IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 25, No. 5, October, 2020, pp. 609-617
Journal homepage: http://iieta.org/journals/isi

Analysis on Course Scores of Learners of Online Teaching Platforms Based on Data Mining

Nan Zhou*, Zhaofeng Zhang, Jing Li

Shijiazhuang University of Applied Technology, Shijiazhuang 050081, China

Corresponding Author Email: rocket_2010@sina.com

https://doi.org/10.18280/isi.250508

Received: 1 June 2020
Accepted: 24 August 2020

Keywords:
course score analysis, online teaching platform (OLP), expectation maximization (EM) clustering, support vector machine (SVM) classifier

ABSTRACT

After years of development, online teaching platforms (OLPs) have accumulated a huge amount of data on student scores. To effectively mine out the useful knowledge and information behind the massive data, this paper puts forward a course score analysis model for OLP learners based on data mining. Firstly, the score features of OLP learners were classified, and the calculation method of computational features was presented. Then, the score features were clustered through expectation maximization (EM) clustering, which has the advantage of unsupervised learning. Moreover, the salient features were obtained through principal component analysis (PCA). Finally, the support vector machine (SVM) prediction algorithm, a supervised learning method, was constructed, and merged with the clustering algorithm to realize accurate classification of the course scores of OLP learners. The effectiveness of the proposed method was proved through experiments. Based on the correlation between learner scores and courses, this research enables teachers to improve current teaching models and methods.

1. INTRODUCTION

The rapid development of information technology (IT) and the Internet has greatly promoted the reform and innovation of traditional teaching models, resulting in better teaching effect [1-4]. More and more schools and teachers have turned to online teaching platforms (OLPs) that integrate online and offline teaching based on Internet Plus. After years of development, OLPs have accumulated a huge amount of data on student scores. To optimize the teaching strategies, contents, and plans, it is not enough to conduct simple statistical analysis and summary of these data [5-9]. Instead, the massive data on student scores need to be analyzed by big data technology, aiming to cluster the students and extract group features. This will enable the teachers to propose pertinent teaching strategies, and design effective teaching contents and plans, thereby improving the learning methods and scores of students.

Data mining is widely used in such fields as finance, science and technology, and agriculture. It is also increasingly applied in education and teaching [10-14]. The existing studies on data mining of education data mainly focus on the prediction of student scores, and the correlation between course scores [15-18].

On the prediction of student scores, Dahri et al. [19] constructed a total score prediction model based on long short-term memory (LSTM) network, and compared it with Bayesian algorithm and decision tree (DT) to verify its accuracy in predicting the scores of graduate and undergraduate students. Using effectively collected enrollment data, Syahidi and Asyikin [20] contrasted the effects of DT, logistic regression, and backpropagation neural network (BPNN) in freshman scores, and discovered that the BPNN achieved the highest prediction accuracy. Hamdi and Kartowagiran [21] preprocessed enrollment data (e.g. enrollment rate) through linear regression, and combined DT with naive Bayes to predict undergraduate scores on the Weka machine learning workbench.

On the correlation between course scores, Arami and Wiyarsi [22] identified the key factors affecting the student scores of universities for nationalities by k-means clustering (KMC), and obtained the highly correlated course scores. With the aid of association rules, Wibawa [23] analyzed the correlations between the factors affecting the total score of students, mined out the useful rules, and optimized the current allocation of teaching resources. Gkontzis et al. [24] applied the improved association rule Apriori algorithm to evaluate the correlations between course scores of computer application majors, and quantified the degrees of correlation between the scores of professional courses. Northey et al. [25] relied on the Iterative Dichotomiser 3 (ID3), a DT algorithm, to examine the factors affecting student scores, and optimized the design of student score mining system, using improved association rule mining and clustering algorithms.

Based on data mining, this paper puts forward a course score analysis model for OLP learners. The purpose is to effectively mine out the useful knowledge and information behind the massive data on OLP learner scores, and to accurately evaluate the rationality of teaching plans, the distribution of course scores, and the importance of each course.

The remainder of this paper is organized as follows: Section 2 classifies the score features of OLP learners, and details the calculation method of the computational features; Section 3 analyzes the learner scores through expectation maximization (EM) clustering, which has the advantage of unsupervised learning, and obtains the salient features by principal component analysis (PCA); Section 4 combines the clustering algorithm with the support vector machine (SVM), a supervised learning method, to accurately categorize the features of OLP learner scores; Section 5 verifies the effectiveness of our method through experiment; Section 6

puts forward the conclusions.

2. FEATURE DEFINITION AND EXTRACTION

According to the data fusion theory on the big data of education and teaching, the multi-source data of OLP learners were divided into five categories: trajectories, social behaviors, resource learning, evaluation & reflection, and basic information. The features of these types of data were defined, extracted, and fused (Figure 1).

