Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 285–298, DOI 10.12694/scpe.v25i1.2295

RESEARCH ON DETECTION TECHNOLOGY OF ABNORMAL DATA IN COLLEGE PHYSICAL EDUCATION NETWORK TEACHING TEST RESULTS

FENG SHAN*AND DONGQI LI†

Abstract. In recent years, with the rapid development of big data technology, more and more data are continuously generated with the summary of university systems. Therefore, how to use these educational data to provide more scientific decision-making information for university information builders is very important. This research collected various educational data through the network teaching system and campus information system. After that, the collected data was used to analyze the students' online behavior and to excavate valuable behavioral characteristics. In addition, the experiment also proposed indicators to describe students' behavior, such as network course behavior, network viscosity and life regularity, to provide the basis for the subsequent abnormal performance prediction model. Finally, the experiment used DBSCAN algorithm based on distance optimization for clustering analysis, and constructed a NA model based on multiple classifiers. The research results showed that when, the SC value was 0.711, which was the optimal solution of D-DBSCAN algorithm. At this time, the corresponding number of clusters was 4. When N=2, that was, the base classifier of NA model is composed of C4.5 model and SVM model, the prediction accuracy and time consumption are the most appropriate. The accuracy, recall and F1 values of NA model were 98.16%, 97.26% and 0.958 respectively, which was better than that of single model. To sum up, the NA model based on classifiers proposed in the study had higher accuracy and better model performance, can effectively reflect students' academic level, and could provide accurate abnormal performance data for college sports online teaching tests.

Key words: Abnormal data; Achievements; Network education; Distance optimization; Cluster analysis

1. Introduction. With the rapid development of modern information technology, the construction of network education has gradually improved. Therefore, a large number of education data are gathered [1]. How to excavate valuable information from these educational data has become a problem faced by university informatization builders. Education Data Mining (EDM) can provide more scientific and accurate basis for university administrators' management decisions [2]. Social psychology theory shows that human behavior is determined by subjective norms, attitudes and perceived behavior control [3]. Therefore, through the analysis of students' behavior, we can reflect on individual behavior attitudes and tendencies. In the construction of university informatization, there will be a large number of student behavior data such as achievement data, library related data and network log data. These data reflect students' learning attitude and learning status, and can provide data basis for the analysis of students' behavior logs [4].

At present, most colleges and universities have realized the construction of online teaching platform, so that students can obtain tutorial resources. For students with strong self-discipline ability, this can help them improve their academic level [5, 6]. However, for students with poor self-control, disordered lifestyles and improper internet use have caused significant negative impacts that cannot be ignored.

Students spend a lot of time on socializing, entertainment, and games, exacerbating the risk of entrepreneurial failure, and in severe cases, even psychological problems may occur [7]. In addition, most of the current research on the analysis of academic performance in various universities is based on shallow level analysis of simple data and models, with a single field oriented approach, such as campus card consumption data and online education platform data.

There has been no research on the impact of combining online behavior with other data on academic performance. Therefore, it is very important to accurately predict students' academic performance and pay attention to students' abnormal performance data. Through the above means, students' online behavior can

* School of Physical Education, Shanghai University of Sport, Shanghai 200438, China
†College of Physical Education and Health, Hunan University of Technology and Business, Changsha 410000, China; College of Education, Emilio Aguinaldo College, Manila 1000, Philippines (lidongqi123456789@163.com)

also be predicted and potential risks can be effectively prevented.

In order to carry out personalized teaching for different student groups, discover the abnormal behavior of students in time, and evaluate students comprehensively and objectively, this study conducted data mining on the network use behavior and campus behavior log of students in a university's physical education network teaching. The experiment digitally analyzes the focus of students' academic level analysis, and finally proposes a N-Adaptive Boosting (NA) model based on multiple classifiers to predict students' grades.

2. Related Work. In recent years, the rapid development of big data technology has promoted changes again and again, laying a technical foundation for data mining in the field of education. EDM focuses on students and changes the traditional collective education mode to personalized learning mode. This method prepared individual learning reports by recording students' learning behaviors and data in online education. EDM mainly analyzed the potential laws in students' learning process, so as to achieve the goal of promoting students' effective learning [8]. To make effective project decisions, Yahya A A first thoroughly understood the internal relationship and mutual relationship between the project education objectives (PEO) and student outcomes (SO), and proposed a method based on data mining to mine relevant knowledge. In the experiment, Apriori algorithm was applied to the dataset to generate management rules. The experiment finally confirmed the effectiveness of the mined knowledge for engineering education decision-making [9].

To explore the relationship between students' online courses and final exam scores, Kerzic D and others selected first-year undergraduate students from the School of Administration of the University of Ljubljana to carry out the experiment. Orange data mining software was used for two prediction modeling tasks. The research results showed that there was a strong relationship between students' performance in online education tests and their final results [10].

To analyze the utility and applicability of deep learning in EDM and learning analysis, Doleck T and others compared the prediction accuracy of current mainstream deep learning algorithms. The research results showed that the deep learning method showed the same performance as other machine learning methods [11]. Fan J et al. applied data mining technology to the development of university information management system based on the role of modern management in cultivating talents and serving the society. The research results showed that data mining could greatly improve the data analysis ability and management level of managers in the application of university informatization [12].

With the improvement of university information system construction, it has accumulated a huge amount of student learning data. This provides a data basis for the analysis and modeling of students' learning behavior under the condition of big data [13]. However, how to use a large number of student behavior data for modeling to further achieve the analysis and evaluation of academic level is still concerned by many researchers. Joshi A et al. proposed a new integrated machine learning model (CatBoost) to predict students' academic performance. The experimental results showed that the accuracy of the model is 92.27%, which verified its reliability. The proposed model helped educators identify students at early risk [14].

Ade R proposed a classifier that combines fuzzy ARTMAP and Bayesian ARTMAP classifiers, and predicted students' learning achievements. The experimental results verified the good accuracy of this method in predicting students' performance [15]. Deepika K proposed a hybrid feature selection method of random forest (RFBF-FE) based on unused education data, which combined Relief-F and budget tree. Compared with the existing logistic regression model, the SAP accuracy of this method had increased by 6.85% [16]. To improve students' academic performance, Yusuf A proposed a performance prediction model using stack classifier and composite minority oversampling technology. The research results showed that this technology improved the performance of data mining models [17].

