Academic Early Warning Model for Students Based on Big Data Analysis

https://doi.org/10.3991/ijet.v18i12.41087

Kun Wang(✉)
Art Design Department, Zibo Vocational Institute, Zibo, China
wangkun@zbvc.edu.cn

Abstract—How to identify in advance and help college students with academic difficulties is an important topic for current education departments and universities. Academic early warning system based on big data analysis comprehensively analyzes the learning, life and psychological data of college students, effectively identifies potential academic problems, and helps teachers and student managers take measures in advance to improve the education quality. The existing academic warning models of college students based on big data analysis often have defects, such as data quality issues, lack of key variables, nonlinear problems, and human factors. Therefore, this paper aimed to study the academic early warning model of college students based on big data analysis. After elaborating on the key points of collecting the academic early warning model data based on big data analysis, this paper explained the reasons of calculating the Pearson correlation coefficient of collected big data. This paper constructed an academic early warning model of college students based on deep self-coding network, provided the construction process, and explained its working principle. After optimizing the model parameters, this paper analyzed the model reconstruction error based on sliding window statistical method, and further improved the prediction ability and generalization performance of evaluating the deep self-coding network model, thus obtaining higher academic early warning accuracy. The experimental results verified that the constructed model was effective.

Keywords—big data analysis, college students, academic early warning, correlation analysis

1 Introduction

With the continuous enlargement of enrollment size in colleges and universities, the number of college students has shown explosive growth, which poses enormous challenges for the allocation and management of educational resources. How to identify in advance and help college students with academic difficulties has become an important topic for current education departments and universities [1, 2]. Academic early warning system based on big data analysis emerged in this context. Traditional academic early warning methods are often based on subjective judgment and experience of teachers, and sometimes overlook important student information [3–5]. However, the academic early warning system based on big data analysis comprehensively analyzes the learning, life and psychological data of college students, effectively identifies potential academic problems, and helps teachers and student managers take measures in advance, thus improving the education quality [6–10]. The research on academic early warning based on big data analysis helps deeply understand the cause and rule of academic problems, and provides scientific basis for education policy formulation and university education management [11, 12]. In addition, algorithm optimization, model construction, and other aspects involved in the research process also promote the technological development of data mining, machine learning, and other fields.

Academic early warning and help are key measures for universities to improve the talent cultivation quality in the new era. Big data generated in educational informatization practice shows its value with the support of artificial intelligence technology. In the context of digital transformation of local colleges and universities, Xu [13] examined the status quo and characteristic trends of academic problems using scientific data analysis methods, and determined key tasks. With the support of algorithms recommended by artificial intelligence, the study introduced three application scenarios with comprehensive data, integrated teaching and learning behavior data and teaching goal achievement with the second classroom, thus helping precise academic assistance, self-improvement of students, and scientific prediction and decision-making. Dong et al. [14] applied machine learning technology to the academic early warning research, and constructed an academic early warning model, thus helping university education managers comprehensively understand students, accurately predict them, and provide personalized services for them. Based on the academic performance data and library data of college students in the first three years, the study predicted whether students would graduate smoothly in the future using Support Vector Machine (SVM) algorithm based on improved Fruit Fly Optimization Algorithm (FOA), and sounded academic early warnings to students who may not be able to graduate smoothly. Experiments showed that the SVM early warning model based on improved FOA was superior to the traditional SVM model, decision tree and random forest in accuracy. Chen [15] proposed an improved fuzzy algorithm, based on fuzzy performance evaluation of composite elements, and applied it to a performance evaluation system to solve complex problems in performance evaluation. In the process of building smart campus in universities, academic early warning, as a main component of smart campus, mainly aimed to ensure that students successfully completed their studies using data mining technology, and provided certain decision-making support for universities.

The existing academic warning models of college students based on big data analysis often have defects, such as data quality issues, lack of key variables, nonlinear problems, and human factors. If the data is missing, incorrect, or incomplete, it affects the prediction results of the model. If key variables are missing, such as academic performance of students, attendance rate, homework completion, and learning behavior, the early warning model cannot accurately predict the performance of students. The prediction process of academic performance is very complex and may involve the interaction of multiple nonlinear factors, which may make the prediction model not accurately capture the impact of all these factors. In addition, academic performance may be affected by several factors, such as psychological health status and social life, but it is difficult for the prediction model to capture them. Therefore, this paper studied the academic early warning model of college students based on big data analysis.