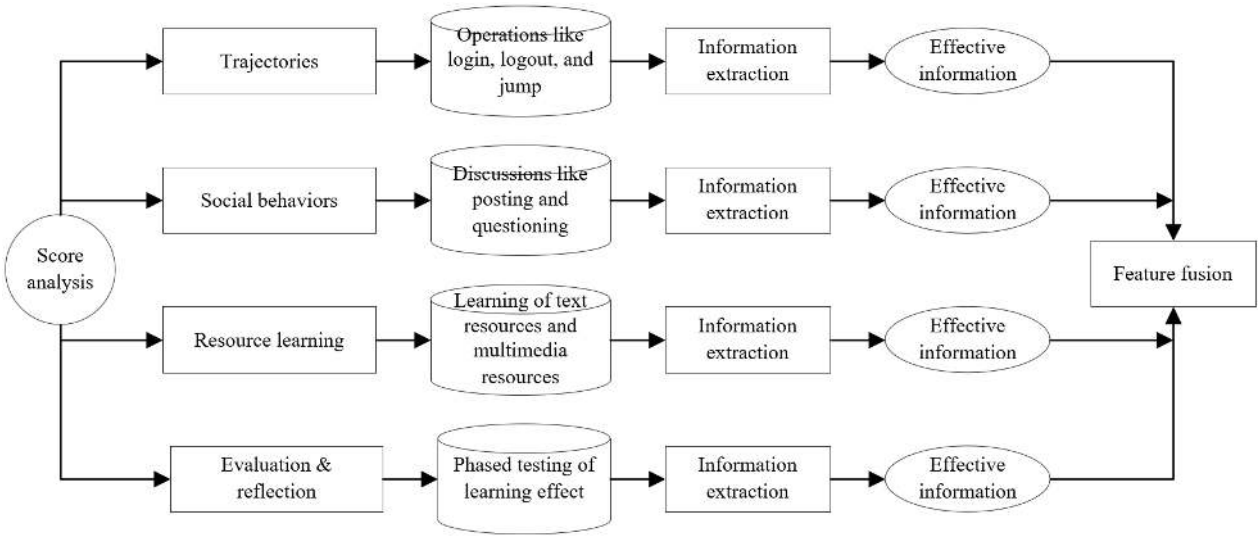


Figure 1. The feature definition, extraction, and fusion of multi-source OLP data

Table 1. The features extracted from OLP learner scores

Computational features	Statistical features
Stay time in school hours	Number of views of learning resources for major courses
Stay time on holidays	Number of views of learning resources for other resources
Number of resource learnings on holidays	Number of assignments
Number of learnings at night	Test score
Attendance rate of compulsory courses	Number of discussions

Note: School hours, a.k.a. regular study hours, refer to 7:30-17:30 on every workday.

The trajectories, the most basic and common data in the OLP, refer to the various operations by learners in the OLP, including login, logout, jump between webpages, and closing webpages.

In terms of social behaviors, every learner can communicate with teachers and other learners in classroom learning on the OLP. The main social behaviors are discussions like real-time posting, questioning, replying, and commenting.

Resource learning dominates the behaviors of OLP learners. The OLP mainly provides two kinds of learning resources: text resources (e.g. PDF and Word), and multimedia resources (e.g. videos, and micro lectures. The learning of text resources involves operations like start, exit, and stay. The learning of multimedia resources involves operations like start, pause, fast forward, loop playback, and stop.

Evaluation & reflection stands for the phased testing of the learning effect on quizzes, assignments, and evaluations. The learning situation of a learner can be reflected comprehensively, if the his/her answering process is preserved on the OLP.

The basic information covers the following personal information of learners: major, student identity (ID), and timetable. Table 1 lists the features extracted from OLP learner scores.

To extract the computational features from the above data, sets A and B were defined as the set of learning resources provided by the OLP, and the set of learning resources for the major, respectively:

$$A = \{a_1, a_2, ..., a_i, a_{i+1}, ..., a_N\} \quad (1)$$

$$B = \{b_1, b_2, ..., b_j, b_{j+1}, ..., b_M\} \quad (2)$$

where, a_i and b_i are the i-th learning resource provided by the OLP and the learning resources for the major of the j-th learner, respectively. In a D-day-long q-th period, the trajectories of the j-th learner can be described by a two-tuple (P^{qj}, t^{qj}), where P^{qj} is the stop position of the learning trajectories after the fusion of multi-source learner data, and t is the time that the learner appears in that position. Then, the stop positions of all the trajectories of the j-th learner in the q-th period can be expressed as a set P:

$$P = \{p_1, p_2, ..., p_i, p_{i+1}, ..., p_K\} \quad (3)$$

where, p_i is the i-th stop position with a stay time longer than 10min in the trajectories. Next, the OLP learner score features were defined:

(1) Stay time in school hours T_{1q}

In the q-th period of school hours, the positions of learner trajectories that stop at the learning resources of his/her major can be expressed as the following set:

$$P_{maj1}^{qj} = \{ p_i \,|\, for\ each\ p_i \in P \cap B\ and\ t^{qj} \in [7:30, 17:30] \} \quad (4)$$

The stay time at each stop position can be expressed as:

$$T_{maj1}^{qj} = \{ t_1, t_2, ..., t_i, t_{i+1}, ..., t_L \} \quad (5)$$

where, t'_i is the stay time at the i-th stop position.

The capacity of P^{qj}_{maj1} can be described as:

$$C_1 = \left| P_{maj1}^{qj} \right| \quad (6)$$

Thus, T_{1q} can be defined as:

$$T_{1q} = \sum_{i=1}^{L} t_i \quad (7)$$

(2) Stay time on holidays T_{2q}

Similar to T_{1q}, in the q-th period of holidays, the positions of learner trajectories that stop at the learning resources of his/her major can be expressed as the following set:

$$P_{maj2}^{qj} = \{ p_i \,|\, for\ each\ p_i \in P \cap A\ and\ t^{qj} \in holiday \} \quad (8)$$

The stay time at each stop position can be expressed as:

$$T_{maj2}^{qj} = \{ t'_1, t'_2, ..., t'_i, t'_{i+1}, ..., t'_L \} \quad (9)$$

where, t'_i is the stay time at the i-th stop position.