To sum up, data mining in the field of education has important practical significance for teaching management and the prediction of students' academic level. With the continuous improvement of online education and university information construction, online education behavior has also become an important influencing factor. In order to explore the impact of college students' online behavior data on the prediction of students' abnormal academic performance, this experiment depicts students' behavior portraits from a new perspective. The experiment uses DBSCAN clustering method to classify different student behavior portraits, and finally constructs a prediction model of student performance anomalies based on multiple classifiers.

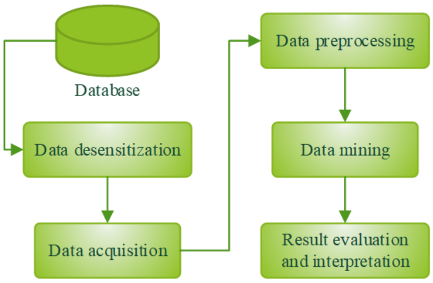


Fig. 3.1: EDM Process

3. The Construction of Student Achievement Anomaly Prediction Model Based on Multiple Classifiers.

3.1. Analysis of College Students' Behavior Data. With the continuous promotion of digital construction in colleges and universities, online education has rapidly entered the plan of college administrators[18]. Student behavior data analysis mainly refers to the use of data mining technology to mine the hidden information and patterns in student behavior log data, and then extract effective features to predict academic level. EDM is a unique branch of the education field. It is a mining technology that uses computer science and data mining technology to obtain special types of data in the education system. This technology extracts and analyzes valuable information under the guidance of psychology and planned behavior, and then discovers students' learning patterns.

EDM inherits the complete process of data mining technology, including data collection, desensitization, preprocessing, mining, and result evaluation and interpretation. See Figure 3.1 for the detailed process. According to data types and mining purposes, EDM methods are mainly divided into four types: clustering algorithm, association algorithm, classification algorithm and regression algorithm. The classification algorithm belongs to supervised learning. It mainly mines the relevance between daily data records and their labels through the known target categories in the data, and classifies them into corresponding categories [19]. After comprehensive consideration, the current mainstream decision tree model (DT), support vector machine (SVM) and integrated learning algorithm (IL) are used in this study [20, 21, 22]. The most classical algorithm in DT is ID3 algorithm. However, in the case of obvious differences and small sample size, the characteristics will be ignored. To solve the above problems, the experiment uses C4.5 algorithm to optimize ID3 algorithm, and uses information gain rate as the standard of feature selection. C4.5 algorithm not only improves the prediction ability of feature points in missing value processing, but also can process and predict discrete and continuous eigenvalues. C4.5 algorithm uses information gain rate to select the optimal partition attribute, see 3.1.

$$\begin{cases} GR(B|A) = \frac{IG(B|A)}{IV(A)} \\ IV(A) = \sum_{m=1}^{M} \frac{|B_m|}{|B|} \log_2 \frac{|B_m|}{|B|} \end{cases} \tag{3.1}$$

In equation 3.1, B is the sample set; $A = a_1, a_2, \ldots, a_m$ has a total of values. If B is divided by A , M branch nodes will be generated. Among them, the m^{th} node contains samples with the value of a_m on all attributes A in B , which is recorded as B_m. $IGB|A)$ is defined as the information gain of attribute A to B. $IV(A)$ is the intrinsic value of property a. The larger the value of M , the greater the value is. C4.5 algorithm can provide effective decision-making for students' behavior analysis, but small changes in data will cause changes in feature selection, and ultimately lead to sudden changes in decision-making logic. SVM is a kind of supervised learning model, which can be used not only for classification problems, but also for nonlinear regression problems. The classification principle of SVM is shown in Figure 3.2.

For linearly separable samples, the calculation of the optimal hyperplane HP is actually a convex quadratic programming problem. In Figure 3.2, there is a sample data set D. And hyperplane HP_1 and HP_2 are expressed

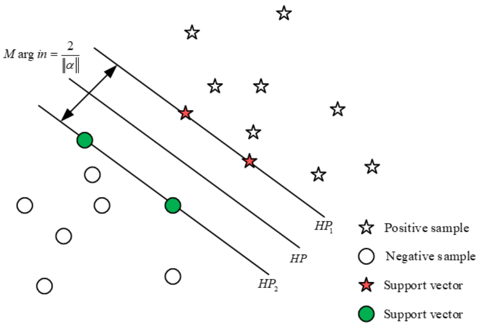


Fig. 3.2: Principle of SVM Model

as equation 3.2.

$$\begin{cases} \text{HP:} & \alpha^T u + g = 0 \\ \text{HP}_1: & \alpha^T u + g = 1 \\ \text{HP}_2: & \alpha^T u + g = -1 \end{cases} \tag{3.2}$$

In equation 3.2, α is the normal vector of the hyperplane; g is the distance between the hyperplane and the coordinate origin. The support vector is consistent with HP_1 and HP_2. The classification interval $Margin$ is the projection of the difference of heterogeneous support vectors at α, as shown in equation 3.3.

$$M = \frac{2}{\|\alpha\|} \tag{3.3}$$

To maximize $Margin$, it is need to solve the convex quadratic programming problem. For optimization problems with constraints, Lagrange function optimization is usually used, that is, adding Lagrange multiplier $\delta_j \geq 0$ to each constraint. The original problem can be transformed into equation 3.4.

$$\begin{cases} \max \sum_{j=1}^{n} \delta_j - \frac{1}{2} \sum_{j=1}^{n} \sum_{k=1}^{n} \delta_j \delta_k v_j v_k u_j u_k \\ \text{s.t.} \quad \sum_{j=1}^{n} \delta_j v_j = 0, \qquad\qquad j = 1, 2, \ldots, n. \end{cases} \tag{3.4}$$

According to equation 3.4, the expression of SVM model can be obtained, see equation 3.5.