2 Model data collection and correlation analysis

As for the academic early warning model of college students based on big data analysis, its implementation process is an iterative process, which requires continuous data collection and cleaning, feature engineering, model training, model testing and optimization, prediction and interpretation, feedback and improvement. The various academic and behavioral data collected include academic performance, course information, attendance, student behavior, personal information, student evaluation, and other related data. Specifically, first, academic performance data includes academic performance data and grade point average (GPA) of each course for students. Second, course information data includes course difficulty, learning quality, and teacher evaluation and so on. Third, attendance data includes attendance situations and leave records of students, and other data. Fourth, student behavior data includes social network behavior, book borrowing records, extracurricular activity participation of students and other data. Fifth, personal information data includes basic information on students, such as gender, age, major, class, etc. Sixth, student evaluation data includes self-evaluation of students, evaluation from teachers and classmates, and other data. Seventh, other relevant data include family background of students, rewards, punishments, and other data. These data usually come from multiple sources, such as university management system, student questionnaire survey, mobile applications and so on. The data from different sources needs to be integrated, cleaned, denoised, and go through the treatment processes of feature extraction and selection, model training and evaluation, thus finally obtaining a reliable academic early warning model of college students.

Big data used for academic early warning of college students usually has several characteristics, such as high dimensionality, large data volume, high complexity, and diverse data sources. As a unsupervised learning method based on neural network, deep self-coding network effectively extracts features and reduces dimensions, and has good data representation ability. Therefore, the deep self-coding network is suitable for training a large amount of academic data in order to learn effective data representation, which provides useful features for subsequent academic early warning model. At the same time, the network adaptively learns data features, which avoids the artificial feature selection problem in traditional feature engineering. In addition, the network constructs deep structures by stacking multiple self-encoders, which further improves feature representation ability and prediction accuracy. Therefore, this paper constructed an academic early warning model framework of college students with deep self-coding network as the core, which provided useful reference and support for student work. Figure 1 shows the basic structure of deep self-coding network.

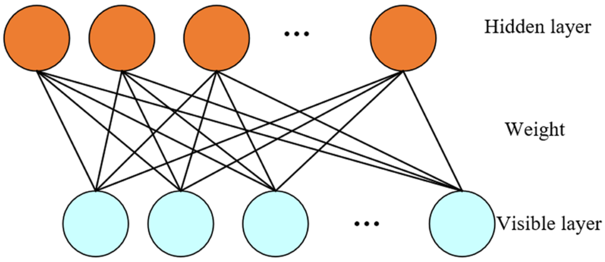


Fig. 1. Structure diagram of deep self-coding network

Pearson correlation coefficient is statistics, which is used to measure the linear relationship strength and direction between two continuous variables. Before conducting feature engineering, the Pearson correlation coefficient calculation of the collected academic early warning big data of college students helps identify the data correlation and then select appropriate features for modeling. Specifically, the reasons and necessity for doing so mainly focus on four aspects, namely, identifying key features, reducing data dimensions, determining data cleaning strategies, and avoiding multicollinearity, thus ultimately improving the prediction accuracy and interpretation ability of the model. Let *f* be the correlation coefficient of the collected data variables, A_j and B_j be the random variables, $\overline{A}$ and $\overline{B}$ be the mean values of random variables, then the calculation formula of *f* was as follows:

$$f = \frac{\sum_{j=1}^{m}(A_j - \overline{A})(B_j - \overline{B})}{\sqrt{\sum_{j=1}^{m}(A_j - \overline{A})^2}\sqrt{\sum_{j=1}^{i}(B_j - \overline{B})^2}} \tag{1}$$

According to the formula, when $f = 1$, *A* and *B* have positive complete correlation and their collected data points are on a straight line, with *B* increasing with the increase of *A*. When $f = -1$, *A* and *B* have negative complete correlation, and their collected data points are also on a straight line, with *B* decreasing with the increase of *A*. When $f = 0$, there is no linear relationship between *A* and *B*.