The capacity of P^{qj}_{maj2} can be described as:

$$C_2 = \left| P_{maj2}^{qj} \right| \quad (10)$$

Thus, T_{2q} can be defined as:

$$T_{2r} = \sum_{i=1}^{l} t'_1 \quad (11)$$

(3) Number of resource learnings on holidays N_{1q}

The number of resource learnings on holidays is the capacity of P^{qj}_{maj2}:

$$N_{1q} = C_2 \quad (12)$$

(4) Number of learnings at night N_{2q}

Based on the definition of school hours, the period of night learning was limited to 19:30-22:30 at night. The state of night learning NS_i was defined as follows: If a learner stops at the learning resources of his/her major in the q-th period and the specified period, then he/she learns at night (NS_i=1); otherwise, he/she does not learn at night (NS_i=0):

$$NS_i = \begin{cases} 1, if\ t^{qj} \in [19:30, 22:30]\ and\ P^{qj} \in A \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases} \quad (13)$$

The number N_{2q} of learnings at night in a semester can be calculated by:

$$N_{2r} = \sum_{i=1}^{D} NS_i \quad (14)$$

(5) Attendance rate of compulsory courses ATT_L

The attendance in public compulsory courses can basically reflect the overall attendance in all courses. Therefore, the attendance rates of OLP learners in compulsory courses were calculated for different periods in one semester. For the j-th learner, the start and end times of compulsory courses in the q-th period can be calculated by:

$$T_{cou}^{qj} = \{ (t_1^b, t_1^e), (t_2^b, t_2^e), ..., (t_i^b, t_i^e), (t_{i+1}^b, t_{i+1}^e), ..., (t_H^b, t_H^e) \} \quad (15)$$

$$C_3 = \left| t_{cou}^{qj} \right| \quad (16)$$

where, t^b_i and t^e_i are the start and end times of the i-th course, respectively. Then, the positions of online courses at each time point in set T^{qj}_{cou} can be described by the following set:

$$W = \{ w_1, w_2, ..., w_i, w_{i+1}, ..., w_H \} \quad (17)$$

For the accuracy of attendance rate, the course duration was expanded 8min before and after the start and end times, creating the range of online time for learners. Then, the attendance rate ATT_l of the l-th compulsory course can be defined as:

$$ATT_l = \begin{cases} 1, if\ t^{qj} \in [t_1^b - 8, t_i^e + 8]\ and\ P^{qj} = w_i \\ 0, \qquad\qquad otherwise \end{cases} \quad (18)$$

$$ATT_L = \left(\sum_{l=1}^{H} ATT_l \right) / C_3 \quad (19)$$

After the computing and extracting the computational and statistical features, the score features of OLP learners can be modelled as Figure 2.

3. FEATURE CLUSTERING AND PCA

The scores on OLPs are private data of the learners. The acquisition of these data brings certain risks and difficulties. Considering the universality of relevant analysis methods and the predefined features of OLP learner scores, this paper decides to analyze learner scores through EM clustering, which has the advantages of unsupervised learning. The workflow of EM clustering is explained in Figure 3.

The EM clustering is an iterative algorithm based on maximum posterior probability and maximum likelihood estimation. Suppose dataset Z=(X, Y) encompasses observed data X and unobserved data Y, and $JPD(X,Y|\Psi)$ be the joint probability density of these data. Then, finding the maximum of the likelihood function LF (X; Ψ) of X is to make the maximum likelihood estimation of Ψ:

$$LF(X;\Psi) = \log JPD(X|\Psi) = \int \log JPD(X,Y|\Psi) dY \quad (20)$$

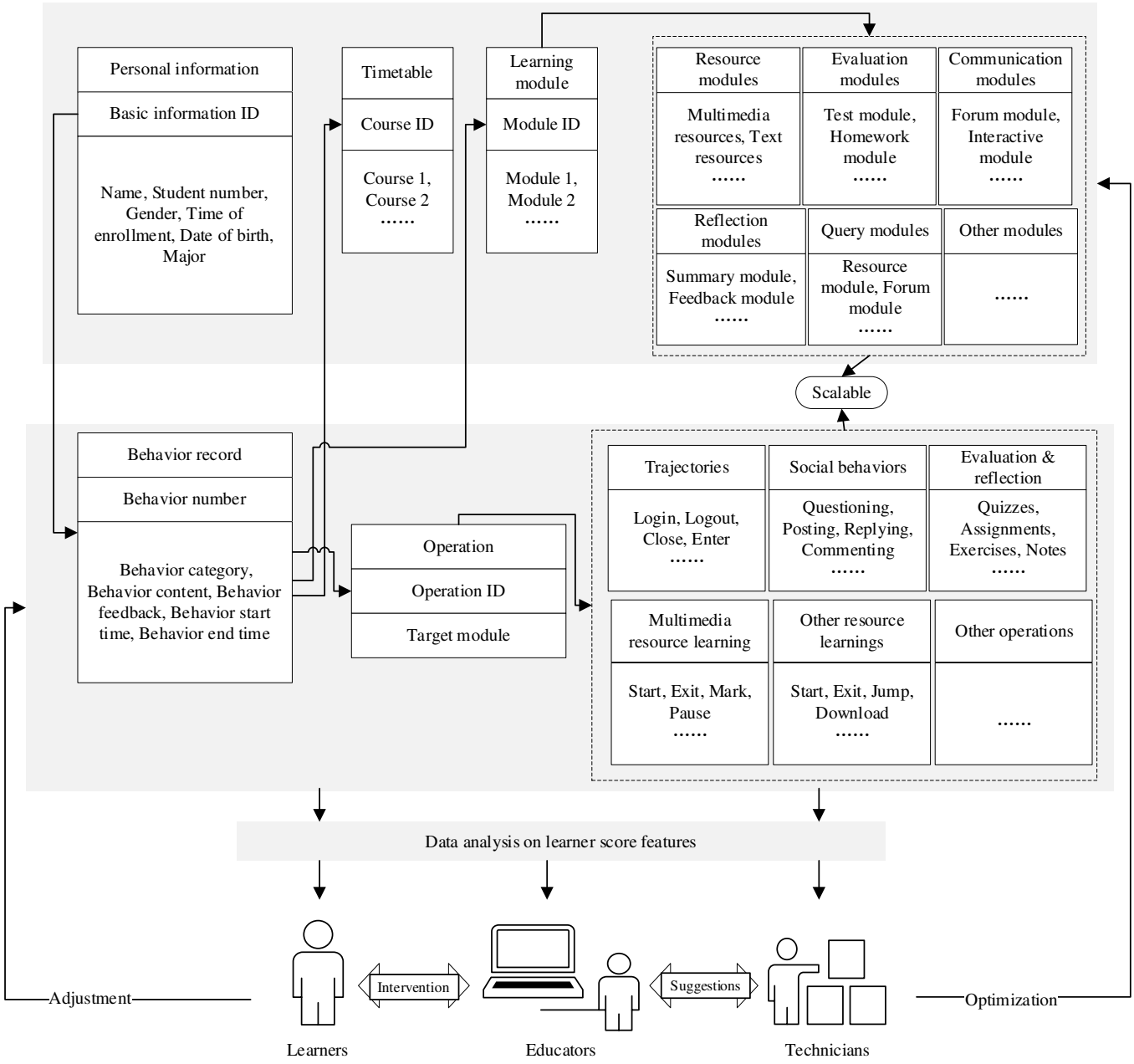


Figure 2. The model of score features for OLP learners

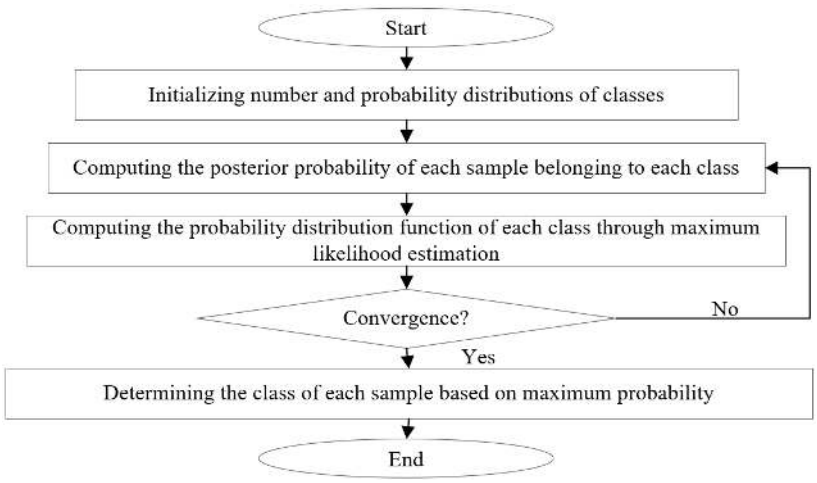


Figure 3. The workflow of EM clustering

In EM clustering, the EM of $LF^*(X; \Psi)$ is iteratively calculated to maximize the log likelihood function of X:

$$LF^*(X;\Psi) = \log JPD(X,Y|\Psi) \tag{21}$$

After t iterations, the likelihood estimate of Ψ can be expressed by Ψ(*t*). During the t+1-th iteration, the likelihood function expectation of Z during the calculation of the posterior probability of each sample belonging to each class can be described as:

$$LFE(\Psi|\Psi(t)) = PPC\{LF^*(\Psi;Z)|X|\Psi(t)\} \tag{22}$$

where, *PPC* {} is the posterior probability calculation function. During the updating of the probability distribution function of each class through maximum likelihood estimation, *LFE* (Ψ|Ψ|(*t*)) is maximized, and Ψ is updated.

The above steps are repeated iteratively to optimize and update model parameters. In this way, the likelihood probability of the training samples and model parameters continues to increase until reaching the extreme point.

After clustering, it is necessary to select suitable features of OLP learner scores. Otherwise, there will be no basis for subsequent data mining. Inspired by the idea of dimensionality reduction, this paper transforms multiple features of OLP learner scores into a few representative features through PCA.

Let $Z=(z_1,z_2,\ldots,z_g)^T$ be a g-dimensional random vector obtained by standardizing the original observed data. Then, a matrix of *U* samples $z_i=(z_{i1},z_{i2},\ldots,z_{ig})^T$ was constructed, and the matrix elements were standardized by:

$$S_{ij} = \frac{z_{ij} - \overline{z}_j}{\sqrt{\sum_{j=1}^{U}(z_{ij} - \overline{z}_j)^2 / (U-1)}} \tag{23}$$

$$\overline{z}_j = \frac{\sum_{i=1}^{U} z_{ij}}{U} \tag{24}$$

The correlation coefficient matrix can be established as:

$$R = [r_{ij}]_g = \frac{S^T S}{U-1} = \left[\frac{\sum s_{kj} \cdot s_{jk}}{U-1}\right]_g \tag{25}$$

Solving the characteristic equation $|R-\lambda I_g|$ of formula (25), a total of g characteristic roots can be obtained. The PCA needs to satisfy the following inequality:

$$\sum_{j=1}^{d}\lambda_j \Big/ \sum_{j=1}^{g}\lambda_j \geq 0.8 \tag{26}$$

Then, the d value of principal components with a utilization rate greater than 0.8 was calculated. Solving the equation set $Rb=\lambda_j b$, the unit eigenvector σ^0_j was obtained. Next, the standardized data features were converted into principal components:

$$PC_{ij} = S_i^T \sigma_j^0 \tag{27}$$

Taking the variance contribution rate as the weight, the d principal component features were weighted and summed to derive the comprehensive evaluation of OLP learner scores.