$$f(u) = \sum_{j=1}^{n} \delta_j v_j u_j^T u + g \tag{3.5}$$

According to equation 3.2, the result of SVM model is only related to support vector. However, in practical applications, there are many factors that can cause nonlinear classification of sample data. Therefore, nonlinear SVM model came into being. The sample data of low latitude is transformed into high dimensional space through kernel function, so that the sample data becomes linearly separable in high dimensional space. The principle of nonlinear SVM model is as follows. Assuming that there is a mapping function $\zeta(u)$ for vector u , the hyperplane expression can be obtained, as shown in equation 3.6.

$$f(u) = \alpha^T \zeta(u) + g \tag{3.6}$$

Then the operation is similar to linear SVM model, add δ_j . The minimum value problem can be transformed into the maximum value problem under limited conditions. It is difficult to calculate $\zeta(u)$ in the feature space, so the mapping relationship can be transformed. That is, the inner product of u_j and u_k in the mapping space is equal to the function value calculated by function $\psi(u)$ in the original space. Finally, the nonlinear SVM model can be obtained, see equation 3.7.

$$f(u)=\sum_{j=1}^{n}\delta_j v_j \psi(u_j,u_k)+g \tag{3.7}$$

In equation 3.7, $\psi(u)$ is the kernel function. Different kernel functions can be used for micro-mapping in experiments. At present, the commonly used kernel functions include Gaussian kernel function and sigmoid kernel function. The principle of IL algorithm is to generate multiple weak classifiers through training data. Then, according to a certain rule or integration strategy, multiple weak classifiers are combined into a strong classifier, and then the final decision is made. At present, the most common IL algorithms are Bagging algorithm and Boosting algorithm. For the classification problem of class prediction, the IL algorithm integrates the results of the base classifier into a voting strategy, including simple voting and weighted voting. See equation 3.8 for corresponding discrimination results.

$$\begin{cases} L_1(u)L_1(u)=\mathrm{argmax}(N_j) \\ L_2(u)=\mathbf{sign}(\sum_{j=1}^{n}\vartheta_j l_j(u)) \end{cases} \tag{3.8}$$

In equation 3.8, $L_1(u)$ and $L_2(u)$ are the final results of simple voting and weighted voting respectively; N_j is the number of base classifiers whose output result is category j ; ϑ_j is the weight assigned to the j^{th} base classifier; l_j is the judgment result of the j base classifier.

4. Design of Student Achievement Anomaly Prediction Model Based on Distance Optimization and Multiple Classifiers. The data of the study comes from the campus all-in-one card data of a university student and the test result data of online physical education teaching. The study collected relevant data from 53 universities across the province from October 2022 to June 2023 for four grades, making the behavioral analysis results more effective. After data acquisition, desensitization of data records is required. The desensitization process includes five types of fields: name, electronic account, student number, IP address and URL. After data desensitization, data pretreatment is also required. The preprocessing process is as follows: First, delete the vacant data record in the dataset directly; Then filter the duplicate values and keep the first record. The test result data of students' physical education online teaching includes the test results and credits of each course.To carry out a comprehensive assessment and evaluation index of students' academic level, the research selects the test results of compulsory and optional courses in physical education network teaching for weighted average processing. According to the final weighted average score, students are divided into four performance groups, namely, abnormal group, passing group, excellent group and non-excellent group. The abnormal group is the student with less than 60 scores, and the passing group is the student with more than 60 scores; The excellent group is the students with more than 90 points, and vice versa. The abnormal group can detect the students who have the risk of abnormal test results; The excellent group can test the behavioral characteristics of students with excellent academic level. Finally, according to the above standards, students' grades are given corresponding labels. Through the above analysis, we can get three behavioral construction indicators of network behavior, network viscosity and life regularity, and then build a feature library of student behavior portraits. See Table 4.1 for details.

According to the above student portrait description indicators, the research uses the Density-Based Spatial Clustering of Applications with Noise based on Distance Optimization (D-DBSCAN) algorithm to cluster college student groups. This method divides students into groups with different performance differences to explore the differences between students' network behaviors under the condition of different academic achievements. The principle of DBSCAN algorithm is shown in Figure 4.1.

In Figure 4.1, u_1 and u_2 are the core points; u_3 and u_4 are boundary points; u_2 direct from u_1 density; u_3 direct from u_2 density; u_4 direct from u_1 density; u_4 connected to (A^{u_3}) density. DBSCAN algorithm has the following advantages: the clustering process is not affected by the noise in the sample set; The number of clusters does not need to be given in advance; The clustering results are not biased. However, DBSCAN algorithm has the following shortcomings: it is difficult to select the initial parameter neighborhood radius χ

Table 4.1: Three Aspects of Behavioral Evaluation Indicators

Behavior	**Field name**	**Field type**	**Field Description**
-	Network behavior	Enum	Evaluate students' preference for online physical education courses
-	Network stickiness	Enum	Evaluate students' dependence on the network
-	Regularity of life	Enum	Evaluate whether students' self-study is regular
Network behavior	Frequency of sports video courses	Numerical type	Frequency of visits to online physical education courses every month
	Knowledge frequency	Numerical type	Frequency of visiting knowledge websites every month
	Game frequency	Numerical type	Frequency of visiting game websites every month
	Social frequency	Numerical type	Monthly visits to social networking sites
Network stickiness	Online time	Numerical type	Distribution range of online time of students every month
	Online duration	Numerical type	Average online time of students per month
	Online days	Numerical type	Number of days students are online per month
Regularity of life	Self-study duration	Numerical type	Average time spent on self-study in the online library every month
	Number of days to enter the library website	Numerical type	Number of days to access the library website every month

