Eng Int Syst (2023) 3: 215–220
© 2023 CRL Publishing Ltd

Engineering
Intelligent Systems

Model for predicting students' academic performance in tertiary English courses

Liu Yang

Foreign Language Department, Hebei University of Architecture, Zhangjiakou 075000, Hebei, China

A student's academic performance is an outcome indicating the effect of teaching and learning practices. It is important for students, teachers and school administrators that students' academic performance be predicted accurately. Taking the learning outcome data of students enrolled in English courses at the Hebei University of Architecture in 2020 and 2021, this paper analyzed students' English grades, taking five factors into consideration including gender and number of books borrowed, and established a particle swarm optimization-radial basis function (PSO-RBF) model for predicting students' performance in English courses. In the proposed model, the RBF neural network in data mining was combined with PSO. It was found that the model had the best performance on data set 3, which had the largest volume of data among three distinct data sets. The PSO-RBF model had a root-mean-square error (RMSE) value of 0.8237, a mean absolute error (MAE) value of 0.6255, and a mean absolute percentage error (MAPE) value of 0.2014, which were smaller than decision tree (DT) and k-nearest neighbor (KNN) models. The experimental results verify the reliability of the PSO-RBF model in predicting students' academic performance in a tertiary English course, and its applicability in a real-world university setting.

Keywords: data mining, English performance, prediction model, neural network

1. INTRODUCTION

With the development of information technology, university education is gradually moving toward informatization. whereby large volumes of education-related data are accumulated in the institution's management system (Thakar, 2015). These data contain a great deal of information that is useful for teaching practices (Asif et al., 2017), and can provide valuable guidance for teachers, students, administrators, and other relevant stakeholders. To utilize these data, educational data mining (EDM) has been extensively studied (Dutt et al., 2017). EDM can mine useful information from the teaching platform through data mining (Kamthania, 2016) and various related technologies (Moscoso-Zea and Lujan-Mora, 2016). The data derived from this mining has been applied in practical teaching activities (Ramos et al., 2016). Costa et al. (2017), predicted students who were likely to fail in an introductory programming course, and evaluated four prediction methods applied to data obtained from a Brazilian Public University. They found that the support vector machine had better validity. Raju et al. (2015) examined student data from a university in the southeastern United States and, using decision trees and logistic regression, found that the first-semester grade point average (GPA) and first semester credits had great impacts on the student retention rate and the likelihood that they would graduate. Hussain et al. (2018) collected data from three universities in Assam, India, mined these data using association rules, and predicted students' final grades. They found that the random forest method worked best. Alturki et al. (2021) collected data from 300 undergraduate students at a university in Saudi Arabia to predict academic achievement, compared six data mining methods, and found that naive Bayes performed well in predicting students' learning outcomes. This paper

*Corresponding address: No. 13, Chaoyang West Street, Zhangjiakou 075000, Hebei, China. Email: d769yl@yeah.net

Table 1 Course learning data.

Data number	1	2	3	4	5	……	1250
Gender	Male	Female	Female	Male	Female	……	Female
Number of books borrowed	1	5	27	51	34	……	46
Number of absences	11	4	5	0	2	……	0
Number of assignments not submitted	12	5	3	1	2	……	0
Average quiz score	46	65	70	97	92	……	98
Final grade	55 points	70 points	72 points	96 points	90 points	……	96 points

Table 2 Distribution of final grades under different factors.