3 Construction of academic early warning model of college students

The implementation process of academic early warning model based on deep self-coding network includes encoding and decoding processes. After being input into the model, the collected raw data *A* passes through two layers of restricted Boltzmann machines and achieves data feature transmission layer by layer. *B*, which is the deep data feature of the collected academic early warning data, is ultimately output through the high level. This process is called encoding. Decoding is the process of obtaining $\widehat{A}$, where *B* transmits backward layer by layer according to the encoding method while passing through the restricted Boltzmann machines. Figure 2 shows the structure diagram of the early warning model. Let *z* be the connection weight value between the visible layer and the hidden layer, *x* be the offset value of the visible layer, and *y* be the offset value of the hidden layer, then the expressions of two processes were given as follows:

$$B = E(z_B A + x_B) \tag{2}$$

$$\widehat{A} = C(z_{\hat{A}} B + y_{\hat{A}}) \tag{3}$$

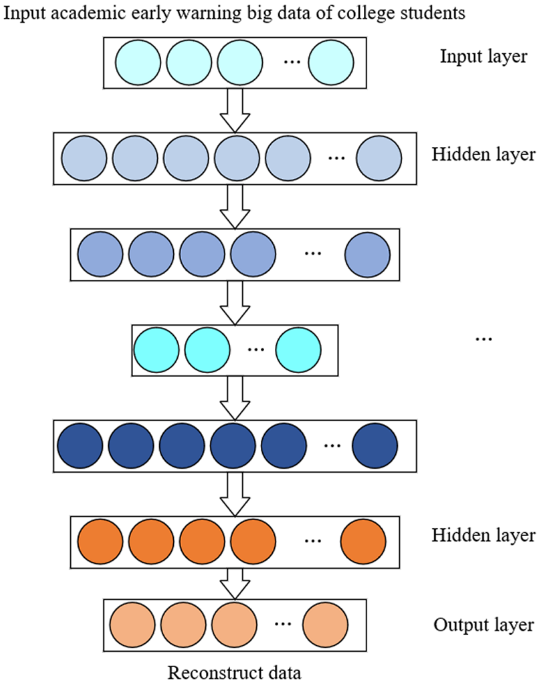


Fig. 2. Structure diagram of early warning model

The model initialized the weight and offset values between constrained Boltzmann machines layer by layer using unsupervised feature optimization algorithm, thus completing its training.

It is usually difficult to construct the likelihood function of big data used for academic early warning of college students, because the big data distribution is irregular. In the academic early warning of college students, this paper considered the visible layer and the hidden layer as different types of data features, and then trained the restricted Boltzmann machine to learn the correlation and probability distribution between different features. In this way, this paper calculated the probability distribution between different features by calculating the energy function, thus sounding academic early warning of college students. Therefore, for big data used for academic early warning of college students, this paper introduced a constrained Boltzmann machine energy model to describe the energy possessed by interlayer unit nodes. Let ϕ be the parameter of constrained Boltzmann machine, z_{nm} be the weight between visible and hidden units, x_n be the offset value of visible unit, y_m be the offset value of the hidden unit, l_n be the state of the visible unit, c_m be the state of the hidden unit, I be the number of visible units, and J be the number of hidden units. The function expression was given as follows:

$$H_{\varphi}(l,e) = -\sum_{n=1}^{I} x_n l_n - \sum_{m=1}^{J} y_m c_m - \sum_{n=1}^{I}\sum_{m=1}^{J} l_n z_{nm} c_m \tag{4}$$

$$\varphi = \{z, x, y\} \tag{5}$$

This paper obtained the joint probability distribution of (*l*, *c*) based on the energy model function of the constrained Boltzmann machine. Let $W(\phi)$ be the normalization factor, and $W(\phi) = \sum_{l,c} g^{-H\phi(l,c)}$, then there was:

$$T_{\varphi}(l,c) = \frac{g^{-H_{\varphi}(l,c)}}{W(\varphi)} \tag{6}$$

This paper integrated the hidden and visible layers, and then obtained their conditional probability $T_{\phi}(c)$ and $T_{\phi}(l)$ distribution. The calculation formulas were given as follows:

$$T_{\varphi}(c) = \frac{\sum_{l} g^{-H_{\varphi}(l,c)}}{\sum_{l,c} g^{-H_{\varphi}(l,c)}} \tag{7}$$

$$T_{\varphi}(l) = \frac{\sum_{c} g^{-H_{g}(l,c)}}{\sum_{l,c} g^{-H_{g}(l,c)}} \tag{8}$$