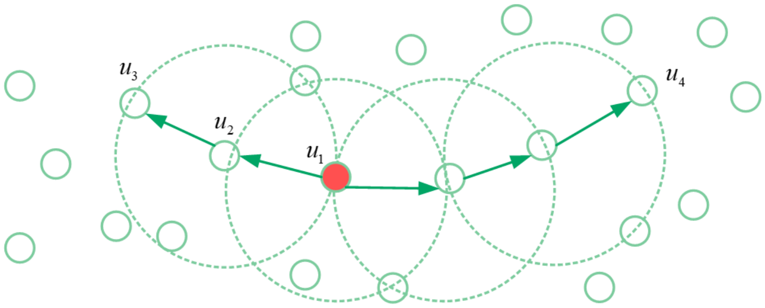


Fig. 4.1: Principle of Dbscan Algorithm

and density threshold *minPts* ; Not suitable for sample sets with uneven density and large distance space; In the case of high dimension of sample set, accurate clustering cannot be achieved. In view of the above problems, the D-DBSCAN algorithm is proposed. The algorithm automatically selects the value of χ according to the characteristics of *minPts* and the data distribution density in the data set. Assuming the existence of sample set $Q = u_1, u_2, \ldots, u_n$, the density can be obtained. For $u_j \in Q$, the density of u_j is in the neighborhood of u_j, and the number of data points is shown in equation 4.1.

$$N(u_j) = \Big\{ u_k \in Q | 0 < distance(u_j, u_k) < \chi \tag{4.1}$$

In equation 4.1, $distance(u_j, u_k)$ is the distance between u_j and u_k. For the sample point u_j in the neighborhood of the core point u_k, the distance coefficient between u_k and $u_j(A)$ can be obtained as equation 4.2.

$$\gamma = \frac{N(u_k)}{N(u_j)} \tag{4.2}$$

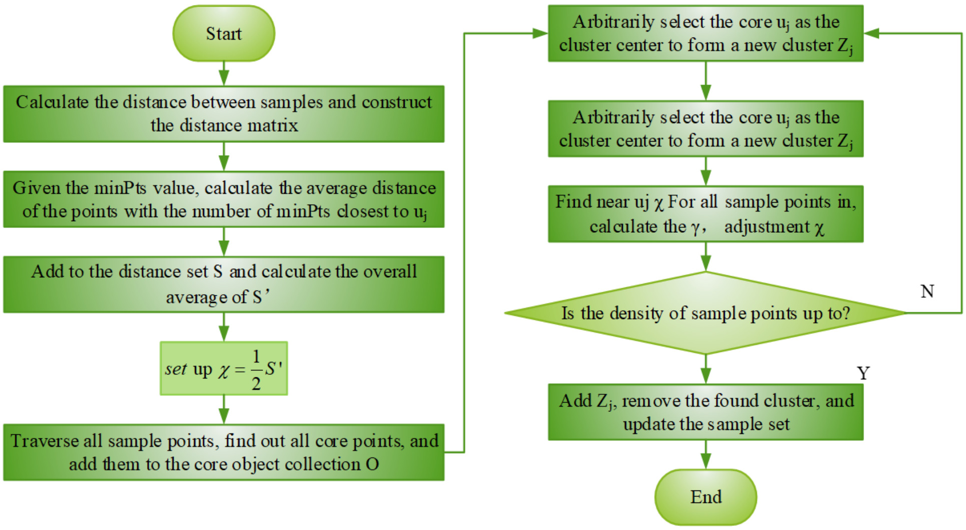

Start
Calculate the distance between samples and construct the distance matrix
Given the minPts value, calculate the average distance of the points with the number of minPts closest to u_j
Add to the distance set S and calculate the overall average of S'
set up $\chi = \frac{1}{2}S'$
Traverse all sample points, find out all core points, and add them to the core object collection O
Arbitrarily select the core u_j as the cluster center to form a new cluster Z_j
Arbitrarily select the core u_j as the cluster center to form a new cluster Z_j
Find near uj χ For all sample points in, calculate the γ, adjustment χ
Is the density of sample points up to?
N
Y
Add Z_j, remove the found cluster, and update the sample set
End

Fig. 4.2: D-DBSCAN Algorithm Flow

According to the above definition, the D-DBSCAN algorithm flow can be obtained, as shown in Figure 4.2.

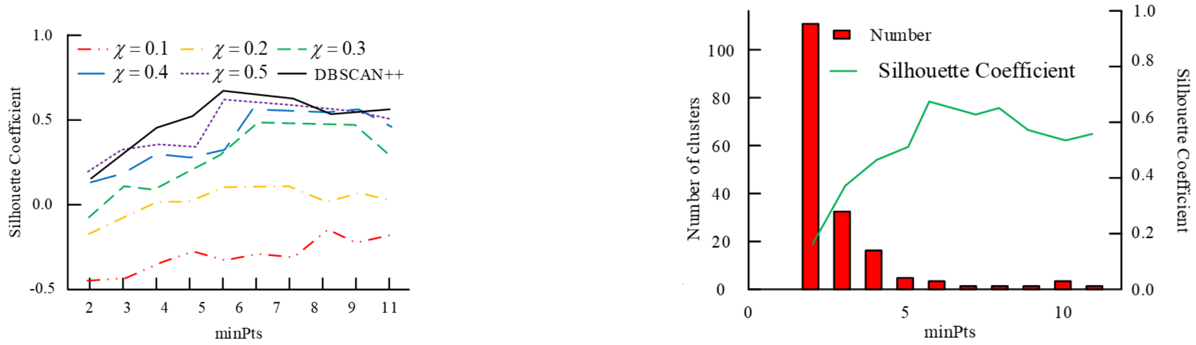
According to the extracted student behavior characteristic data set, the meaning, dimension and order of magnitude of each characteristic index in the multidimensional characteristic data set are different. It is very inappropriate to directly conduct data mining without considering the dimension of feature vectors before knowing the influence of feature vectors on the calculation results. Therefore, in order to ensure the reliability of the results and the validity of the model, before entering the model, it is necessary to standardize the data set of the original features, so that each feature has an equal amount of influence factors in the initial state. The commonly used standardization methods include deviation standardization (DS) and Z-csore standardization. In DS, for sequence $U = u_1, u_2, \dots, u_j$, $j \in 1, 2, \dots, n$, if the characteristic index is the greater the better type index or the smaller the better type index, ewuation 4.3 can be obtained.

$$\begin{cases} v_j = \frac{u_j - \min(U)}{\max(U) - \min(U)} \\ v_j = \frac{\max(U) - u_j}{\max(U) - \min(U)} \end{cases} \tag{4.3}$$