Characteristics		Final grade interval				
		[0–60)	[60–70)	[70–80)	[80–90)	[90–100]
Gender	Male	2.25%	**55.67%**	23.64%	9.36%	9.08%
	Female	0.98%	13.67%	26.78%	**41.33%**	17.24%
Number of books borrowed	[0-20)	10.87%	**34.56%**	**49.27%**	3.98%	1.32%
	[20–40)	0.37%	16.33%	**38.64%**	**36.78%**	7.88%
	40 and above	0.12%	4.38%	5.33%	**46.32%**	**43.85%**
Number of absences	[0–5)	0.78%	5.64%	12.66%	30.12%	**50.80%**
	[5–10)	30.12%	**35.64%**	22.33%	9.57%	2.34%
	10 and above	**39.64%**	37.64%	12.02%	9.87%	0.83%
Number of assignments not submitted	[0–5)	0.64%	3.64%	5.64%	**46.78%**	43.30%
	[5–10)	**38.77%**	29.36%	25.26%	5.16%	1.45%
	10 and above	**42.36%**	36.78%	17.36%	3.12%	0.38%
Average quiz score	[0–60)	**67.34%**	25.36%	4.33%	2.71%	0.26%
	[60–70)	16.78%	**71.26%**	8.36%	2.97%	0.63%
	[70–80)	9.87%	25.36%	**55.33%**	9.36%	0.08%
	[80–90)	3.36%	10.33%	26.78%	**55.12%**	4.41%
	[90–100]	0.00%	2.36%	8.77%	12.34%	**76.53%**

developed a particle swarm optimization-radial basis function (PSO-RBF) model for English performance prediction by data mining and verified the reliability of the model through experiments and analysis, which is conducive to better application of data mining techniques in the field of education and also makes some contributions to accurate and efficient student performance prediction.

2. ANALYSIS OF ENGLISH PERFORMANCE UNDER UNIVERSITY EDUCATION

In this study, first we collected the learning data of students enrolled in an English course at the Hebei University of Architecture in 2020 and 2021. The data, obtained from the institution's education management system, served as the basis of our prediction model. After eliminating data that were incomplete and/or abnormal, 1248 data items were obtained. The content of the data set is shown in Table 1.

To facilitate model predictions, the data in Table 1 were transformed as follows:

(1) gender; male $= 1$, female $= 0$;

(2) number of books borrowed: $[0-20) = 1$, $[20-40) = 2$, 40 and above $= 3$;

(3) number of absences: $[0-5) = 1$, $[5-10) = 2$, 10 and above $= 3$;

(4) number of assignments not submitted: $[0-5) = 1$, $[5-10) = 2$, 10 and above $= 3$;

(5) average quiz score: [0 points-60 points) $= 1$, [60 points-70 points) $= 2$, [70 points-80 points) $= 3$, [80 points-90 points) $= 4$, [90 points-100 points] $= 5$.

After data processing, the English grades were further analyzed. The distribution of students' final grades and various (possibly) influencing factors is shown in Table 2.

As shown in Table 2, the final grades of males were mainly [60 points-70 points), while the grade of most females was [80 points-90 points), indicating that females were more likely to achieve excellent grades in the university English course. The scores of the students who borrowed [0–20) books were [60 points-70 points) and [70 points-80 points), the scores of the students who borrowed [20–40) books were [70 points-80 points) and [80 points-90 points), and the scores of the students who borrowed 40 or more nooks were [80 points-90 points) and [90 points-100 points], indicating that the higher the number of books borrowed, the higher was the likelihood of a student obtaining higher final grades. The higher the number of absences and the number of assignments not submitted, the higher is the likelihood that students' grades would be [0 points-60 points). Regarding the average quiz score, the