In order to achieve the maximum possible approximation between the spatial distribution of the input samples and the sample distribution represented by the constrained Boltzmann machine, this paper introduced relative entropy *K-L* distance to represent their difference degree. Let Ψ be the sample space, $D(l)$ be the input sample distribution, and $T(l)$ be the edge distribution of the Gibbs distribution represented by the restricted Boltzmann machine, then there was *K-L* distance:

$$KL(D\|T) = \sum_{l\in\psi} D(l)\ln\frac{D(l)}{T(l)} = \sum_{l\in\psi} D(l)\ln D(l) - \sum_{l\in\psi} D(l)\ln T(l) \tag{9}$$

When the input academic early warning big data sample was determined, the second term in the formula could not only be sampled using the Monte Carlo method, and then was estimated using $\sum 1nT(a)/q$. Let q be the number of samples and $T(l, c)$ be the joint probability distribution of the hidden and visible layers in the second term on the right of the formula in order to ensure the similar distribution of D and T, then there was:

$$\frac{\beta \ln T(l)}{\beta\varphi} = -\sum_{c} T(c\,|\,l)\frac{\beta \ln H(l,c)}{\beta\varphi} + \sum_{l,c} T(l,c)\frac{\beta \ln H(l,c)}{\beta\varphi} \tag{10}$$

This paper solved the first term on the right side of the above formula relatively easily. $T(l, c)$ needed to be solved using the following formula:

$$\sum_{l,c} T(l,c)\frac{\beta H(l,c)}{\beta\varphi} = \sum_{l}\sum_{c} T(l)T(c\,|\,l)\frac{\beta H(l,c)}{\beta\varphi} = \sum_{l}\left(T(l)\sum_{c} T(c\,|\,l)\frac{\beta H(l,c)}{\beta\varphi}\right) \tag{11}$$

Therefore, it was transformed into solving $\sum_{c} T(c|l)\partial H(l,c)/\partial\phi$. Partial derivative of $\{z, x, y\}$ was solved respectively and substituted into the following formulas to calculate the partial derivative of $T(l)$:

$$\frac{\partial T(l)}{\partial z_{n,m}} = T(c_n = 1 \mid l) l_m - \sum_l T(l) T(c_m = 1 \mid l) l_m \tag{12}$$

$$\frac{\partial \ln T(l)}{\partial x_n} = l_n - \sum_l T(l) l_n \tag{13}$$

$$\frac{\partial T(l)}{\partial y} = T(c_n = 1 \mid l) - \sum_l T(l) T(c_m = 1 \mid l) \tag{14}$$

Let $k(a)$ be the *sigmoid* activation function, then the calculation formulas of conditional probability density for the visible and hidden layers of the constrained Boltzmann machine were as follows:

$$T(c_m = 1 \mid l) = k(x_m + \sum_l z_{nm} l_n) \tag{15}$$

$$T(l_n = 1 \mid c) = k(y_n + \sum_l z_{nm} c_m) \tag{16}$$

$$k(a) = \frac{1}{1 + e^{-a}} \tag{17}$$

The training result of constrained Boltzmann machine was solving the appropriate parameter ϕ, which was obtained from the likelihood function formula of the visible layer l:

$$Q(\varphi; l) = \prod_l T_\varphi \frac{\sum_c g^{-H_\varphi(l,c)}}{\sum_{l,c} g^{-H_\theta \varphi(l,h)}} \tag{18}$$

Let γ be the learning rate, $<\cdot>_p$ be the expected value of the training data, $<\cdot>_c$ be the expected data value after c-step Gibbs sampling, Δz_{nm}, Δy_n and Δx_m be the updated weights and offset values. The update rules of ϕ were obtained based on the contrast divergence algorithm:

$$\Delta z_{nm} = \gamma \left(\left\langle l_n c_n \right\rangle_p - \left\langle l_n c_m \right\rangle_c \right) \tag{19}$$

$$\Delta y_n = \gamma \left(\left\langle l_n \right\rangle_p - \left\langle l_n \right\rangle_c \right) \tag{20}$$

$$\Delta x_n = \gamma \left(\left\langle c_n \right\rangle_p - \left\langle c_m \right\rangle_c \right) \tag{21}$$

The first restricted Boltzmann machine extracted features of the input big data for the first time, and obtained the hidden layer parameters by adopting the above learning method. Then the output was used as the input of the second restricted Boltzmann machine for further data feature extraction. By analogy, the feature output of the final constrained Boltzmann machine was the internal features of the academic early warning big data of college students obtained through initial training, thus ultimately completing the pre-training of the deep self-coding network.