DS method is simple and easy to implement, but it is easy to be affected by outliers or outliers, and easy to increase the calculation amount. The standardized expression of Z-csore is shown in equation 4.4.

$$v_j = \frac{u_j - \bar{u}}{\sigma} \tag{4.4}$$

In equation 4.4, $\bar{u}$ and σ are the average and standard deviation of sequence U respectively. This method is applicable to the case of outliers in the feature data set. AB (Adaptive Boosting, AB) model is one of the most classical algorithms in Boosting model. It adopts the idea of joint decision to improve the classification accuracy. However, due to the same type of base classifier, the model still has the limitations of a single classifier in the learning process. The principle of AB model is studied, and a NA model based on multiple classifiers is established. NA models no longer use a single classifier as a base learner, but integrate multiple classifier models to avoid the problem that similar classifiers perform well in a certain aspect, where N represents the number of classifier models. In the training process, the base learner is composed of multiple classifier models. Each classifier model will learn and classify the training samples, and the obtained training results are decided by several classifier models using simple voting. During each iteration, the training sample data set is used to pass several classifier models and the model is fitted using the same weights. By integrating different classifier models,



(a) Comparison of contour coefficients of two algorithms under different parameters

(b) Values of D-DBSCAN algorithm under different minPts

Fig. 5.1: The Results of the Contour Coefficients of the Two Algorithms and the Values of D-Dbscan under Different

this model overcomes the classification limitations brought by a single learner and makes the performance of the classifier complementary. The flow of NA model is as follows: there is training data set G_j ; The number of iterations is N , and the weight distribution of the initial training sample is shown in equation 4.5.

$$\begin{cases} W_1 = (w_{1,1}, w_{1,2}, \ldots, w_{1,j}) \\ w_{1,j} = \frac{1}{N}, 1, 2, \ldots, N \end{cases} \tag{4.5}$$

Then construct learning algorithm Γ, which is composed of j classifiers $C_j(u)$. In the algorithm, $C_j(u)$ training data are used for classification prediction, and the classification results are counted. Return the final classification result to Γ through a simple voting algorithm. For $n = 1, 2, \ldots, N$, use the training data set with weight W_n to train the base Γ. Input the integrated classifier model $C(u)$ to get the weak classifier $R_n(u)$, see equation 4.6.

$$R_n(u) = \Gamma(G_j, W_n, C(u)) \tag{4.6}$$

Calculate the classification error rate of $R_n(u)$ for the training data set, and then calculate the weight of $R_n(u)$ in the strong classifier according to the classification error rate, and update the weight distribution of the training sample set. After iteration N of the above process, the final classifier result can be obtained, as shown in equation 4.7.

$$F(u) = \text{sign}\left(\sum_{j=1}^{N} \theta_n R_n(u)\right) \tag{4.7}$$

In equation 4.7, θ_n is the proportion of $R_n(u)$ in the strong classifier. According to the prediction results of the algorithm, students' abnormal grades are filtered.

5. Result Analysis of NA Model. To evaluate the clustering effect of the D-DBSCAN algorithm proposed in the study, and determine the optimal number of clusters. Silhouette Coefficent (SC) was selected for evaluation. SC can evaluate the cohesion and separation between sample data points at the same time, and can evaluate the clustering effect when the formal sample data set category is unknown. The research selects DBSCAN algorithm and D-DBSCAN algorithm for comparison, and obtains the results of the contour coefficients of the two algorithms and the values of D-DBSCAN under different *minPts* conditions, as shown in Figure 5.1. It can be seen from Figure 5.1a - that DBSCAN algorithm needs to constantly find the optimal

Table 5.1: Comparison Results of Operation Time of Two Algorithms

Algorithm	Total running time (s)	Total operation times/time	Total operation times/time
DBSCAN algorithm	2617.00	100	26.17
D-DBSCAN algorithm	2904.60	10	290.46

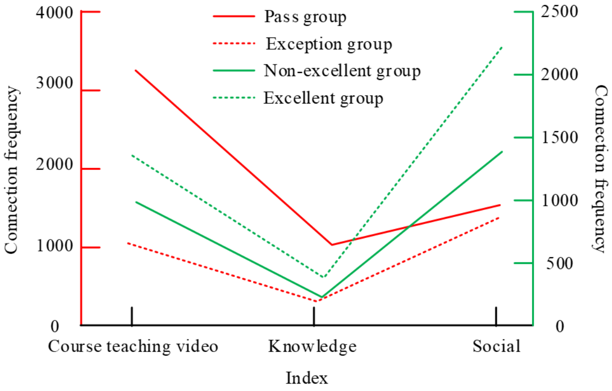


Fig. 5.2: The use of different groups in sports online education

solution of $minPts$ and χ. When $\chi = 0.1$, $minPts$ is any value, and the corresponding SC value is negative. It shows that the clustering effect is the worst at this time. When $\chi = 0.9$ and $minPts = 6$, the SC value is 0.643, and the parameter is the optimal solution. DBSCAN++represents that the algorithm only needs one parameter, so it only needs to find the optimal solution under different $minPts$. D-DBSCAN algorithm has better clustering effect than DBSCAN algorithm for other parameter values except . From Figure 5.1b, when $minPts = 6$, the SC value is 0.711, which is the optimal solution of the D-DBSCAN algorithm. Compared with DBSCAN algorithm, D-DBSCAN clustering performance is improved by 10.6%, and the corresponding number of clusters is 4. The results show that when clustering students' behavior characteristics, the number of clusters is 4, and the corresponding SC value is 0.711, which is the best clustering effect.

Table 5.1 shows the comparison results of the operation time of the two algorithms. It can be seen from Table 2 that DBSCAN algorithm performs 100 operations on the two parameters and ; D-DBSCAN algorithm performs 10 operations on to obtain better clustering results. The average single run time of D-DBSCAN algorithm is 11.10 times that of DBSCAN algorithm, but its total run time is 1.11 times that of DBSCAN algorithm, and the calculation time is 9% longer than DBSCAN algorithm. The above results are due to the fact that compared to other algorithms, the D-DBSCAN algorithm proposed in the study can discover clusters with different shapes and adaptively select appropriate neighborhood radii, making it more suitable for analyzing complex academic behavior of students in universities.