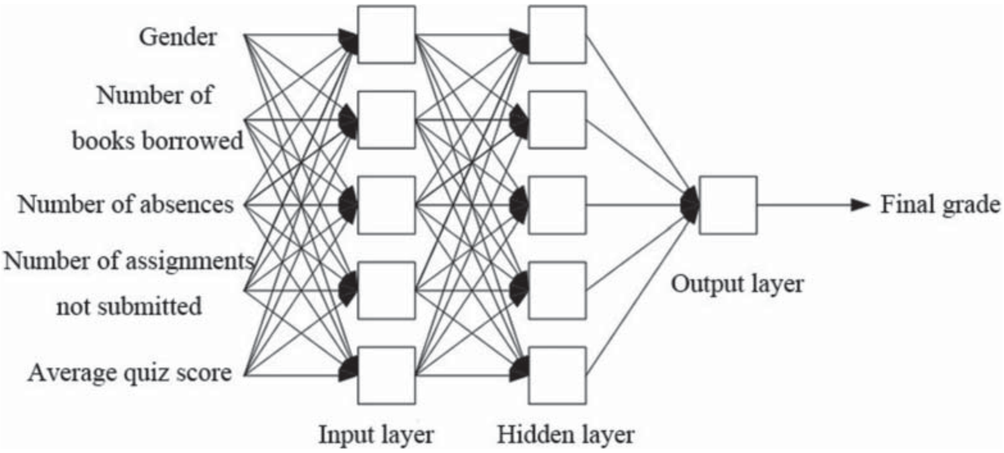

Figure 1 A PSO-RBF model for English performance prediction.

students were most likely to obtain grades in the final exam similar to those they received for regular tests. Students with an average quiz score of [90 points-100 points] were 76.53% more likely to obtain a score of [90 points-100 points] on the final exam.

3. ESTABLISHMENT OF AN ENGLISH PERFORMANCE PREDICTION MODEL

Data mining algorithms commonly used for EDM include classification, clustering and association rule mining. Performance prediction can be considered to be a classification problem. Therefore, the RBF neural network algorithm, a classification algorithm, is selected for this study.

Neural networks perform well in data prediction (Formisano et al., 2021), the most widely used ones being back-propagation neural networks (BPNNs) and RBF neural networks. BPNNs have the disadvantages of slow convergence and poor generalization (Jiang et al., 2017), while RBF neural networks have a simple structure, converge quickly, and are adaptable and efficient (Wang et al., 2016). Therefore, in this paper, an RBF neural network is selected for the building of our prediction model.

An RBF neural network is a three-layer neural network. Let $x_1, x_2, \ldots, x_n$ be input vectors, $y_1, y_2, \ldots, y_m$ be output vectors, and w_{ij} be the weight. The output from the input layer to the hidden layer is the activation function:

$$h_j = exp\left(-\frac{1}{2\sigma_i^2}\|x_n - C_i\|\right),$$

where σ_i is the width parameter, and C_i is the center vector of the Gaussian function.

The output of the output layer is:

$$y_i = \sum_{j=1}^{m} w_{ij}h_j,$$

where w_{ij} is the weight value. In the RBF neural network, the most influential parameters to the model are σ_i, C_i and w_{ij}. The three parameters are optimized using the PSO algorithm.

The PSO algorithm is a simulation of the food-seeking behavior of birds, and is effective in finding globally-optimal solutions (Wang and Liu, 2016). It has few parameters, fast convergence, and easy implementation, and has very wide applications in various industries. Suppose that in a D-dimensional space, the initial position of the particle is $x_i (i = 1, 2, \cdots, N)$, the velocity is v_i, the best position passed by the particle is p_{best_i}, and the best position passed by all particles is g_{best_i}, then in the PSO algorithm, the update formulas for the velocity and position of the particle are:

$$\begin{aligned} v_{id}^{k+1} &= wv_{id}^k + c_1r_1(p_{best_{id}} - x_{id}^k) + c_2r_2(g_{best_{id}} - x_{id}^k), \\ x_{id}^{k+1} &= x_{id}^k + v_{id}^k, \end{aligned}$$

where k is the number of iterations, w is the inertia factor, c_1 and c_2 are the learning factors, and r_1 and r_2 are the random numbers in [0,1]. When determining the value of w, this paper uses a linear decreasing formula to improve the convergence performance of the PSO algorithm:

$$w = w_{max} - \frac{k}{k_{max}}(w_{max} - w_{min}),$$

where w_{max} and w_{min} are the maximum and minimum weights, $w_{max} = 0.9$ and $w_{min} = 0.4$ usually.

The optimal σ_i, C_i and w_{ij} values are obtained using the above PSO algorithm to build a PSO-RBF model for the prediction of student performance in English courses.

The five features described in Section 2 provide the input for the model, and the students' final grades are the output, as shown in Figure 1.

