Predicting Student Performance Using Optimized DBSCAN-K-Means Clustering and RBF Neural Networks

Jia Wang
Department of Information Technology, Shanxi Professional College of Finance, Taiyuan 030000, China
E-mail: wangjia0313@126.com

Keywords: data mining, RBF, grade prediction, cluster analysis, balanced discriminant function

Received: June 28, 2024

A student performance prediction model based on an optimized density-based spatial clustering of applications with noise (DBSCAN)-K-means clustering method and radial basis function neural network (RBFNN) has been proposed. Firstly, in response to the problem of traditional K-means algorithm requiring manual selection of K values in clustering analysis and easily falling into the curse of data dimensionality, this paper introduced the DBSCAN idea for optimization, which improved clustering efficiency and accuracy. Secondly, to solve the problem of low efficiency and non-optimal parameter settings in RBFNN, this paper adopted an improved K-means algorithm and a balanced discriminant function to determine the number of network center points and hidden layer nodes in RBFNN, optimizing the network structure. By inputting the optimized clustering results into RBFNN, high-precision prediction of students' final grades was achieved. The experimental results showed that Model 1 only needed 58 iterations to achieve the target accuracy during the training process, with a high fitting accuracy of 0.957, an accuracy rate of 99.2%, and an average prediction error rate of 0.03. Compared with existing advanced methods, this model shows significant advantages in both prediction accuracy and efficiency, which helps teachers obtain more intuitive teaching feedback, thereby improving teaching methods and enhancing students' learning outcomes.

Povzetek: Študija uvaja model za napoved uspešnosti študentov z optimizirano kombinacijo DBSCAN-K-means gručenja in RBF nevronskih mrež.

1 Introduction

During the epidemic, the teaching mode of online education has become the mainstream. Lots of learning data will be generated [1]. Through data mining (DM), effective information and rules can be obtained from it, so as to predict academic performance. The prediction of academic can facilitate teachers to understand students' learning status, obtain intuitive teaching feedback, and improve teaching methods and teaching effects [2-3]. Cluster analysis is an important means to realize learning DM and predict students' academic performance. The study uses the K-means to mine the student characteristic data. However, the K-means is prone to fall into the disaster of data dimensionality, and needs to manually select the K value, so the clustering accuracy is not high [3-4]. In response to this defect, the study introduces density-based spatial clustering of applications with noise (DBSCAN) to optimize it to improve clustering efficiency and clustering effect. After the characteristics of the student learning data are obtained through cluster analysis, they are input into the radial basis function neural network (RBFNN) for training. The cluster center obtained by the improved K-means is used as the center of the neural network to optimize the network structure of RBFNN. Using the improved RBFNN to classify the output results can realize the prediction of student performance. There are two innovations in the research. The first point is to use DM technology to analyze the characteristics of students' learning data, so as to realize the prediction of students' grades. The second point is to use the clustering center obtained by the improved K-means as the center of the neural network to optimize the RBFNN and improve the classification accuracy. The research predicts students' academic performance more accurately and effectively, which helps teachers obtain more intuitive feedback, so as to carry out targeted educational intervention and teaching improvement, and improve teaching quality and student performance.

2 Related works

Today, with the highly developed Internet technology, there are massive amounts of data in various fields. Using artificial intelligence technology to perform feature mining and association analysis on these data, useful information and rules can be obtained from it, helping users to apply data scientifically and rationally, and improving data utilization efficiency. Nowadays, DM technology has important applications in various fields. Yang et al. introduced several common medical databases and several DM technologies applied to medical databases in their research, providing clinical researchers with a way to quickly understand and apply big data technology [5]. Sungheetha and Rajendran applied DM

technology to the analysis and perturbation of big data. This method had excellent performance in terms of attack resistance, scalability, accuracy, etc., and can effectively protect user privacy [6]. Kumar discussed the application effect of DM technology in the financial field, and built a marketing decision support system based on decision tree to provide support for the decision-making of relevant staff [7]. Sunhare et al. transformed the lots of data accumulated in the Internet of Things (IoT) into useful information and knowledge through DM, making it play an important role in intelligent decision-making and resource optimization. Finally, the study explored the key role of DM in the IoT environment [8]. Mengash analyzed students' learning characteristics based on DM technology, and used artificial neural network (ANN) technology to build models to predict students' performance. Using the student data of Saudi public universities from 2016 to 2019 to verify the model, it was found that the prediction precision rate of the model exceeded 79%, which was exceed other models [4]. Shakya, aimed at the problem that solar power plants are prone to faults and defects, and the efficiency of manual monitoring is low and difficult, which affects the efficiency of power plants, built a self-monitoring system for solar power plants Based on DM technology. The system could obtain and analyze relevant data from the IoT, so as to predict and monitor the status of the solar power plant, quickly monitor the fault area, and improve the power generation efficiency of the power plant [9]. Edastama et al. used the apriori to mine and analyze the glasses transaction data, so as to help enterprises formulate marketing plans and increase the sales of glasses [10]. Sirichanya and Kraisak discussed several common semantic DM methods in the information age and the contributions of these semantic DM methods in various fields, and introduced in detail the framework construction of semantic DM in the application process [11].