See Figure 5.2 for the results of the use of different groups in sports online education. In terms of connection frequency between the abnormal group and the passing group, the passing students have more access to the network physical education curriculum resources than the abnormal students; In terms of social behavior in online classroom, the two groups of students visited the same. In the comparison between the excellent group and the non-excellent group, the excellent students have the highest connection frequency in the social aspects of sports online education class, and the connection frequency in sports knowledge is 1.7 times of the non-excellent students.

Figure 5.3 shows the results of online duration of online physical education courses for different groups of students each month. The students in the abnormal group generally spend less time in physical education courses than the students in the passing group. During the four months of learning, the online duration of

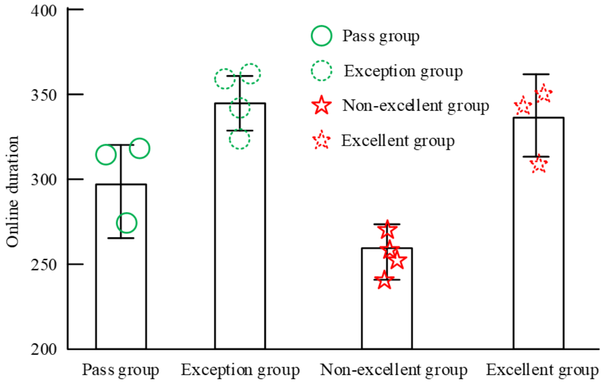


Fig. 5.3: Monthly Online Duration Results of Online Physical Education Courses For Students in Different Groups

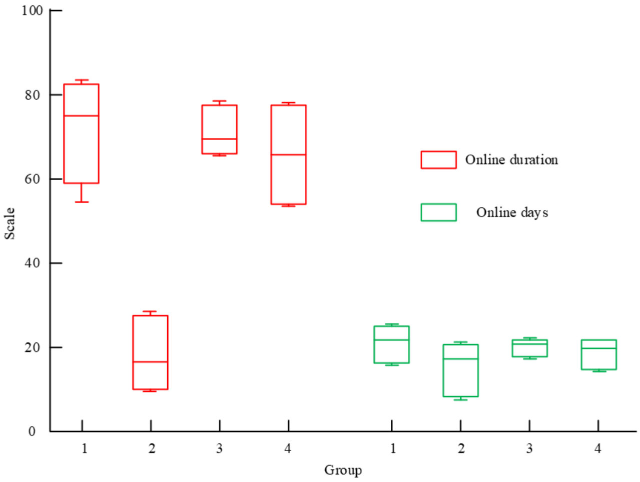
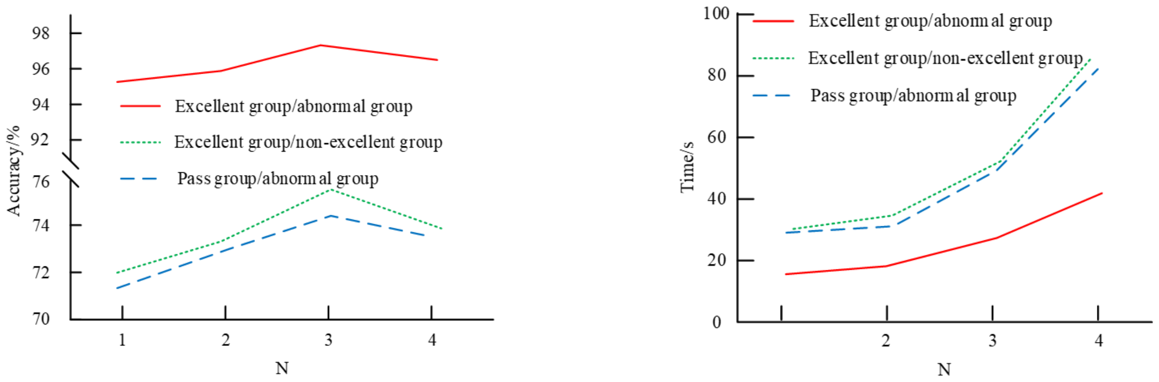


Fig. 5.4: Index Chart of Network Self-Study Of Different Groups Of Students in the Library

students in the abnormal group was more scattered, and the online duration of students in the abnormal group was significantly reduced near the test. The average online duration of the excellent group is the highest, and the distribution is more concentrated, and the overall performance of the semester is also more stable.

Figure 5.4 shows the results of relevant indicators of the self-study network of different groups of students in the library every month. 1-4 corresponds to the passing group, abnormal group, excellent group and non-excellent group respectively. The students in the abnormal group have significantly lower online duration and online days of self-study, which indicates that the students in the abnormal group will not have less time to self-study. The students in the excellent group showed lower variance in both indicators. It shows that the data of the excellent group converges, indicating that the behavior pattern of the excellent students is more fixed. From the overall performance analysis, the difference between the abnormal group and the qualified group is more obvious. Therefore, it is easier to find students with abnormal results in the test in the subsequent prediction.

To verify the performance of the model proposed in the study, the study used the tenfold cross-validation method for training evaluation. The experiment selects the average of accuracy and F1 value as the final result.