4. SVM-BASED CLASSIFICATION MODEL

During the analysis and differentiation of OLP learner scores, unsupervised learning can avoid the disclosure of private information like score ranking. However, unsupervised learning cannot accurately distinguish between learners in different score intervals, making the classified management of learners less scientific and effective. Therefore, this paper combines unsupervised EM clustering with supervised SVM (Figure 4) to re-analyze and predict the OLP learner scores.

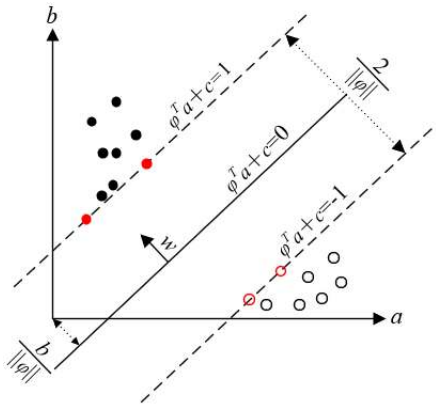


Figure 4. The principle of the SVM algorithm

Let $f(a)=\varphi^T a+c$ be the discriminant function of the SVM, $f(a)=0$ be a two-dimensional (2D) hyperplane that divides all feature samples into two classes, and (a_1, b_1), (a_2, b_2), …(a_V, b_V) be the V feature samples (bi=+1 or -1). Then, the classification formula can be defined as:

$$\begin{cases} \phi^T a + c \geq 0, b_i = +1 \\ \phi^T a + c < 0, b_i = -1 \end{cases} \tag{28}$$

Changing the modulus of the weight vector, the classification rules can be updated as:

$$\begin{cases} \phi^T a + c \geq \delta > 0 \rightarrow \phi^T a + c \geq 1, b_i = +1 \\ \phi^T a + c \leq -\delta < 0 \rightarrow \phi^T a + c \leq -1, b_i = -1 \end{cases} \tag{29}$$

Merging the above formulas:

$$b_i(\phi^T a + c) \geq 1 \tag{30}$$

The distance from a to the hyperplane F must be greater than 1:

$$|g(a_i)| \geq 1 \tag{31}$$

The geometric interval from a sample point to the hyperplane F can be described as:

$$\frac{b_i(\phi^T a + c)}{\|\phi\|} \tag{32}$$

where, the numerator $b_i(\varphi^T a+c)$is the interval from the sample point to the hyperplane function. Let γ be the vector perpendicular to the hyperplane F, and υ be the vertical distance between a and F. Following the principle of vector addition, vector a can be rewritten as:

$$a=\frac{a_p+\upsilon\phi}{\|\phi\|} \tag{33}$$

$$f(a)=\phi^T(a_p+\frac{\upsilon\phi}{\|\phi\|})+c=\upsilon\|w\|\phi \tag{34}$$

To maximize the interval between the two classes, $\|\varphi\|$ should be minimized, that is, $\|\varphi\|^2$ should be minimized. However, there exists a constraint $|f(a)\geq 1|$, indicating that the sample point closest to the hyperplane represented by the support vector needs to satisfy $|f(a)=1|$.

According to the concept of data interval, the distance between two sample classes can be defined as $2/\|\varphi\|$. To minimize $\|\varphi\|/2$, i.e. maximize the inter-class distance, it is necessary to configure the optimal hyperplane constraint, that is, the optimization function of SVM:

$$\min_{\phi}\frac{1}{\|\phi\|_2}\quad s.t.\ b_i(\phi^T a+c)\geq 1 \tag{35}$$

After finding the support vector and maximizing the interval, φ and c can be determined. Then, the optimization function is equivalent to:

$$\arg\max_{\phi,\ c}\left\{\min_{\phi}(\phi^T a_i+c))\frac{1}{\|\phi\|}\right\} \tag{36}$$

Since linear classification cannot differentiate all actual samples, the few misclassified samples were separated with a slack variable:

$$\min\frac{1}{2}\|\phi\|^2+\eta\sum_{i=1}^{n}\mu_i\quad s.t.\ b_i(\phi^T a+c)\geq 1-\mu_i,\mu_i\geq 0 \tag{37}$$

where, μ_i is the slack variable; η is a constant, representing the penalty function. If misclassified, sample a can be discarded if the η value is sufficiently small. The smaller the η value, the wider the hyperplane, and the more the misclassified sample points. This problem can be solved by introducing the Lagrangian factor to hyperplane optimization. The Lagrangian optimization function can be constructed as:

$$\begin{aligned}&\max L(\phi,c,\alpha)=\frac{1}{2}(\phi^T\phi)-\sum_{i=1}^{V}\alpha_i\left[b_i(\phi^T a_i+c)-1\right]\\&s.t.\quad 0\leq\alpha_i\leq\eta\end{aligned} \tag{38}$$

Finding the partial derivatives:

$$\begin{cases}\dfrac{\partial L(\phi,c,\alpha)}{\partial\phi}=0\ \Leftrightarrow\ \phi-\sum_{i=1}^{V}\alpha_i b_i a_i=0\\ \dfrac{\partial L(\phi,c,\alpha)}{\partial c}=0\ \Leftrightarrow\ \sum_{i=1}^{V}\alpha_i b_i=0\end{cases} \tag{39}$$

Combining formulas (38) and (39):

$$L(\phi,c,\alpha)=\sum_{i=1}^{V}\alpha_i-\frac{1}{2}\sum_{i=1}^{V}\sum_{j=1}^{V}\alpha_i\alpha_j b_i b_j a_i^T a_j \tag{40}$$

The optimal solution α_i^* can be obtained by taking the maximum of the above formula. The optimal constraint can be described as:

$$\phi^*=\sum_{i}^{V_S}\alpha_i^* b_i a_i \tag{41}$$

where, V_S is the number of support vectors. The optimal bias can be obtained by:

$$c^*=b_i-\phi^{*T}a_i \tag{42}$$

Solving the Lagrangian factor α_i, the optimal hyperplane can be easily obtained.