RBFNN is a feedforward neural network proposed at the end of the last century. It has the advantages of high efficiency, less time required for training, strong data processing ability, and strong applicability. Because of this, RBFNN has been favored by many researchers and has been applied to various fields. Zhou et al. proposed an RBF-based state-related autoregressive (RBF-AR) model with regression weights, and used this model to realize online estimation of parameters. Simulation results denoted taht the online parameter estimation precision rate of the model exceeded 90% [12]. Li et al. proposed RBF Extreme Learning Machine (RBF-ELM) for the existing hybrid data classification model that did not consider the compatibility of data processing and data classification, resulting in poor data classification effect. model to achieve efficient classification of mixed data. Results showed that on 34 data sets, this model was better than that of several existing data classification models [13]. Yang et al. proposed an improved RBFNN model and applied the model to solar power generation prediction. Through the simulation analysis of the real data of a solar power station in the Netherlands, it was found that the prediction precision rate of the model was high, and it could effectively realize the prediction of solar power generation [14]. Chen et al. used the wavelet model to filter the meteorological data and constructed an RBF-long short-term memory model (RBF-LSTM) to make real-time predictions of local PM2.5. During the experiment, the researchers used data in Taiwan to verify the precision rate of the model [15]. Sun et al. designed a supervisory controller based on RBF and applied it to the supervisory control of maglev vehicles on elastic tracks. Results validated the effectiveness and robustness of the proposed approach in terms of latency [16]. Izonin et al. discussed the problem that the health decision support system is difficult to handle lots of classification and regression tasks when the data is limited in the short medical data set, and proposed a data analysis regression method based on the improved RBF [17]. Under the premise of considering the effect of gravity and constant payload, Liu et al. proposed a control scheme of RBF manipulator based on adaptive bias. In the simulation experiment, the controllability of the scheme to the robot manipulator was verified [18]. Fath et al combined multi-layer perceptron (MLP) and RBFNN to construct the MLP-RBFNN model. Based on the MLP-RBFNN model, a solution gas-oil ratio prediction model of the crude oil system was constructed to evaluate the pressure-volume-temperature characteristics of the reservoir oil and ensure the safety of the system. Outcomes indicated that the model had higher accuracy and efficiency than existing models [19].

In the above discussion, the application of DM technology and RBFNN is very extensive and mature, and lots of research results have been derived. Among them, some studies explore the application of RBF and DM technology in the field of education, including the mining and analysis of students' learning behavior characteristics and performance prediction. However, these research results have a common problem, that is, the efficiency of the model is low, and the prediction accuracy is not ideal. To this end, the study proposes an improved K-means for DM, and uses the improved K-means to optimize RBFNN. It aims to build a student performance prediction model, help teachers obtain more intuitive teaching feedback, and then improve teaching quality and student learning score.

Table 1: Summary of related work

Authors	Methods	Dataset	Results	Key contributions
Mengash [4]	ANN technology	Saudi Arabian public university	Prediction accuracy exceeds 79%	Student learning feature analysis and performance

		student data		prediction model
Yang et al. [5]	DM technology	Medical database	Quickly understand the big data technology approach	Introduction to DM technology for medical databases
Sungheetha and Rajendran [6]	DM applied to big data analysis	Big data	Anti-attack, scalability, and excellent accuracy	DM methods for user privacy protection
Kumar [7]	ANN and decision tree	Financial data	Marketing decision support system	Application of DM in the financial field
Sunhare et al. [8]	DM	IoT data	Intelligent decision-making, optimized resources	Conversion of data into information and knowledge in the IoT
Shakya [9]	DM	Solar power plant data	Quickly monitor fault areas	Solar power station self-monitoring system
Edastama et al.[10]	Apriori algorithm	Glasses transaction data	Marketing plan formulation	Increase in sales of glasses
Sirichanya and Kraisak [11]	Semantic DM method	Multiple fields	Semantic DM contribution and framework	Introduction to semantic DM methods in the information age
Zhou et al. [12]	RBF-AR model	State-related data	Online parameter estimation accuracy exceeds 90%	Parameter online estimation method
Li et al. [13]	RBF-ELM model	Mixed data	Efficient classification	Mixed data classification model
Yang et al.[14]	Improved RBFNN model	Solar power generation data	Efficient classification	Short-term solar power generation prediction
Chen and Li [15]	RBF-LSTM model	Meteorological data	High prediction accuracy	Meteorological data filtering and prediction model
Sun et al. [16]	RBF-based supervisory controller	Magnetic levitation vehicle data	Real-time PM2.5 prediction	Magnetic levitation vehicle control under network delay
Izonin et al. [17]	Improved RBF data analysis regression method	Short medical data set	Validity and robustness verification	Health decision support system for short medical data collection
Liu et al. [18]	RBF with adaptive bias	Robot control	Validity verification of data analysis regression method	Manipulator control scheme
Fath et al. [19]	MLP-RBFNN model	Crude oil system data	Verification of robot manipulator controllability Accuracy and efficiency improvement	Prediction model for gas-oil ratio of crude oil system solution

In Table 1, although existing SOTA methods have made some progress in DM and neural network applications, they still have shortcomings in efficiency and prediction accuracy. Especially when dealing with large-scale datasets, existing models often require a significant amount of computing resources and time, and there is still room for improvement in the accuracy of their prediction results. In addition, the parameter setting

process of RBFNN usually relies on manual experiments and experience, which is not only inefficient but also difficult to ensure finding the optimal solution. To address these issues, this study proposes a student performance prediction model based on an improved K-means algorithm and RBFNN. By combining the ideas of DBSCAN to optimize the K-means algorithm, the efficiency and accuracy of clustering analysis can be improved. In addition, optimizing the parameters of RBFNN using improved K-means algorithm and balanced discriminant function can further improve the training efficiency and prediction accuracy of the model.

3 Construction of academic performance prediction model based on improved RBFNN

3.1 Data preprocessing

In the process of online learning, a large amount of data containing students' learning behaviors and grades is generated. To ensure the accuracy and efficiency of DM, this paper first thoroughly cleans the data and removes null values, outliers, and duplicate records to ensure the consistency and reliability of the data set. Subsequently, considering the privacy of the data and the relevance to the research objectives, this paper removes personal information such as student names, ID numbers, and attributes that have a low correlation with academic performance prediction, such as political affiliation and home address. The preprocessing flow of online learning data is in Figure 1.

