane, M. (2022). Learning behaviours data in programming education: Community analysis and outcome prediction with cleaned data. *Future Generation Computer Systems*, *127*, 42–55. http://dx.doi.org/10.1016/j.future.2021.08.026.

Mallik, P., Roy, C., Maheshwari, E., Pandey, M., & Rautray, S. (2019). Analyzing student performance using data mining. In Y.-C. Hu, S. Tiwari, K. Mishra, & M. Trivedi (Eds.), *Ambient communications and computer systems, Vol. 904* (pp. 307–318). Singapore: Springer, http://dx.doi.org/10.1007/978-981-13-5934-7_28.

Manoharan, J. J., Ganesh, S. H., Felciah, M. L. P., & Banu, A. K. S. (2014). Discovering students' academic performance based on GPA using K-Means clustering algorithm. In *World congress on computing and communication technologies* (pp. 200–202). USA: IEEE, http://dx.doi.org/10.1109/WCCCT.2014.75.

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, *8*, 55462–55470. http://dx.doi.org/10.1109/ACCESS.2020.2981905.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, *29*(3), 241–288. http://dx.doi.org/10.1016/0004-3702(86)90072-X.

Przepiorka, A., Blachnio, A., Cudo, A., & Kot, P. (2021). Social anxiety and social skills via problematic smartphone use for predicting somatic symptoms and academic performance at primary school. *Computers & Education*, *173*, Article 104286. http://dx.doi.org/10.1016/j.compedu.2021.104286.

Ramze Rezaee, M., Lelieveldt, B. P. F., & Reiber, J. H. C. (1988). A new cluster validity index for the fuzzy *c*-mean. *Pattern Recognition Letters*, *19*(3–4), 239–246. http://dx.doi.org/10.1016/S0167-8655(97)00168-2.

Riestra-Gonzalez, M., Paule-Ruiz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, *163*, Article 104108. http://dx.doi.org/10.1016/j.compedu.2020.104108.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. http://dx.doi.org/10.1016/j.eswa.2006.04.005.

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12–27. http://dx.doi.org/10.1002/widm.1075.

Schapire, R. E. (1989). The strength of weak learnability. In R. Rivest, D. Haussler, & M. K. Warmuth (Eds.), *Proceedings of the 2nd annual workshop on computational learning theory* (pp. 197–227). San Francisco (CA): Morgan Kaufmann, http://dx.doi.org/10.1016/B978-0-08-094829-4.50030-1.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, *143*, Article 103676. http://dx.doi.org/10.1016/j.compedu.2019.103676.

Turabieh, H., Al Azwari, S., Rokaya, M., Alosaimi, W., Alharbi, A., Alhakami, W., et al. (2021). Enhanced Harris Hawks optimization as a feature selection for the prediction of student performance. *Computing*, *103*(7), 1417–1438. http://dx.doi.org/10.1007/s00607-020-00894-7.

Varela, N., Montero, E. S., Vásquez, C., Guiliany, J. G., Mercado, C. V., et al. (2019). Student performance assessment using clustering techniques. In Y. Tan, Y. Shi (Eds.), *Data mining and big data* (pp. 179–188). Singapore: Springer, http://dx.doi.org/10.1007/978-981-32-9563-6_19.

Yadav, R. S. (2020). Application of hybrid clustering methods for student performance evaluation. *International Journal of Information Technology*, *12*(3), 749–756. http://dx.doi.org/10.1007/s41870-018-0192-2.

Zaffar, M., Hashmani, M. A., Habib, R., Quraishi, K. S., Irfan, M., et al. (2022). A hybrid feature selection framework for predicting students performance. *Computers, Materials & Continua*, *70*(1), 1893–1920. http://dx.doi.org/10.32604/cmc.2022.018295.

Zhang, R. C., Lai, H. M., Cheng, P. W., & Chen, C. P. (2017). Longitudinal effect of a computer-based graduated prompting assessment on students' academic performance. *Computers & Education*, *110*, 181–194. http://dx.doi.org/10.1016/j.compedu.2017.03.016.