5. EXPERIMENTS AND RESULT ANALYSIS

The massive data on OLP learner behaviors were collected and processed through the flow in Figure 5, and the preprocessed features were stored in a non-relational database. The learner behavior data involve attributes like StudentID, SessionID, Verb, Object, and Context, which help to judge whether a learner enters a session or exit the system.

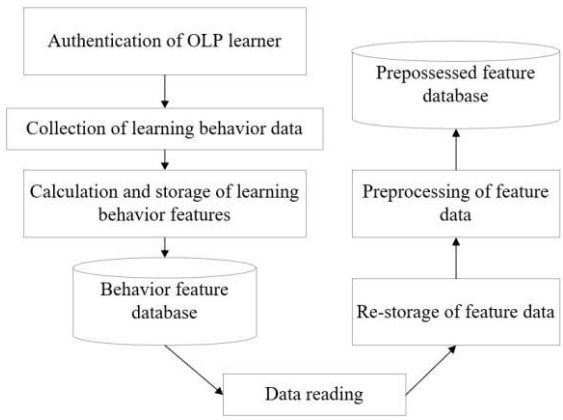


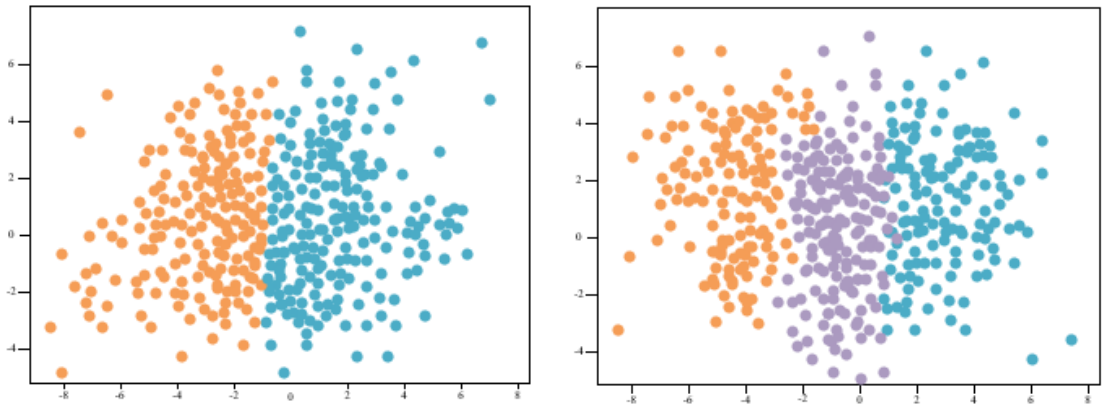
Figure 5. The collection and preprocessing of learner behavior data

The originally extracted data contain lots of redundant information. The high-quality features need to be selected from the original data, such that the model will not face problems induced by the curse of dimensionality: poor generalization, high complexity, and long training. Here, the correlation coefficients of score features in each category are calculated by the correlation coefficient method. The calculated results are recorded in Table 2.

If the Pearson correlation coefficient falls in 0.5-1.0, the features are strongly correlated; if the coefficient falls in 0.3-0.5, the features are moderately correlated; if the coefficient falls in 0.1-0.3, the features are slightly correlated; if the coefficient falls in 0-0.1, the features are basically uncorrelated. This paper chooses the features with Pearson correlation coefficient greater than 0.3 for clustering analysis.

Table 2. The salient features of OLP learner scores and their Pearson correlation coefficients

Computational features	Correlation coefficient	Statistical features	Correlation coefficient
Stay time in school hours	0.431	Total number of visits to multimedia resources	0.321
Stay time on holidays	0.411	Total number of visits to text resources	0.311
Number of resource learnings on holidays	0.382	Score of unit quiz	0.316
Number of learnings at night	0.341	Number and quality of assignments	0.357
Attendance rate of compulsory courses	0.391	Forum activity	0.382
		Number of interactions with teachers	0.326
		Number of notes and feedbacks	0.346
		Self-evaluation	0.391
		Student-student mutual evaluation	0.325



(a) Two classes (b) Three classes

Figure 6. The visualized results of EM clustering of selected features

From the multi-source data of OLP learners throughout the 20-week semester, the features closely correlated with learner scores were extracted, and grouped by EM clustering into two classes (I and II) or three classes (I, II, and III). To give a visual display of the clustering results, the features to be clustered were subject to PCA and dimensionality reduction. The visualized results of EM clustering are displayed in Figure 6.

Following the idea of equal division, the results of three-class EM clustering before and after feature selection were divided into three levels, namely, 1-200, 200-400, and 400-600, and the clustering accuracy in each level was calculated in turn. Table 3 compares the accuracies of EM clustering before and after feature selection.

From Figure 6 and Table 3, the EM clustering achieved a good effect on the OLP learner scores. The accuracy before feature selection was higher than that after feature selection. The three-class clustering was more accurate than two-class clustering, with accuracy falling between 82.76% and 86.74%. Therefore, our clustering analysis method excels in distinguishing fine-grained features of learner scores.

Next, three SVM classifiers were designed by integrating multiple binary classifiers through one-against-one. Their performance was cross validated, and evaluated. Then, the SVM classifiers were adopted to predict the multiple classes of learner scores, based on the feature subset of the multi-source data on OLP learner scores and the features closely correlated with learner scores. Table 4 compares the classification results of the three SVM classifiers.