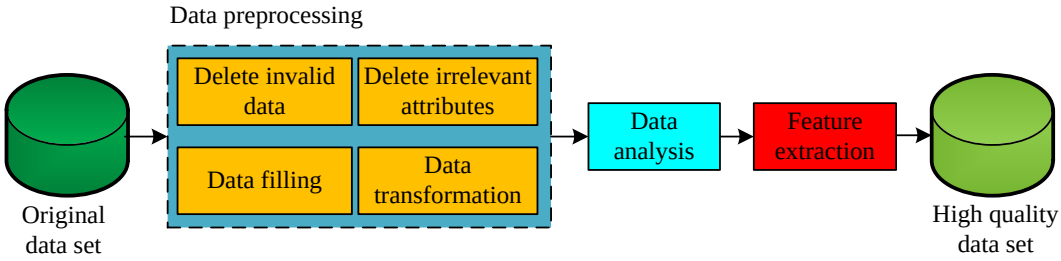

Figure 1: Preprocessing of online learning data

To unify the metrics of the data and reduce the computational complexity, this paper digitizes and standardizes the remaining numerical data, and uses the minimum-maximum normalization method to scale the data to the [0, 1] interval. This processing not only solves the problem of comparing data of different dimensions and magnitudes, but also facilitates subsequent model training. Furthermore, this paper analyzes the correlation between each learning behavior feature and the final grade by calculating the Pearson correlation coefficient. Based on this analysis, this paper retains features that are strongly correlated with the final grade, such as the number of learning behaviors, online days, and learning completion, as well as moderately correlated learning time, to ensure that the model can capture the key factors that affect students' grades. Before preprocessing, the presentation form of the data table is in Table 2.

Table 2: Data sheet form before preprocessing

Attribute	Code	Student ID			
		1	2	3	4
Student ID	xh	20187762	20181403	20183342	20180326
Name	name	-	-	-	-
Gender	sex	f	m	m	f
Age	age	20	twenty-one	twenty-two	20
Identity cards	id_card	-	-	-	-
Region	local	-	-	-	-
Study major	major	3	5	1	12
Political outlook	zzmm	2	1	1	1
Semester	semester	Autumn 2020	Autumn 2020	Autumn 2020	Autumn 2020
Credit	xuefen	81	76	85	78
Learning duration	Study_time	12	8	15	9
Number of exams	ks_num _	3	3	3	3
Examination results	ks_grade _	82	79	81	64
Posts	post_num	84	72	86	64
Replies	repost_num	40	36	44	35
Attend a branch school	jdfx	10633	10633	10633	10633
Student status	xj	1	1	1	1
final exam	wxya	86	78	91	73

Number of studies	activity_num	36	25	54	20
Percentage of courses completed	wxya	90%	85%	100%	60%

After completing preprocessing, the data table presentation form of these 4 students is in Table 3.

Table 3: Data sheet form after preprocessing

Attribute	Code	Student ID			
		1	2	3	4
Student ID	xh	20187762	20181403	20183342	20180326
Name	name	-	-	-	-
Gender	sex	f	m	m	f
Age	age	20	twenty-one	twenty-two	20
Learning duration	Study_time	12	8	15	9
Number of exams	ks_num	3	3	3	3
Examination results	ks_grade	82	79	81	64
Posts	post_num	84	72	86	64
Replies	repost_num	40	36	44	35
Student status	xj	1	1	1	1
Final exam	wxya	86	78	91	73
Number of studies	activity_num	36	25	54	20
Percentage of courses completed	wxya	90%	85%	100%	60%

After preprocessing, the amount of data is reduced by nearly half. In this way, the data is greatly simplified and the amount of subsequent model calculations is reduced. In the original data set, the units, magnitudes, and dimensions of many data are not uniform, or there are cases where the data types are different. Therefore, these data cannot be calculated uniformly, which increases the difficulty of calculation. To solve this problem, the original data is numericalized and standardized. This step is implemented using the minimum-maximum normalization method, as shown in formula (1).

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

In the formula (1), x_i represents the i th data value of the first category; $x_{\max}$ represents the maximum value in the sample; $x_{\min}$ represents the minimum value in the sample; y_i represents the learning effect, that is, the final grade. Based on the above content, the preprocessing of the data of the original data set is completed.

3.2 Learning behavior feature extraction algorithm based on improved K-Means

The Pearson correlation coefficient is utilized to analyze the correlation between the relevant attributes and the final grade, as shown in formula (2).

$$r = \frac{n\sum x_i y_i - \sum x_i y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}} \tag{2}$$

In formula (2), r is the Pearson correlation coefficient. n indicates the number of samples. The Pearson correlation analysis is carried out on the students' learning behavior and final grades. By extracting these learning behavior characteristics and passing certain rules, the prediction of students' academic performance can be realized. The feature extraction algorithm based on K-means is utilized to calculate the characteristic attributes of students' learning behavior, as shown in formula (3).

$$w_i = \sum_{x\in H(c)} -\frac{diff\left(t, d_i, x\right)}{nl} + \sum_{c\notin class(d_i)}\left[\frac{p(c)}{1-p\left(class\left(d_i\right)\right)}\sum_{x\in M(c)} diff\left(t, d_i, x\right)\right] / nl \tag{3}$$

In formula (3), w_i is the weight t of the first type of data; i represents the characteristics; d_i represents the first type i of data object; n represents the number of sampling; l represents the number of samples $H(c)$ and $M(c)$; c indicates the category of the data object; p indicates the probability of occurrence. If the feature is a continuous feature, there is formula (4).

$$diff\left(t, d_i, d_j\right) = \left|\frac{d_i - d_j}{\max_t - \min_t}\right| \left(1 \le i \ne j\right) \tag{4}$$

If the feature is a discrete feature, there is formula (5).

$$diff\left(t, d_i, d_j\right) = \begin{cases} 0 & d_i = d_j \\ 1 & d_i \ne d_j \end{cases} \tag{5}$$

The core idea of the above content is to use the

Euclidean norm to calculate the distance, and to evaluate the correlation between learning behavior and students' final grades based on the distance. Then data in different categories of learning behaviors are selected, and the weight of each category of learning behaviors is calculated. It selects the learning behavior characteristics that have a strong correlation with the final grades, randomly selects the student's data samples for weight calculation, calculates 20 times in total, and takes the average of the 20 calculation results as the weight of the learning feature. The general K-means is easy to fall into the disaster of data dimensionality, and it needs to manually select the K value. The clustering efficiency and clustering accuracy are not ideal. In response to this problem, the study proposes strategies to optimize and improve it. First, refer to the DBSCAN method, it changes the traditional K-means to randomly select the cluster center strategy, and select high-density points as the cluster center. To achieve the above goals, the distance between data objects needs to be measured by Euclidean distance, as denoted in formula (6).