(a) (b)

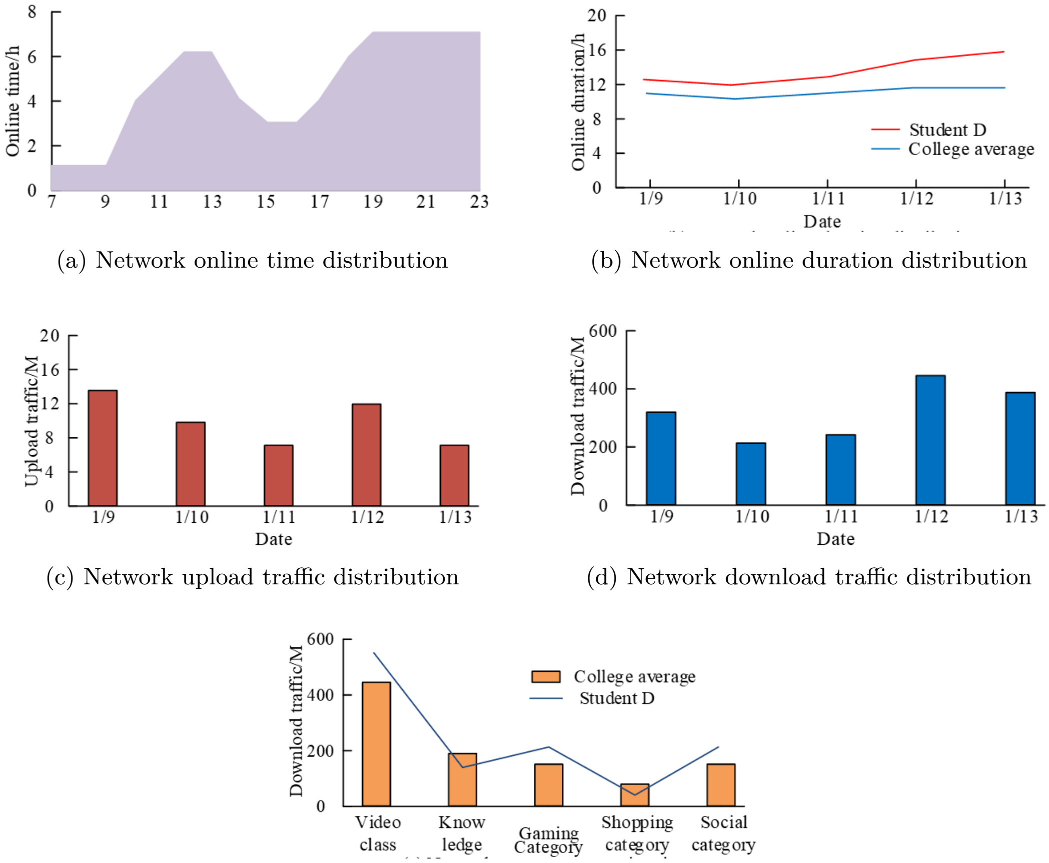
Fig. 5.5: Prediction Accuracy and Consumption Time Results of Each Group of Samples at Different N

Table 5.2: Prediction Results of Different Models for Each Group in Physical Education Network Teaching Test

Group	Model	Accuracy (%)	Precision (%)	Recall (%)	F1
Abnormal group/pass group	C4.5	66.34	79.89	68.30	0.743
	SVM	67.57	83.20	69.06	0.756
	IL	60.18	77.03	63.20	0.691
	NA	97.93	98.63	83.27	0.801
Abnormal group/excellent group	C4.5	72.83	70.42	75.63	0.743
	SVM	73.35	73.86	77.49	0.757
	IL	65.40	66.34	69.73	0.672
	NA	98.16	98.58	97.26	0.958

Figure 9 shows the prediction accuracy and consumption time results of each group of samples at different N. Figure 5.5a- shows that in the abnormal group and the qualified group, the excellent group and the non-excellent group, the prediction accuracy rate is the highest when N=3, but the integration rate is lower when compared with N=2. In the excellent group and the abnormal group, the prediction accuracy is the highest when N=3. Figure 5.5b shows that when N=3, it takes 50 seconds. However, combined with the accuracy curve, when N=2, the accuracy can be effectively improved in the case of similar time consumption. Therefore, select N=2, that is, the base classifier is composed of C4.5 model and SVM model, as the NA model. By comparing the results of multiple experiments mentioned above, the effectiveness of the NA model in predicting academic performance in online physical education teaching in universities can be determined. By studying the impact of the value of N in the NA model, a solid data foundation has been laid for predicting academic performance in a wider range of related universities in the future.

Table 5.2 shows the prediction results of different models of each group in the physical education network teaching test. Compared with the other three models, the accuracy of NA model was 97.93% for the students in the abnormal/passing group. The recall rate of SVM represents that 69.06% of the positive samples are accurately predicted; The NA model iteratively trains the base learner, so its accuracy is 31.59% and 30.36% higher than C4.5 model and SVM model. The F1 value of NA model is also the highest, 0.801. In the abnormal group/excellent group, the accuracy rate of the model was 98.16%; The recall rate was 97.26%; F1 value is 0.958. The research results show that the integrated learning algorithm performs better than the single prediction model in the prediction of sports network teaching results, and can accurately predict students' abnormal results. The above results are due to the fact that compared to a single prediction model, the NA



(a) Network online time distribution

(b) Network online duration distribution

(c) Network upload traffic distribution

(d) Network download traffic distribution

(e) Network resource access situation

Fig. 5.6: Data Analysis and Score Risk Prediction Results for Student D

model proposed in the study is no longer limited to the use of a single algorithm. Instead, based on the characteristics exhibited by behavioral data, a NA model composed of multiple classifiers is used for prediction analysis, complementing the advantages of different classification models, improving the shortcomings of the model, and thereby improving the prediction accuracy of the model, which is better used for predicting grades in online physical education teaching in universities. To further validate the effectiveness of the NA model proposed in the study in practical applications, the study randomly selected student D from a certain university for analysis and score risk prediction.

The data analysis and performance risk prediction results of student D are shown in Figure 5.6. From Figure 5.6, it can be observed that the time distribution of student D's online presence during the day is maintained for a long time from 11:00 to 13:00 and from 19:00 to 23:00. And through its network distribution in the Xi'an market within a week, it was found that the student's average online duration was significantly higher than that of the college. In the online upload and download traffic of student D, it was found that the usage was highest on January 12th and lower on January 11th. Through the access to network resources, it was found that video accounted for the most at 23%, while shopping and knowledge accounted for the least at 2% and 7%, respectively. Based on the above analysis of results, the network usage of student D is high, the network viscosity is severe, and their life is irregular. The predicted result is a failure. Through the above analysis and risk warning, administrators can save the risk list of failing students for further offline communication and communication, improve their non-standard life behavior, and enhance their academic level. In summary, the application of NA model in online teaching and testing of physical education in universities can more efficiently manage and communicate with different students, improve their academic performance, and provide higher quality students to society.