It can be seen that the three-class SVM classifier achieved an accuracy of 82-91%, slightly higher than that of two-class SVM classifier. This classifier effectively distinguished between the scores of learners in different levels, suggesting that it is suitable for analyzing and managing OLP learner scores.

Furthermore, the scores of electrical automation majors on 13 professional courses were clustered by the proposed SVM classifier and the rule-based classifier. As shown in Figure 7, the proposed SVM classifier outshined the rule-based classifier in accuracy and recall, despite a slight lag in efficiency. The results demonstrate the accuracy and effectiveness of our method in the classification of OLP learners and their score features.

Table 3. The EM clustering accuracies before and after feature selection

Levels	I		II		III	
	Before feature selection	After feature selection	Before feature selection	After feature selection	Before feature selection	After feature selection
1-200	59	61	20	21	74 (positive samples)	76 (positive samples)
200-400	84 (positive samples)	96 (positive samples)	62	59	52	58
400-600	41	39	80 (positive samples)	92 (positive samples)	24	26
Accuracy	82.76%	82.91%	84.82%	86.74%	82.74%	85.74%

Table 4. The classification results of the 3 SVM classifiers before and after feature extraction

Type	Accuracy		Recall		F1-score		Support	
	Before feature selection	After feature selection	Before feature selection	After feature selection	Before feature selection	After feature selection	Before feature selection	After feature selection
I	0.87	0.89	0.19	0.21	0.24	0.29	16	21
II	0.82	0.85	0.92	0.94	0.78	0.82	58	59
III	0.87	0.91	0.16	0.22	0.34	0.46	22	22

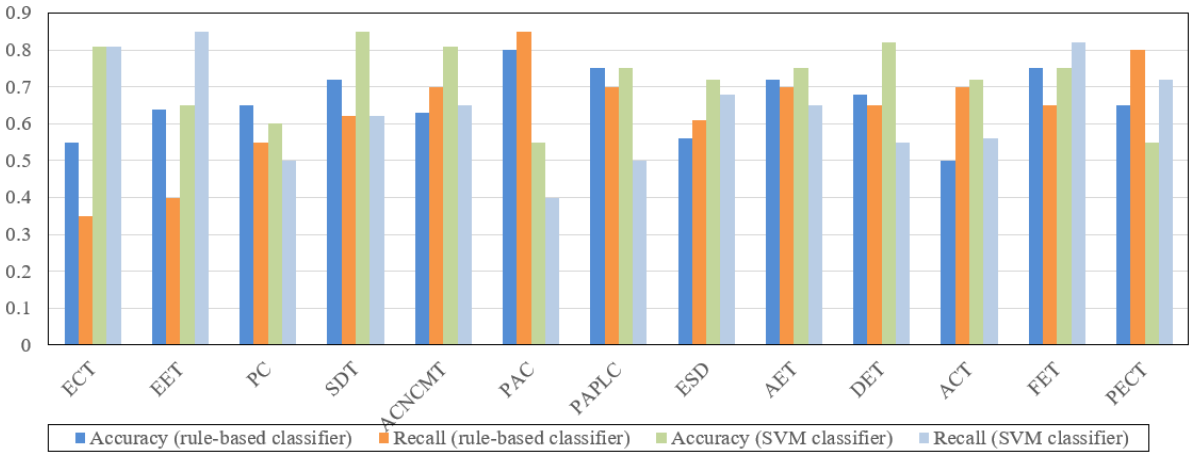


Figure 7. The classification results on professional course scores of rule-based classifier and SVM classifier

Note: ECT is Electrical Control Technology, EET is Electrical and Electronic Technology, PC is Programmable Controller, SDT is Sensing and Detection Technology, ACNCMT is Application of Computer Numerically Controlled (CNC) Machine Tools, PAC is Principle of Automatic Control, PAPLC is Principle and Application of Programmable Logic Controller (PLC), ESD is Electronic System Design, AET is Analog Electronic Technology, DET is Digital Electronic Technology, ATC is Application of Computer Technology, FET is Fundamentals for Electrical Towage, PECT is Power Electronic Conversion Technology.

6. CONCLUSIONS

Based on data mining, this paper develops a course score analysis model for OLS learners. Firstly, the score features of OLS learners were categorized, and the calculation method for computational features was detailed. Then, EM clustering was adopted to cluster the score features of the learners, owing to its advantage of unsupervised learning. The salient features were obtained through PCA. Experimental results demonstrate that our clustering analysis method excels in distinguishing fine-grained features of learner scores. Finally, the authors designed an SVM classifier, a supervised learning tool, and combined it with EM clustering to accurately categorize the score features of OLS learners. The proposed method was compared with the rule-based classifier. The comparison shows that our method achieved the better accuracy and recall, an evidence to its feasibility and accuracy.

REFERENCES

[1] Mahmud, M., Nor, N.M., Jauhari, N.E., Nordin, N.I., Rahman, N.A. (2020). Student engagement and attitude in mathematics achievement using single valued neutrosophic set. Journal of Physics: Conference Series, 1496(1): 012017. https://doi.org/10.1088/1742-6596/1496/1/012017

[2] Faber, J.M., Luyten, H., Visscher, A.J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. Computers & Education, 106: 83-96. https://doi.org/10.1016/j.compedu.2016.12.001

[3] Yildirim, I. (2017). The effects of gamification-based teaching practices on student achievement and students' attitudes toward lessons. The Internet and Higher Education, 33: 86-92. https://doi.org/10.1016/j.iheduc.2017.02.002

[4] Masson, A.L., Klop, T., Osseweijer, P. (2016). An analysis of the impact of student–scientist interaction in a technology design activity, using the expectancy-value model of achievement related choice. International Journal of Technology and Design Education, 26(1): 81-104. https://doi.org/10.1007/s10798-014-9296-6

[5] Giridharan, P.K., Raju, R. (2016). The impact of experiential learning methodology on student achievement in mechanical automotive engineering education. International Journal of Engineering Education, 32(6): 2531-2542.