$$d\left(x_i, x_j\right)=\left(\left(x_{i1}-x_{j1}\right)^2+\ldots+\left(x_{ip}-x_{jp}\right)^2\right)^{\frac{1}{2}} \quad (6)$$

In formula (6), x_i, x_j denote the i th and j th data objects. At this time, the definition of the Euclidean distance density function of the *density* data object in the dataset x_i is as indicated in formula (7).

$$density\left(x_i\right)=\sum_{j=1}^{n}\left(d\left(x_i, x_j\right)\sum_{i=1}^{n}d\left(x_i, x_j\right)^{-1}\right) \quad (7)$$

If there is a data object x_i whose area radius is R_i, when the data object is used as the center of the circle, the calculation method of the density of data points in the circular area $Y\left(x_i\right)$ is shown in formula (8).

$$Y\left(x_i\right)=\left|\left\{p\left|d\left(x_1, p\right)\le R_i, p\in X\right.\right\}\right| \quad (8)$$

In formula (8), X is the dataset where the data object x_i is located. The average density of data points in this dataset can be deduced from formulas (6), (7), and (8), as shown in formula (9).

$$PY\left(x\right)=\frac{1}{n}\sum_{x\in X}Y\left(x\right) \quad (9)$$

In summary, the process of improving the K-means is in Figure 2.

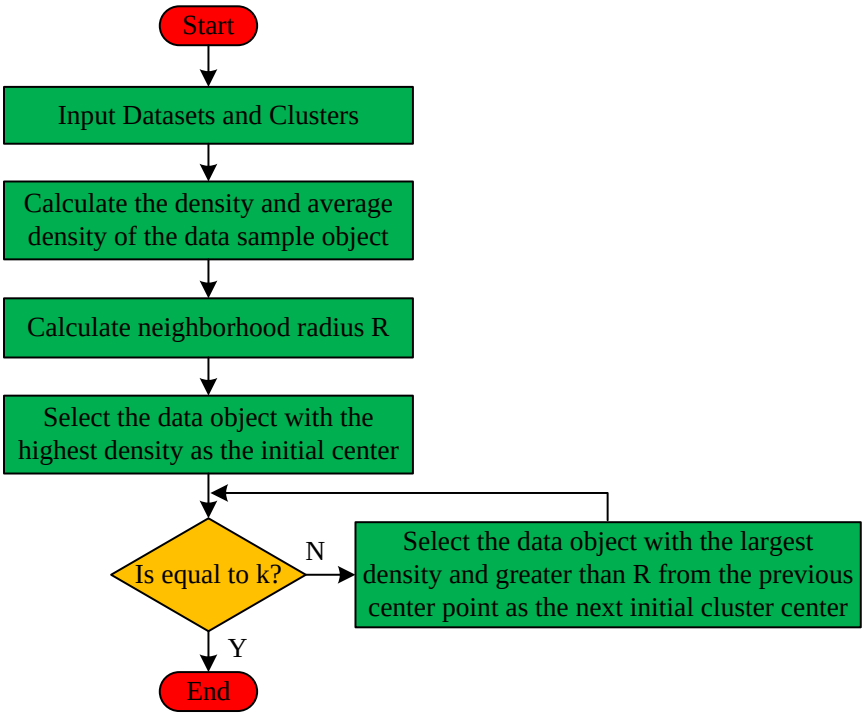

Figure 2: The flow of the improved K-means

In addition to the above operations, the research also uses the balanced discriminant function to calculate the optimal number of clusters, thereby improving the clustering quality of the K-means and the quality of learning behavior feature extraction, as shown in formula (10).

$$J_{\min}\left(c,k\right)=\sqrt{w\left(c\right)^2+b\left(c\right)^2} \quad (10)$$

In formula (10), $b\left(c\right)$ is the distance between two different class centers. When the k value of the number of clusters makes formula (10) the minimum, k is the optimal number of clusters. Based on the above content, the optimization of the K-means is completed, and the learning behavior feature extraction algorithm is constructed.

3.3 Construction of learning performance prediction model based on improved RBF

RBFNN is a feedforward neural network proposed at the end of the last century. It has the advantages of high efficiency, less time required for training, strong data processing ability, and strong applicability. Therefore, the study uses RBFNN to classify the extracted learning behavior features and build a learning performance prediction model. The study determines the weight of each student's learning behavior through formula (3), and selects the learning features whose weight value ranks in the top five as the input of the RBFNN model, and the student's academic performance as the output of the model. The input and output relationship of the model is expressed as formula (11).

$$G(x)=\{x_1,x_2,x_3,x_4,x_5\} \tag{11}$$

In formula (11), $x_1 \sim x_5$ refers to the learning features whose weight values are ranked in the top five. The basic principle of RBFNN is to input the feature vector into a hidden layer space constructed by the RBF, so that the input feature vector can be upgraded to a high-dimensional space, so that the original low-dimensional vector problem that cannot be linearly divided can be transformed into a linear partitioning of high-dimensional vector problems. The mathematical model of RBFNN can be expressed by formula (12).

$$y_i=\sum_{k=1}^{M} w_{ij}h(\|x_i-c_k\|)+\theta_j \tag{12}$$

In the formula (12), $h(\|x_i-c_k\|)$ is the RBF, and the Gaussian function is used as the RBF in the research. $\|x_i-c_k\|$ represents the Euclidean distance between x_i to and c_k. θ_j represents the center of the node in the hidden layer space, and $k(k=1,2,\ldots,M)$ is the threshold of the output node. The performance of the RBFNN model is closely related to the location of the network center point, the number of hidden layer nodes and other parameter settings. The general RBFNN parameter settings are compared by manual experiments, and the network parameters are determined according to the results of multiple experiments. However, this method is inefficient, and the set parameters are often not the optimal parameters. Therefore, the study proposes a strategy to optimize the parameters of RBFNN. First of all, the network center point of RBFNN is obtained by the improved K-means to obtain the optimal network center point, so as to obtain the best learning effect and improve the training rate. When a vector feature is input x_i, the output result of the node in the hidden layer space is as shown in formula (13).