4. PREDICTION RESULTS AND ANALYSIS

The experiments were conducted in a MATLAB environment. The operating system was Windows 10 × 64. The processor was an Intel i7. The memory was 16 GB. The programming language was Python. There were 1248 experimental data, as shown in Table 3.

The 1248 data were divided according to three forms, as shown in Table 4.

Table 3 Experimental data.

Experimental data number	1	2	3	4	5	……	1248
Gender X1	0	1	0	1	1	……	0
Number of books borrowed X2	3	1	2	2	1	……	2
Number of absences X3	1	3	2	2	3	……	1
Number of assignments not submitted X4	1	2	2	1	3	……	1
Average quiz score X5	5	2	4	5	1	……	4
Final grade Y	97	64	83	98	46	……	88

Table 4 Division of three data sets.

	Training data	Test data
Data set 1	312	312
Data set 2	642	312
Data set 3	936	312

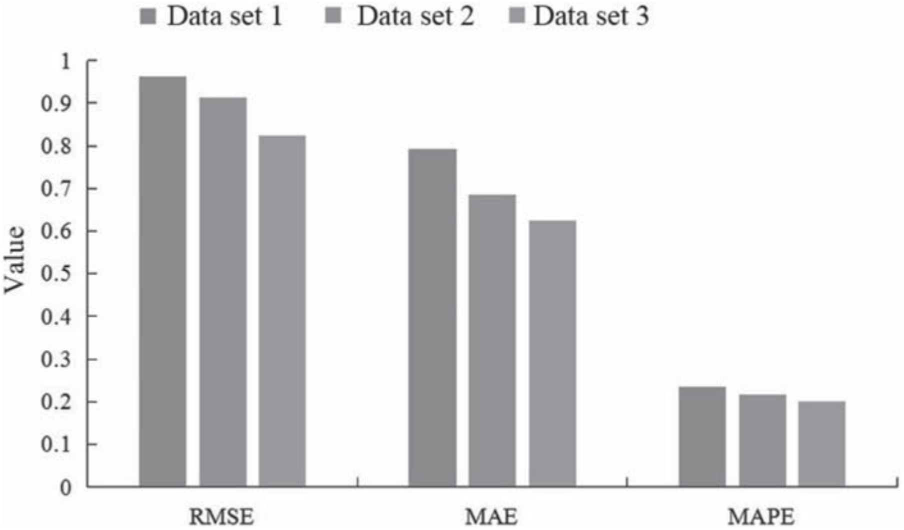

Figure 2 Performance of the PSO-RBF model on different data sets.

The total number of samples is represented by n, y'_i is the predicted value, and y_i is the actual value. The performance of the model in predicting English performance was evaluated using the following three indicators.

(1) RMSE, used for measuring the deviation between predicted and actual values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y'_i - y_i)^2}.$$

(2) MAE, used for reflecting the actual situation of the error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y'_i - y_i|.$$

(3) MAPE, used for measuring the strengths and weaknesses of the model. If MAPE is 0%, then the model is perfect.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y'_i - y_i|}{y_i}$$

The PSO-RBF model was trained and tested using the three data sets. The results are presented in Figure 2.

Figure 2 indicates that there were some differences in the prediction performance of the model on different datasets. On data set 1, the model had an RMSE value of 0.9621, an MAE value of 0.7936, and an MAPE value of 0.2356; on data set 2, the model had an RMSE value of 0.9123, MAE value of 0.6848, and MAPE value of 0.2171; on data set 3, the model had an RMSE value of 0.8237, an MAE value of 0.6255, and an MAPE value of 0.2014. Overall, the model had the smallest RMSE, MAE and MAPE values on data set 3, indicating good performance. This suggests that the larger the volume of data, the better is the performance of the model. Therefore, in the subsequent experiment, data set 3 was selected for the experiment.

To further demonstrate the performance of the PSO-RBF model, it was compared with the following models:

(1) decision tree (DT) model (Dadsena, 2021);

(2) K-nearest neighbor (KNN) model (El-Magd et al., 2021);

(3) BPNN (Zhu et al., 2021);

(4) RBF neural network (Antos et al., 2022).