[6] Asarta, C.J., Schmidt, J.R. (2017). Comparing student performance in blended and traditional courses: Does prior academic achievement matter? The Internet and Higher Education, 32: 29-38. https://doi.org/10.1016/j.iheduc.2016.08.002

[7] Popyack, J.L. (2016). UPSILON PI EPSILON UPE 2016 national meeting, and celebrating outstanding student achievement. ACM Inroads, 7(2): 28-30. https://doi.org/10.1145/2926716

[8] Mabed, M., Köhler, T. (2016). Aligning performance assessments with standards: A practical framework for improving student achievement in vocational education. Knowledge, Information and Creativity Support Systems, 416: 575-585. https://doi.org/10.1007/978-3-319-27478-2_45

[9] Rahman, K., Qodriyah, K., Bali, M.M.E.I., Baharun, H., Muali, C. (2020). Effectiveness of android-based

mathematics learning media application on student learning achievement. Journal of Physics: Conference Series, 1594(1): 012047. https://doi.org/10.1088/1742-6596/1594/1/012047

[10] Apriesnig, J.L., Manning, D.T., Suter, J.F., Magzamen, S., Cross, J.E. (2020). Academic stars and energy stars, an assessment of student academic achievement and school building energy efficiency. Energy Policy, 147: 111859. https://doi.org/10.1016/j.enpol.2020.111859

[11] Wibowo, A.N., Wibowo, Y.E. (2020). Implementing student teams-achievement division to improve student's activeness and achievements on technical drawing courses. Journal of Physics: Conference Serie, 1446(1): 012033. https://doi.org/10.1088/1742-6596/1446/1/012033

[12] Apuanor, S., Yuniarsih, R.O. (2020). The influence of library and internet utilization of student achievement index. Journal of Physics: Conference Series, 1477(4): 042026. https://doi.org/10.1088/1742-6596/1477/4/042026

[13] Rustana, C.E., Andriana, W., Serevina, V., Junia, D. (2020). Analysis of student's learning achievement using PhET interactive simulation and laboratory kit of gas kinetic theory. Journal of Physics: Conference Series, 1567(2): 022011. https://doi.org/10.1088/1742-6596/1567/2/022011

[14] Wasil, M., Sudianto, A. (2020). Application of the decision tree method to predict student achievement viewed from final semester values. Journal of Physics: Conference Series, 1539(1): 012027. https://doi.org/10.1088/1742-6596/1539/1/012027

[15] Sebillo, M., Vitiello, G., Di Gregorio, M. (2020). Maps4Learning: Enacting geo-education to enhance student achievement. IEEE Access, 8: 87633-87646. https://doi.org/10.1109/ACCESS.2020.2993507

[16] Sudipa, I.G.I., Wijaya, I.N.S.W., Radhitya, M.L., Mahawan, I.M.A., Arsana, I.N.A. (2020). An android-based application to predict student with extraordinary academic achievement. Journal of Physics: Conference Series, 1469(1): 012043. https://doi.org/10.1088/1742-6596/1469/1/012043

[17] Lerche, T., Kiel, E. (2018). Predicting student achievement in learning management systems by log data analysis. Computers in Human Behavior, 89: 367-372. https://doi.org/10.1016/j.chb.2018.06.015

[18] Setyowati, C.S.P., Louise, I.S.Y. (2018). Implementation of reflective pedagogical paradigm approach on the rate of reaction to student achievement. Journal of Physics: Conference Series, 1097(1): 012057. https://doi.org/10.1088/1742-6596/1097/1/012057

[19] Dahri, S., Yusof, Y., Chinedu, C. (2018). TVET lecturer empathy and student achievement. Journal of Physics: Conference Series, 1049(1): 012056. https://doi.org/10.1088/1742-6596/1049/1/012056

[20] Syahidi, A.A., Asyikin, A.N. (2018). Applying student team achievement divisions (STAD) model on material of basic programme branch control structure to increase activity and student result. IOP Conference Series: Materials Science and Engineering, 336(1): 012027. https://doi.org/10.1088/1757-899X/336/1/012027

[21] Hamdi, S., Kartowagiran, B. (2020). Learning achievement of Elementary School student of mathematics using the Testlet model instrument: A comparison between the 2006 Curriculum and the 2013 Curriculum. Journal of Physics: Conference Series, 1581(1): 012055. https://doi.org/10.1088/1742-6596/1581/1/012055

[22] Arami, M., Wiyarsi, A. (2020). The student metacognitive skills and achievement in chemistry learning: correlation study. Journal of Physics: Conference Series, 1567(4): 042005. https://doi.org/10.1088/1742-6596/1567/4/042005

[23] Wibawa, S.C. (2018). The impact of fashion competence and achievement motivation toward college student's working readiness on "Cipta Karya" subject. In Materials Science and Engineering Conference Series, 296(1): 012017. https://doi.org/10.1088/1757-899X/296/1/012017

[24] Gkontzis, A.F., Kotsiantis, S., Tsoni, R., Verykios, V.S. (2018). An effective LA approach to predict student achievement. Proceedings of the 22nd Pan-Hellenic Conference on Informatics, New York, US, pp. 76-81. https://doi.org/10.1145/3291533.3291551

[25] Northey, G., Govind, R., Bucic, T., Chylinski, M., Dolan, R., Van Esch, P. (2018). The effect of "here and now" learning on student engagement and academic achievement. British Journal of Educational Technology, 49(2): 321-333. https://doi.org/10.1111/bjet.12589