4 Model parameter tuning and early warning realization

A deep self-coding network model typically consists of multiple self-encoders, each of which can be trained and optimized using the back propagation (BP) algorithm. When carrying out the reverse fine tuning, this paper first stacked the pre-trained self-encoders to build a depth structure. Then this paper fine tuned the parameters of the entire network using the BP algorithm in order to improve the training efficiency and performance of the deep self-coding network, and reduce the overfitting risk of the model, thus further optimizing the prediction results of the academic early warning model of college students.

In order to effectively improve the learning rate of deep learning network, this paper used cross entropy function as the cost function of BP neural network. Let *i* be the total number of training samples, b_i be the actual data, and $\hat{b}_i$ be the reconstructed data, then the function expression was given as follows:

$$M = -\frac{1}{i}\left[b_i \log \hat{b}_i + (1 - b_i)\log(1 - \hat{b}_i) \right] \tag{22}$$

After the BP algorithm was used, the deep self-coding network model achieved the optimal network parameters, under the condition that the global error met the accuracy requirements. Since this process was completed based on unsupervised learning training, the obtained network parameters needed to be adjusted and optimized only, which had better algorithm timeliness.

Based on sliding window statistical method, this paper analyzed the model reconstruction error in order to further improve the predictive ability and generalization performance of the deep self-coding network model, thus obtaining higher accuracy of academic early warning. Figure 3 shows the schematic diagram of sliding window. Specifically, the raw data was first divided into sliding windows in chronological order, with each window containing data of a time period. This paper reconstructed the data within each window using the deep self-coding network model, and calculated the reconstruction error, which was the error between raw data and the reconstructed data. Then this paper obtained a set of reconstruction error sequences through statistics of the reconstruction errors within each window. After analyzing the reconstruction error sequences, this paper found abnormal windows, whose reconstruction errors significantly deviated from the normal range. For abnormal windows discovered, early warning was sounded in order to notify relevant personnel to intervene, thus improving the early warning effects of academic performance. Let $\hat{A}$ be the reconstructed data of the academic early warning monitoring variables of college students, and *A* be the raw data, then the reconstruction error was calculated by the following formula:

$$Fh = (\hat{A} - A)^2 \tag{23}$$

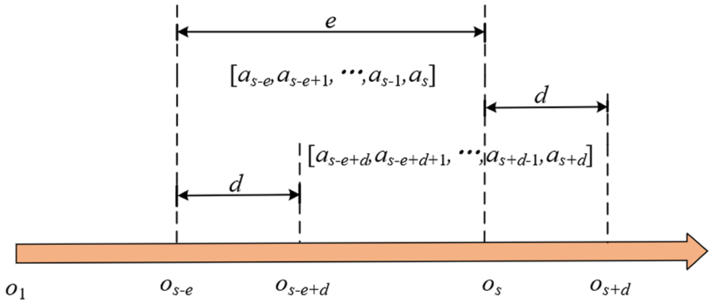

e
$[a_{s-e}, a_{s-e+1}, \cdots, a_{s-1}, a_s]$
d
$[a_{s-e+d}, a_{s-e+d+1}, \cdots, a_{s+d-1}, a_{s+d}]$
d
o_1
o_{s-e}
o_{s-e+d}
o_s
o_{s+d}

Fig. 3. Schematic diagram of sliding window

Let *e* be the window length of the sliding window model used, *d* be the increment, and *i* be the time steps, then the data in the window at *s* time was obtained by the following formula:

$$P_S = \left(a_{s-e}, a_{s-e+1}, a_s\right)^D = \begin{bmatrix} a_{s-e,1} & a_{s-e,1} & \cdots & a_{s-w,m} & \cdots \\ a_{s-e+1,1} & a_{s-e+1,2} & \cdots & a_{s-e+1,m} & \cdots \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ a_{s1} & a_{s2} & \cdots & a_{sm} & \cdots \end{bmatrix} \tag{24}$$

The value of *s* was related to *e* and the increment *d*, satisfying $s = e + id$. When data processing was completed, both ends of the window moved with the same increment *d* along the increasing direction of the time axis at o_s time, and the window data matrix to be processed changed into P_{S+d}. Figure 4 shows the schematic diagram of early warning process of the model.