Table 5 Comparison of experimental results.

	RMSE	MAE	MAPE
DT	0.9234	0.9346	0.2481
KNN	0.8921	0.9254	0.2354
BPNN	0.8747	0.8245	0.2267
RBF	0.8536	0.7548	0.2112
PSO-RBF	0.8237	0.6255	0.2014

For comparison, the results of the experiments are shown in Table 5.

Table 5 shows that of the five models, the DT, KNN and BPNN models performed poorly. For the DT model, the RMSE value was 0.9234, the MAE value was 0.9346, and the MAPE value was 0.2481, indicating that its performance in predicting English performance was moderate. The performance of the RBF model was better than the above three models, achieving RMSE, MAE and MAPE values of 0.8536, 0.7548 and 0.2112, respectively, which were 7.56%, 19.24% and 14.87% less than those of the DT model. The comparison verified that the RBF model was reliable in predicting data. Compared with the RBF model, the RMSE, MAE and MAPE of the PSO-RBF model were 3.5%, 17.13% and 4.64% less, indicating that the model performed significantly better and obtained more accurate results after being combined with the PSO algorithm.

5. DISCUSSION

The current application directions of EDM include student performance prediction (Salih and Khalaf, 2021), student psychological cognition (Shi, 2019), evaluation of teaching practices (Wang, 2019), student recommendation systems and visualization (Kumar, 2016). Through EDM, students can acquire a better understanding of their current academic performance, which may encourage them to improve their learning efficiency. By means of EDM, teachers can receive timely feedback about the effectiveness of their teaching practices and acquire a better understanding of their students' needs and capabilities, which will inform their subsequent teaching approach (Chamizo-Gonzalez et al., 2015). EDM enables educational institutions to understand and monitor their education system, thereby facilitating staff and student management. In this paper, we focused on the prediction of students' academic performance through EDM. We took students' performance in a university English course as the data sample and designed a model for experimental analyses.

The experimental results showed that the performance of the PSO-RBF model differed, depending on the size of the data sets. In general, on data set 3, which had the largest amount of data, the PSO-RBF model had the smallest RMSE, MAE, and MAPE values, indicating that the PSO-RBF model had the best prediction performance on data set 3. Therefore, data set 3 was also used in the subsequent experiments. It was seen from the comparison between the RBF model and the other models that the RMSE, MAE and MAPE of DT, KNN and BPNN models were above 0.85, 0.8 and 0.22, respectively, indicating that the performance of all these data mining models was poorer than that of the RBF model. It was seen from the comparison between the RBF model and the PSO-RBF model that the RMSE, MAE and MAPE values of the PSO-RBF model were 0.8237, 0.6255 and 0.2014, respectively, which were smaller than those of the RBF model. The results verified the reliability of the PSO-RBF model.

In this paper, we used data mining methods to predict students' academic performance in a university English course. Although our proposed model achieved some promising results, it has several shortcomings. In future research, other data mining methods should be studied and applied in order to optimize the performance of the PSO-RBF model. Also, experiments could be conducted using larger data sets, and models could be applied and tested in real-world applications for the purpose of predicting students' learning outcomes.

6. CONCLUSION

In this paper, a PSO-RBF model was designed to predict students' performance in English courses. The model was based on the RBF neural network, and the performance of the model was tested on a sample of students enrolled in 2020 and 2021 English courses at a university. The results showed that the RMSE, MAE and MAPE values of the PSO-RBF model were 0.8237, 0.6255 and 0.2014, respectively, which were superior to those of the other models. The experimental results verify the reliability of the PSO-RBF model. The PSO-RBF model can be further refined and applied in actual university education management.