$$\varphi_{ij}=\Phi(\|x_i-c_i\|) \tag{13}$$

In formula (13), Φ is the output matrix of the hidden layer, as shown in formula (14).

$$\Phi=\begin{pmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1h} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2h} \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_{n1} & \varphi_{n2} & \cdots & \varphi_{nh} \end{pmatrix} \tag{14}$$

In formula (14), h is the number of hidden layers, and n is the number of nodes in the hidden layer. The output of RBFNN Y can be calculated by formula (15).

$$Y=\Phi w \tag{15}$$

In formula (15), w is the network weight. The optimal center point and the number of optimal hidden layer nodes of RBFNN are respectively determined by the improved K-means and the balanced discriminant function, and the optimal hidden layer node width σ is solved by formula (16).

$$\sigma=\frac{d_{\max}}{\sqrt{2h}} \tag{16}$$

In the formula (16), $d_{\max}$ represents the maximum value of the distance between the center points of the network. To sum up, the basic process of constructing a performance prediction model based on improved RBFNN is in Figure 3.

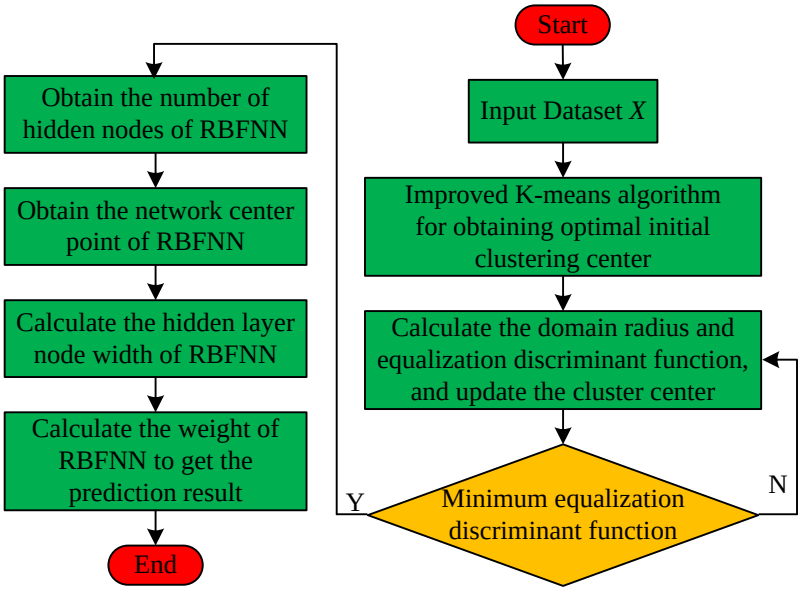

Figure 3: Performance prediction model based on improved RBFNN

4 Performance analysis of the performance prediction model based on improved RBFNN

4.1 Experimental setup and parameter optimization

This study introduced enhancements to the K-means clustering algorithm by integrating the density peak concept from DBSCAN, thereby optimizing cluster center selection through data point density. The balanced discriminant function was employed to ascertain the optimal cluster count, minimizing inter-class distances. For the RBFNN, parameters were set based on the improved K-means outcomes, with the Gaussian function designated as the radial basis, and node widths calibrated to the maximum center point distances.

The experiment utilized a setup with a Win10 64bit OS, 64GB RAM, Intel Core i7 CPU, Python programming, PyCharm IDE, and a MySQL5.6 database. It leveraged a dataset encompassing 20,000 student learning records from 2018 to 2021, partitioned into an 80% training set and a 20% test set. Parameter tuning commenced with adjusting K-means for accurate clustering, followed by cluster count optimization using the balanced discriminant function. For the RBFNN, network configuration and node widths were experimentally determined post-improvement of K-means. These methodical adjustments led to the identification of optimal parameters, corroborating the model's proficiency in student grade prediction with high precision and efficiency.

4.2 Performance evaluation of student performance prediction model based on improved algorithm

To make better use of online learning data, allow teachers to accurately understand students' learning status, and facilitate the improvement of teaching methods, the research based on the improved K-means and improved RBFNN built a grade prediction model (Model 1) to predict students' final exam grades. Experiments were designed to verify the performance of the model. The experimental environment is as follows. The performance of Model 1 was compared with several common grade prediction models, including the grade prediction model (Model 2) based on the back propagation neural network (BPNN) optimized by the genetic algorithm (GA), the model constructed by optimizing the BPNN based on the adaptive differential evolution algorithm (Model 3), and an end-to-end two-layer self-attention network (DEAN)-based grade prediction model (Model 4).

First, the impact of data preprocessing on the training effect of the model was verified. Before and after data preprocessing, the training process of several models is in Figure 4. In the comparison of Figure 4(a) and Figure 4(b), after preprocessing the data, the number of iterations required for the model to achieve the best accuracy was significantly reduced. Taking model 1 as an example, without data preprocessing, it took 78 iterations for model 1 to reach the target accuracy. After data preprocessing, model 1 only needed 58 times to achieve the target accuracy, which was 20 times below before data preprocessing. Furthermore, Model 1 required significantly fewer iterations to reach the target accuracy than the other 3 models. After data preprocessing, the number of iterations required for Models 2, 3 and 4 to reach the target accuracy was 79, 100, and 122 times, which were 21, 42, and 64 times more than Model 1, respectively. Therefore, data preprocessing was necessary, and the training efficiency of Model 1 was exceed that of

the other three models.