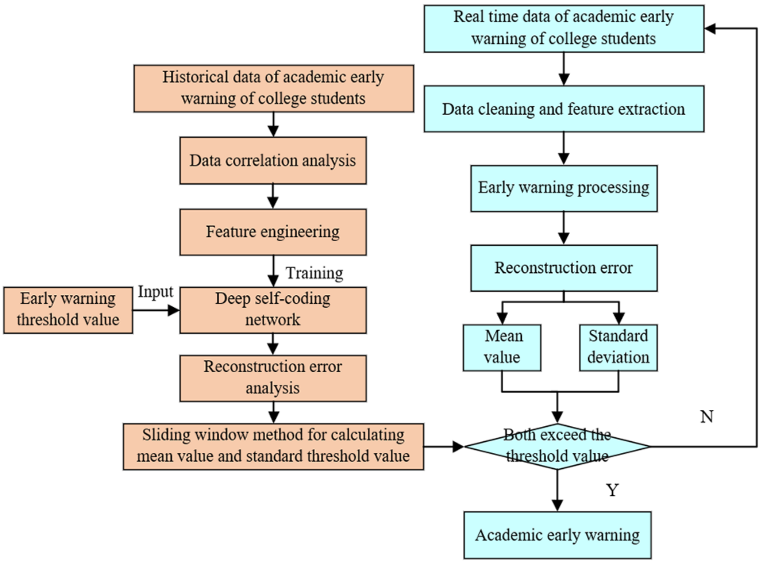

Real time data of academic early warning of college students
Historical data of academic early warning of college students
Data cleaning and feature extraction
Data correlation analysis
Early warning processing
Feature engineering
Reconstruction error
Training
Early warning threshold value
Input
Deep self-coding network
Mean value
Standard deviation
Reconstruction error analysis
N
Sliding window method for calculating mean value and standard threshold value
Both exceed the threshold value
Y
Academic early warning

Fig. 4. Schematic diagram of early warning process of the model

5 Experimental results and analysis

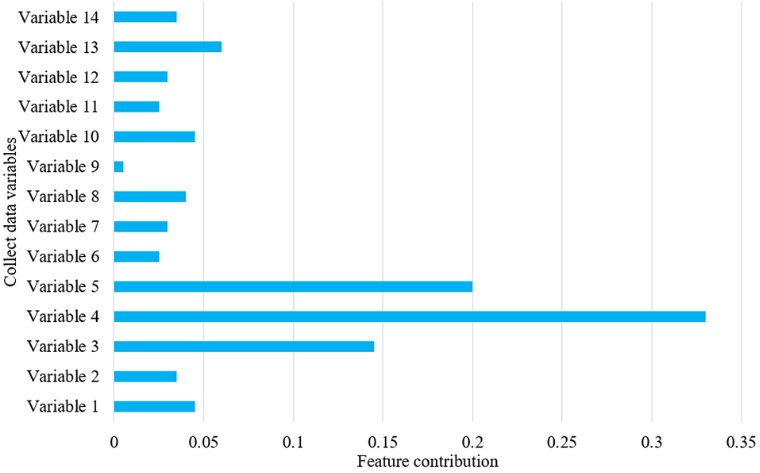


Fig. 5. Contribution diagram of each feature

This paper extracted features of various academic and behavioral data in different categories. Figure 5 compares the contributions of corresponding features of collected data variables to the academic early warning model of college students. Variable 4 (learning quality, 0.33) has the highest feature contribution and plays the most critical role in the early warning model. Variable 3 (course difficulty, 0.145) and Variable 5 (online learning participation, 0.2) also have relatively high feature contributions, and have a significant impact on the model. Variable 9 (book borrowing records, 0.005) has the lowest feature contribution and can be almost ignored in the model, indicating that Variable 9 has minimal impact on the early warning results. Therefore, when constructing and optimizing the early warning model, this paper should focus on Variables 4, 3 and 5, because they have the greatest impact on the early warning results.