6. Conclusion. The development of education informatization has made a huge collection of education data. How to mine valuable information from a large number of data and accurately classify and predict the students with abnormal results in the physical education network teaching test is an important means to help managers make scientific decisions. This research was based on the feature library of student behavior portrait, and used D-DBSCAN algorithm for clustering analysis. The experiment constructed a multi-classification based NA model to predict the abnormal scores of students in college physical education network test. The experimental results showed that when, the SC value was 0.711, which was the optimal solution of D-DBSCAN algorithm. At this time, compared with DBSCAN algorithm, D-DBSCAN clustering performance was improved by 10.6%, and the corresponding number of clusters was 4. When N=2, that was, the base classifier of NA model was composed of C4.5 model and SVM model, the prediction accuracy and time consumption were the most appropriate. Compared with C4.5 model, SVM model and IL model, the accuracy, recall rate and F1 value of NA model were 98.16%, 97.26% and 0.958 respectively. To sum up, the NA model based on classifiers proposed in the study had better performance and could accurately predict the abnormal performance of students in college physical education network teaching test. However, there are still shortcomings in this study. For example, the amount of research data is not very large. With the development of high-performance computing technology, high-performance platforms can be used to build a distributed cluster environment in future research. And then realize the parallelization of student behavior data processing and calculation, and improve the operation efficiency of the overall model.

REFERENCES

[1] Vatsalan, D., Rakotoarivelo, T., Tyler, P., Ladjalet, D. & Bhaskar, R. Privacy risk quantification in education data using Markov model. *British Journal Of Educational Technology.* **53**, 804-821 (2022)

[2] Dhanalakshmi, R., Muthukumar, B. & Canessane, R. Analysis of Special Children Education Using Data Mining Approach. *International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems.* **30** pp. 125-140 (2022)

[3] Banerjee, D., Gidwani, C. & Rao, T. The role of "Attributions" in social psychology and their relevance in psychosocial health: A narrative review[J]. *Indian Journal Of Social Psychiatry.* **36**, 277-283 (2021)

[4] Qi, S. Approaches to Information Service and Management Construction in University Libraries. *International Journal Of Social Science And Education Research.* **2**, 41-45 (2019)

[5] Gj, A., Dw, A., Dy, A., Dha, B., Qd, A., Role, W. & Domain-Based, O. Access Control Model for Graduate Education Information System. *Procedia Computer Science.* **176** pp. 1241-1250 (2020)

[6] Wang, S., Zhang, F., Gong, Q., Bolati, D. & Ding, J. Research on PBL teaching of immunology based on network teaching platform. *Procedia Computer Science.* **183**, 750-753 (2021)

[7] Grin, N. INFORMATIZATION IN EDUCATION. *Scientific Papers Collection Of The Angarsk State Technical University.*, 348-351

[8] Sang, H. Analysis and Research of Psychological Education Based on Data Mining Technology. *Security And Communication Networks.* **2021**, 1-8 (2021)

[9] Yahya, A. & Osman, A. data-mining-based approach to informed decision-making in engineering education. *Computer Applications In Engineering Education.* **27**, 1402-1418 (2019)

[10] Kerzic, D., Aristovnik, A., Tomazevic, N. & Lan, U. Assessing the impact of students' activities in e-courses on learning outcomes: a data mining approach. *Interactive Technology And Smart Education.* **16**, 117-129 (2019)

[11] Doleck, T., Lemay, D., Basnet, R. & Bazelais, P. Predictive analytics in education: a comparison of deep learning frameworks. *Education And Information Technologies.* **25**, 1951-1963 (2020)

[12] Fan, J., Zhang, M., Sharma, A. & Kukkar, A. Data mining applications in university information management system development. *Journal Of Intelligent Systems.* **31**, 207-220 (2022)

[13] Ramanathan, K. & Thangavel, B. Minkowski Sommon Feature Map-based Densely Connected Deep Convolution Network with LSTM for academic performance prediction. *Concurrency And Computation Practice And Experience.* **33** pp. 4 (2021)

[14] Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D. & Khanna, A. CatBoost - An Ensemble Machine Learning Model for Prediction and Classification of Student Academic Performance. *Advances In Data Science And Adaptive Analysis: Theory And Applications.* **13**, 1-21410 (2021)

[15] Ade, R. Students performance prediction using hybrid classifier technique in incremental learning. *International Journal Of Business Intelligence And Data Mining.* **15**, 173-189 (2019)

[16] Deepika, K., Relief-F, S. & Tree, B. Random Forest Based Feature Selection for Student Academic Performance Prediction. *International Journal Of Intelligent Engineering And Systems.* **12**, 30-39 (2019)

[17] Yusuf, A. & And, J. and synthetic minority oversampling techniques for academic performance prediction. *International Journal Of Informatics And Communication Technology (IJ-ICT).* **8**, 122-127 (2019)

[18] Ns, A., Maa, B., Rb, C. & Hh, D. The effect of the virtual social network-based psycho-education on the hope of family caregivers of clients with severe mental disorders. *Archives Of Psychiatric Nursing.* **35**, 290-295 (2021)

[19] Envelope, D., Akbar, M., Bagaskara, A. & Vinarti, R. Improving classification algorithm on education dataset using hyper-parameter tuning - ScienceDirect. *Procedia Computer Science.* **197** pp. 538-544 (2022)

[20] An, Y. & Zhou, H. Short term effect evaluation model of rural energy construction revitalization based on ID3 decision tree algorithm. *Energy Reports.* **8** pp. 1004-1012 (2022)

[21] Sun, F. & Shi, G. Study on the application of big data techniques for the third-party logistics using novel support vector machine algorithm. *Journal Of Enterprise Information Management.* **35**, 1168-1184 (2022)

[22] Tan, F. & Xie, X. Recognition Technology of Athlete's Limb Movement Combined Based on the Integrated Learning Algorithm. *Journal Of Sensors.* **7557**, 1-30575 (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 16, 2023
Accepted: Nov 16, 2023