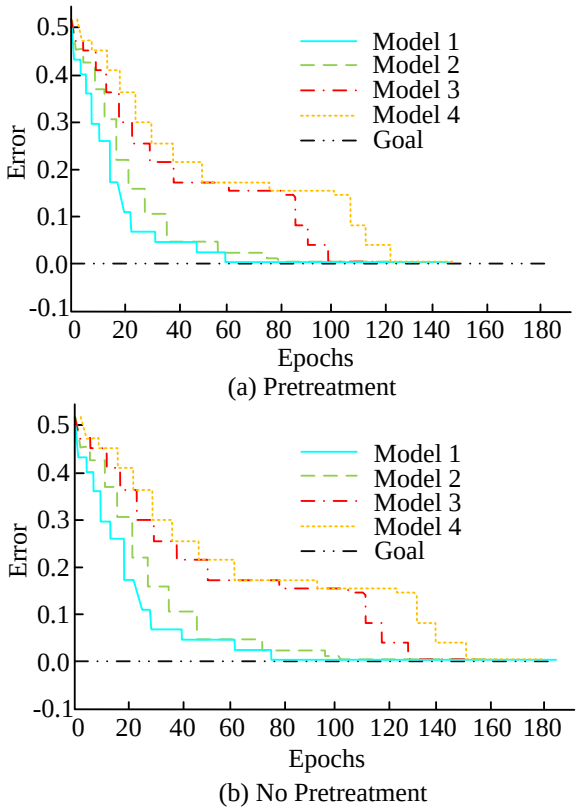

Figure 4: Analysis of data preprocessing effect and model training process

The fitting precision rate of the above four models in the sample data is compared, as shown in Figure 5. The fitting precision rate of Model 1 and sample data in Figure 5 was significantly better than that of Models 2, 3 and 4. Among them, the fitting precision rate of model 1 was 0.957; the fitting precision rate of Model 2 was 0.902, 0.055 lower than Model 1; the fitting precision rate of Model 3 was 0.894, 0.063 lower than Model 1; the fitting precision rate of Model 4 was 0.884, 0.073 lower than Model 1. The above the precision rate of Model 1 was better than the other three models.

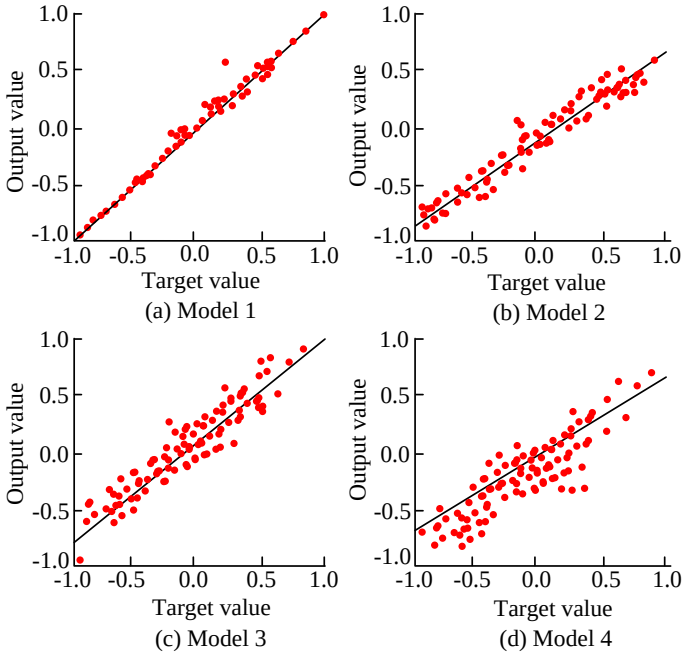

Figure 5: Fitting precision comparison of four models in sample data

On the test dataset, the academic performance prediction precision rate of the four models is in Figure 6. After preprocessing the data, the prediction precision rate of the model for academic performance was significantly improved. Taking Model 1 as an example, before preprocessing, the accuracy rate of Model 1 was 89.8%, and after data preprocessing, Model 1 reached 99.2%, which was 9.4% exceed before preprocessing. In Figure 6(a), after data preprocessing, Model 2, Model 3, and Model 4 were 98.1%, 95.0%, and 92.3%, respectively, which were 1.1%, 4.2%, and 6.9% lower than Model 1, respectively.

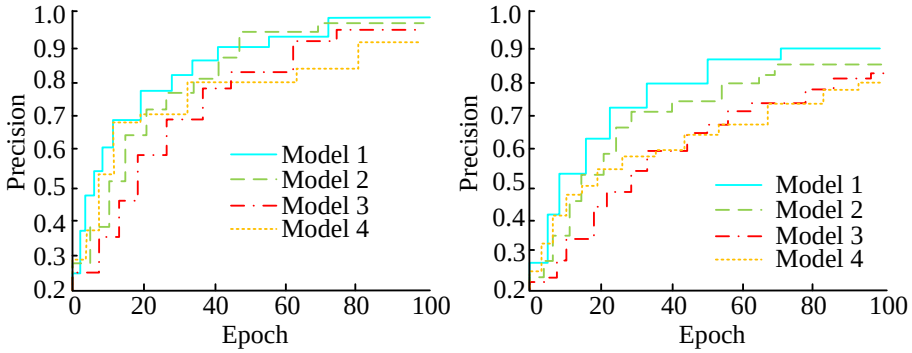

Figure 6: Precision rate of learning performance prediction of four models

60 experiments were carried out on the four models, and the error rate between the model prediction results and the actual performance was compared, as shown in Figure 7. The highest error rate of Model 1 was 0.09, the lowest error rate was 0.01, and the average error rate was 0.03. The error rate of Model 1 was generally lower than that of the other three models. Model 2 was 0.06, 0.03 exceed Model 1, Model 3 was 0.08, 0.05 exceed Model 1. Model 4 was 0.14, 0.11 exceed Model 1.

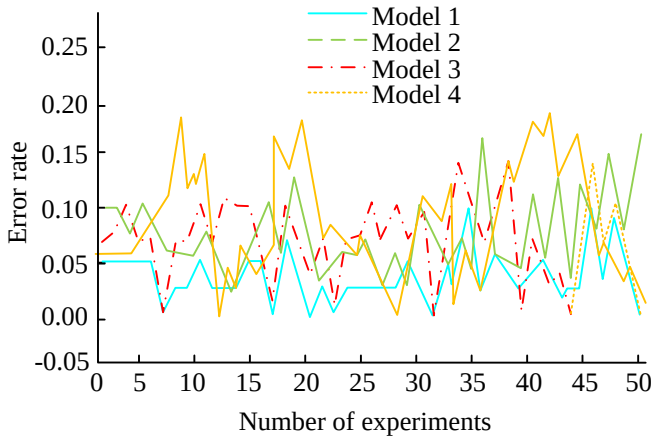

Figure 7: Error rate between model prediction results and actual scores

The performance of the model was comprehensively evaluated using indicators such as recall rate (Recall), precision rate (Precision), and F1, as shown in Table 4. In Table 4, Dataset 1, Dataset 2, Dataset 3, and Dataset 4 represent the original training set, preprocessed training set, original test set, and preprocessed test set, respectively. On each dataset, the indicators of Model 1 exceeded those of the other three models.