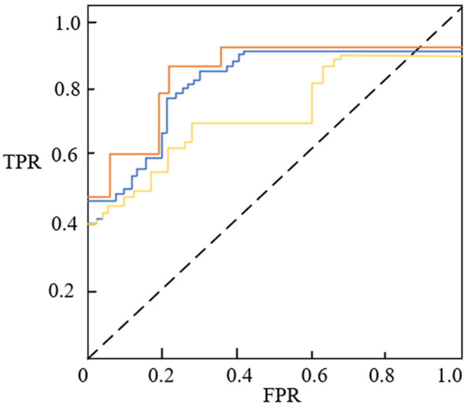


Fig. 6. Experimental results of receiver operating characteristic (ROC) curve

Figure 6 shows the experimental results of the ROC curve. It can be seen that the constructed early warning model has the largest ROC area of 0.88, which verifies the effectiveness of applying deep self-coding network to the academic early warning model.

Table 1. Early warning classification of students

Group	Early Warning Level	Range of Scores	Percentage
1	High early warning	[0,50]	8%
2	Middle early warning	[50,62]	16%
3	Low early warning	[62,68]	18%
4	No early warning	[68,100]	58%

Table 1 shows early warning classification of students output by the model constructed in this paper. It can be seen that the proportion of students in the high early warning group is 8%, with their scores ranging from 0 to 50, these students are faced with serious academic risks. The proportions of students in the middle, low and no early warning groups are 16%, 18%, and 58%, respectively. For students with high and middle early warning, universities and teachers should focus on them and provide targeted tutoring and psychological support, thus reducing their academic risks. For students with low early warning, universities and teachers should still pay attention to their learning status and provide necessary guidance and support, though their academic risks were relatively low. For students in the no early warning group, universities and teachers should continue to monitor their learning progress to ensure that they can continue to maintain good academic performance.

Table 2 further provides the statistical description of students with high early warning. The following conclusions were drawn based on the table:

Table 2. Statistical description of students with high early warning

Data Category	Minimum Value	Maximum Value	Average Value	Standard Deviation
Course scores	42.3	51.0	46.66	2.41
GPA	42.8	77.4	60.21	12.22
Course difficulty	221	548	358.42	106.42
Learning quality	28	53	41.65	8.12
Online learning participation	19	44	30.31	7.21
Attendance	35	71	58.21	14.22
Leave records	191	450	344.51	110.29
Social network behavior	14.33	26.91	18.71	3.61
Book borrowing records	15.84	24.68	20.21	2.65

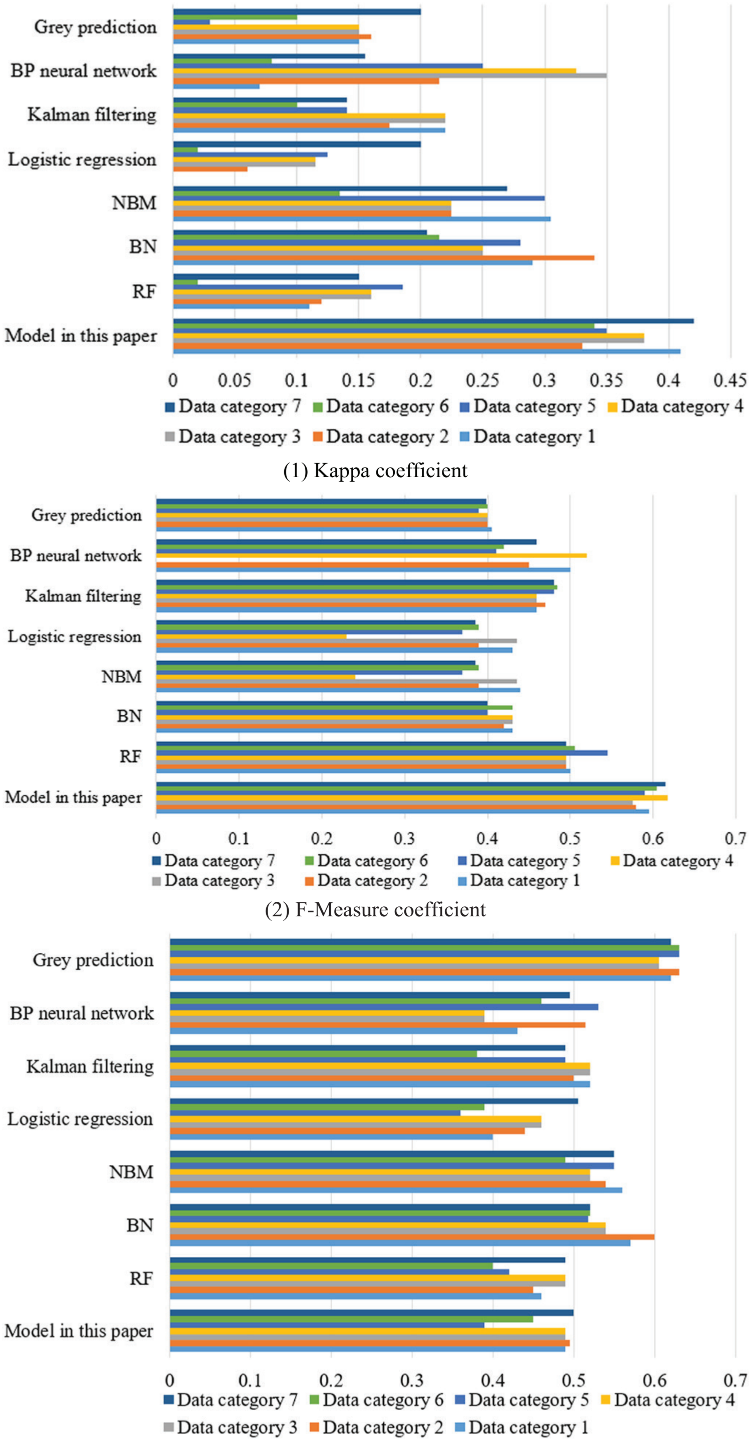
(Continued)