REFERENCES

1. Alturki, S., Alturki, N. & Stuckenschmidt, H. (2021). Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions. *Journal of Information Technology Education Innovations in Practice*, *20*, 121–137.
2. Antos, J., Kubalcik, M. & Kuritka, I. (2022). Scalable Non-dimensional Model Predictive Control of Liquid Level in Generally Shaped Tanks Using RBF Neural Network. *International Journal of Control, Automation and Systems*, *20*(3), 1041–1050.
3. Asif, R., Merceron, A., Ali, S.A. & Haider, N.G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*(oct.), 177–194.
4. Chamizo-Gonzalez, J., Cano-Montero, E.I., Urquia-Grande, E. & Munoz-Colomina, C.I. (2015). Educational data mining for improving learning outcomes in teaching accounting within

higher education. *International Journal of Information & Learning Technology*, *32*(5), 272–285.

5. Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F. & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*(Aug.), 247–256.
6. Dadsena, D.R. (2021). Rotational Moment Shape Feature Extraction and Decision Tree Based Discrimination of Mild Cognitive Impairment Conditions Using MR Image Processing. *Biomedical Sciences Instrumentation*, *57*(2), 228–233.
7. Dutt, A., Ismail, M.A. & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, 15991–16005.
8. El-Magd, S.A., Ali, S.A. & Pham, Q.B. (2021). Spatial modeling and susceptibility zonation of landslides using random forest, naive bayes and K-nearest neighbor in a complicated terrain. *Earth Science Informatics*, 1–17.
9. Formisano, A., D'Addona, D.M., Durante, M. & Langella, A. (2021). Evaluation and neural network prediction of the wear behaviour of SiC microparticle-filled epoxy resins. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *43*(5), 1–9.
10. Hussain, S., Dahan, N.A., Ba-Alwib, F.M. & Najoua, R. (2018). Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447–459.
11. Jiang, G., Luo, M., Bai, K. & Chen, S. (2017). A Precise Positioning Method for a Puncture Robot Based on a PSO-Optimized BP Neural Network Algorithm. *Applied Sciences*, *7*(10), 969.
12. Kamthania, D. (2016). Educational Data Mining – A Case Study. *International Journal of Applied Decision Sciences*, *8*(2), 187.
13. Kumar, S.A. (2016). Edifice an Educational Framework using Educational Data Mining and Visual Analytics. *International Journal of Education & Management Engineering*, *6*(2), 24–30.
14. Moscoso-Zea, O. & Lujan-Mora, S. (2016). Educational data mining: An holistic view. *Iberian Conference on Information Systems & Technologies*, 1–6.
15. Raju, D. & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. *Journal of College Student Retention Research Theory & Practice*, *16*(4), 563–591.
16. Ramos, J.L.C., e Silva, R.E.D., Silva, J.C.S., Rodrigues, R.L. & Gomes, A.S. (2016). A Comparative Study between Clustering Methods in Educational Data Mining. *IEEE Latin America Transactions*, *14*(8), 3755.
17. Salih, N.Z. & Khalaf, W. (2021). Prediction of student's performance through educational data mining techniques. *Indonesian Journal of Electrical Engineering and Computer Science*, *22*(3), 1708.
18. Shi, X. (2019). Emotional Data Mining and Machine Learning in College Students Psychological Cognitive Education. *Engineering Intelligent Systems*, *27*(4), 167–175.
19. Thakar, P. (2015). Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue. *Computer Science*, *2*(2), 400–412.
20. Wang, C.F. & Liu, K. (2016). A Novel Particle Swarm Optimization Algorithm for Global Optimization. *Computational Intelligence and Neuroscience*, *2016*(2–3), 1–9.
21. Wang, L. (2019). Evaluation of Web-Based Teaching Based on Machine Learning and Text Emotion. *Engineering Intelligent Systems*, *27*(3), 111–119.
22. Wang, W.Y., Guo, G.L., Jiang, B. & Wang, L. (2016). Discovering WDMS with Automatic Classification System Based on RBF Neural Network. *Guang Pu Xue Yu Guang Pu Fen Xi*, *36*(10), 3360–3363.
23. Zhu, W., Wang, H. & Zhang, X. (2021). Synergy evaluation model of container multimodal transport based on BP neural network. *Neural Computing and Applications*, *33*(2), 1–9.