Table 4: Comparison of various indicators of the model

Model	Indicator	Dataset			
		Dataset 1	Dataset 2	Dataset 3	Dataset 4
Model 1	Recall	0.892	0.986	0.889	0.991
	Precision	0.883	0.991	0.891	0.996
	F1	0.893	0.979	0.894	0.982
Model 2	Recall	0.854	0.963	0.852	0.984
	Precision	0.843	0.982	0.838	0.975
	F1	0.866	0.952	0.869	0.962
Model 3	Recall	0.817	0.944	0.819	0.951
	Precision	0.802	0.958	0.805	0.942
	F1	0.819	0.950	0.824	0.944
Model 4	Recall	0.784	0.910	0.781	0.906
	Precision	0.768	0.916	0.775	0.913
	F1	0.782	0.924	0.781	0.930

The statistical analysis in Table 5 revealed the performance differences of the four models in predicting student scores. Model 1 led with an accuracy of 99.23%, a confidence interval of 98.76% to 99.70%, and a standard deviation of 0.37%, showing its high predictive stability. After 10-fold cross-validation, the average accuracy of Model 1 dropped slightly to 98.65% with a standard deviation of 0.45%, confirming its excellent generalization ability. Model 2 had an accuracy of 98.15% and a standard deviation of 0.48%, and an average accuracy of 97.25% and a standard deviation of 0.56% after 10-fold cross-validation. Although slightly inferior to Model 1, it still performed well. Model 3 and Model 4 had an accuracy of 95.08% and 92.34%, respectively, with large standard deviations of 0.62% and 0.83%, indicating that the prediction results were highly volatile. This was further confirmed by the cross-validation results, with average accuracies of 94.31% and 91.52%, respectively, and standard deviations of 0.71% and 0.97%, respectively.

Table 5: Model performance statistics and cross-validation results

Model number	Accuracy (%)	Confidence interval of accuracy (95%)	Standard deviation (%)	Cross-validation (10 folds) average accuracy (%)	Cross-validation (10 folds) standard deviation (%)
Model 1	99.23	98.76 - 99.70	0.37	98.65	0.45
Model 2	98.15	97.42 - 98.88	0.48	97.25	0.56
Model 3	95.08	94.21 - 95.95	0.62	94.31	0.71
Model 4	92.34	91.24 - 93.44	0.83	91.52	0.97

5 Discussion

The model proposed in this paper demonstrated significant improvements over existing work in the literature in terms of efficiency and prediction accuracy. Specifically, compared with the student performance prediction model built by Mengash H A et al. based on ANN technology, the proposed model improved both fitting accuracy and accuracy rate, while reducing the number of iterations required for model training. The prediction accuracy of the model proposed by Mengash H A et al. on student data from Saudi public universities exceeded 79%, while the accuracy of the model in this paper reached 99.2%, with an average prediction error rate of only 0.03, showing higher prediction accuracy. Furthermore, compared with the RBF-based supervisory controller proposed by Sun Y et al., the model in this paper realized automation in parameter setting and reduced the dependence on manual experience. Although the research by Sun Y et al. verified the effectiveness and robustness of the RBF method in terms of delay, it did not conduct an in-depth discussion on parameter optimization. By improving the K-means algorithm and the balanced discriminant function, this paper not only optimized the network center point and the number of hidden layer nodes of RBFNN, but also automatically determined the optimal hidden layer node width, thereby improving the model's generalization ability and prediction accuracy. In addition, compared with the RBF manipulator control scheme based on adaptive bias proposed by Liu Q et al., the model in this paper was more targeted and practical in application in the field of education. Although the research by Liu Q et al. verified the controllability of the solution for the robot manipulator in simulation experiments, its application scope was relatively limited. In contrast, this model focused on the field of education and provided teachers with a new tool that helps to understand students' learning status more intuitively and formulate more effective teaching strategies.

In this study, the processing of student data strictly complied with ethical standards to ensure privacy protection and data security. All data were anonymized before analysis, removing all personal identifying information, such as name and ID number, to protect student privacy. Access and storage of data were encrypted and restricted to authorized personnel. The purpose of the study, how the data would be used, and potential benefits were transparently explained to the participating educational institutions and students. The research process was reviewed by the institutional ethics review board to ensure ethical compliance. These measures reflect a serious attitude towards research ethics and a commitment to data protection.

Although the model in this paper achieved positive results on a single university dataset, its generalization ability still needs to be verified on a wider range of datasets. Future research should expand the number and sources of samples and consider the differences in different educational backgrounds and learning environments to evaluate the applicability and robustness of the model in different situations. At the same time, in-depth analysis of the data preprocessing and feature selection process is also necessary, because these steps directly affect the upper limit of model performance. Through these efforts, the contribution of this method in the field of educational DM can be further consolidated and the technological progress in this field can be promoted.

6 Conclusion

Using DM technology to analyze students' online learning to predict students' final grades has become a hot spot in academic research. Aiming at the low efficiency and precision rate of the existing final grade prediction methods, the study improved the K-means to extract learning behavior features, optimized the RBFNN by using the improved K-means and the balanced discriminant function, and finally constructed the improved RBFNN-based grade prediction model. Compared with the original dataset, after data preprocessing, the number of iterations of Model 1 was reduced by 20 times, and the prediction accuracy rate was increased by 9.4%. Model 1 only needed 58 times to reach the target accuracy, which was 21, 42 and 64 times below Model 2, Model 3 and Model 4, respectively. The fitting precision rate of Model 1 was 0.957, 0.055 exceed Model 2, 0.063 exceed Model 3, and 0.073 exceed Model 4. The precision rate of Model 1 reached 99.2%, which was 1.1%, 4.2% and 6.9% exceed Model 2, Model 3 and Model 4, respectively. The average prediction error rate of Model 1 was 0.03, 0.03 lower than Model 2, 0.05 lower than Model 3, and 0.11 lower than Model 4. The Recall, Precision, F1 and other indicators of Model 1 exceeded those of the other three models. The proposed model could efficiently and accurately predict students' final grades, and help teachers understand students' learning conditions, so as to make targeted teaching improvements. The study only used the data of one university in the experimental process, and the limitations of the experimental samples were relatively large, which may have a certain impact on the results. In the future, the research needs to expand samples' number and sources of samples to reduce experimental errors.

Funding Statement

The research is supported by Philosophy and Social Science Research Projects of Shanxi Provincial Colleges and Universities, Research on the Empowerment of Career Growth and Development Mechanism for Higher Vocational Counselors by Big Data--An Evaluation Based on Competency and Job Performance, 2023zsszsx183.

References

[1] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," Applied Sciences, vol. 11, no. 1, pp. 237-237, 2020. https://doi.org/https://doi.org/10.3390/app11010237

[2] D. Shin and J. Shim, "A systematic review on data mining for mathematics and science education," International Journal of Science and Mathematics Education, vol. 19, no. 4, pp. 639-659, 2021. https://doi.org/10.1007/s10763-020-10085-7

[3] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," Smart Learning Environments, vol. 9, no. 11, pp. 1-19, 2022. https://doi.org/10.1186/s40561-022-00192-z

[4] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," IEEE Access, vol. 8, no. 3, pp. 55462-55470, 2020. https://doi.org/10.1109/ACCESS.2020.2981905

[5] J. Yang, Y. Li, Q. Liu, L. Li, A. Feng, T. Y. Wang, S. Zheng, A. D. Xu, and J. Lyu, "Brief introduction of medical database and data mining technology in big data era," Journal of Evidence-Based Medicine, vol, 13, no. 1, pp. 57-69, 2020. https://doi.org/10.1111/jebm.12373

[6] A. Sungheetha and R. S. Rajendran, "Big data analysis and perturbation using data mining algorithm," Journal of Soft Computing Paradigm (JSCP), vol. 3, no. 1, pp. 19-28, 2021. https://doi.org/10.36548/jscp.2021.1.003

[7] T. S. Kumar, "Data mining based marketing decision support system using hybrid machine learning algorithm," Journal of Artificial Intelligence, vol. 2, no. 3, pp. 185-193, 2020. https://doi.org/10.36548/jaicn.2020.3.006

[8] P. Sunhare, R. R. Chowdhary, and M. K. Chattopadhyay, "Internet of things and data mining: An application-oriented survey," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 6, pp. 3569-3590, 2022. https://doi.org/10.1016/j.jksuci.2020.07.002

[9] S. Shakya, "A self monitoring and analyzing system for solar power station using IoT and data mining algorithms," Journal of Soft Computing Paradigm, vol. 3, no. 2, pp. 96-109, 2021. https://doi.org/10.36548/jscp.2021.2.004

[10] P. Edastama, A. S. Bist, and A. Prambudi, "Implementation of data mining on glasses sales using the apriori algorithm," International Journal of Cyber and IT Service Management, vol. 1, no. 2, pp. 159-172, 2021. https://doi.org/10.34306/ijcitsm.v1i2.46

[11] C. Sirichanya and K. Kraisak, "Semantic data mining in the information age: A systematic review," International Journal of Intelligent Systems, vol. 36, no. 8. pp. 3880-3916, 2021. https://doi.org/10.1002/int.22443

[12] Y. Zhou, X. Zhang, and F. Ding, "Hierarchical estimation approach for RBF-AR models with regression weights based on the increasing data length," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 12, pp. 3597-3601, 2021. https://doi.org/10.1109/TCSII.2021.3076112.

[13] Q. Li, Q. Xiong, S. Ji, Y. Yu, C. Wu, and H. L. Yi, "A method for mixed data classification base on RBF-ELM network," Neurocomputing, vol. 431, no. 3, pp. 7-22, 2021. https://doi.org/10.1016/j.neucom.2020.12.032

[14] Z. Yang, M. Mourshed, K. Liu, X. Z. Xu, and S. Z. Feng, "A novel competitive swarm optimized RBF neural network model for short-term solar power generation forecasting," Neurocomputing, vol. 397, no. 7, pp. 415-421, 2020. https://doi.org/10.1016/j.neucom.2019.09.110

[15] Y. C. Chen and D. C. Li, "Selection of key features for PM2. 5 predictions using a wavelet model and RBF-LSTM," Applied Intelligence, vol. 51, no. 4, pp. 2534-2555, 2021. https://doi.org/10.1007/s10489-020-02031-5

[16] Y. Sun, J. Xu, G. Lin, W. Ji, and L. Wang, "RBF neural network-based supervisor control for maglev vehicles on an elastic track with network time delay," IEEE Transactions on Industrial Informatics, vol. 18, no. 1, pp. 509-519, 2020. https://doi.org/10.1109/TII.2020.3032235

[17] I. Izonin, R. Tkachenko, I. Dronyuk, P. Tkachenko, M. Gregus, and M. Rashkevych, "Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method," Mathematical Biosciences and Engineering, vol. 18, no. 3, pp. 2599-2613, 2021. https://doi.org/10.3934/mbe.2021132

[18] Q. Liu, D. Li, S. S. Ge, R. H. Ji, Z. Ouyang, and K. P. Tee, "Adaptive bias RBF neural network control for a robotic manipulator," Neurocomputing, vol. 447, no. 8, pp. 213-223, 2021. https://doi.org/10.1016/j.neucom.2021.03.033

[19] A. H. Fath, F. Madanifar, and M. Abbasi, "Implementation of multilayer perceptron (MLP) and radial basis function (RBF) neural networks to predict solution gas-oil ratio of crude oil systems," Petroleum, vol. 6, no. 1, pp. 80-91, 2020. https://doi.org/10.1016/j.petlm.2018.12.002