Table 2. Statistical description of students with high early warning *(Continued)*

Data Category	Minimum Value	Maximum Value	Average Value	Standard Deviation
Extracurricular activity participation	0	4	0.66	1.45
Self-evaluation	2	31	19.99	9.71
Evaluation from teachers	2	21	8.45	7.22
Evaluation from classmates	14	149	93.99	58.61
Rewards	0	/	0	5.21
Punishments	3	8	6.61	2.03
Grade ranking	431.9	492.4	457.45	22.87
Major ranking	430.5	491.6	460.51	24.55
Class ranking	437.9	491.3	464.62	19.61

1. Course scores: these students generally had lower scores, with 46.66 average scores and 2.41 standard deviation, indicating that their scores concentrated in the lower score range.
2. GPA: these students also had relatively low GPA, with 60.21 average value and 12.22 standard deviation, also reflecting these students had poor academic performance.
3. Course difficulty: 358.42 average value and 106.42 standard deviation indicated that the difficulty of the courses selected by these students fluctuated greatly.
4. Teaching quality and evaluation from teachers: their average values and standard deviations were 4165, 8.45, 8.12 and 7.22, respectively, indicating that these students may not have received high-quality teaching support in some courses.
5. Extracurricular activity participation: the average value and standard deviation were 0.66 and 1.45, indicating that these students had a low participation degree in extracurricular activities.
6. Self-evaluation and evaluation from classmates: their average values and standard deviations were 19.99, 93.99, 9.71 and 58.61, respectively, indicating that certain differences existed in the two types of evaluations of these students.
7. Attendance, leave records, and punishments: the average values and standard deviations were 58.21, 344.51, 6.61, 14.22, 110.29 and 2.03, respectively, indicating that these students had unstable attendance and leave situation, and may have a certain degree of disciplinary offence.
8. Social network behavior and book borrowing records: the average values and standard deviations were 18.71, 20.21, 3.61 and 2.65, respectively, indicating that these students had lower participation degree in social network behavior and book borrowing.
9. Grade ranking, major ranking, and class ranking: the ranking data indicated that these students had lower overall academic performance level.

Figure 7 compares the academic early warning effects of different warning models in different data categories, and specific indicators include Kappa coefficient, F-Measure coefficient, and RMSE. The seven data categories include academic performance, course information, attendance, student behavior, personal information, student evaluation, and other related data.



(1) Kappa coefficient

(2) F-Measure coefficient

(3) Root mean square error (RMSE)

Fig. 7. Comparison of academic early warning effects of different data categories

According to the Kappa coefficient of various academic prediction models in different data categories, it can be seen that among all data categories, the prediction effects of the model in this paper are generally good, and the Kappa coefficient shows a high level in all data categories, indicating that the model in this paper provides relatively accurate academic predictions in different data categories. Logistic regression, Naive Bayesian Model (NBM), and grey prediction model perform well in some data categories, but their prediction effects still lag behind that of the model in this paper. Models, such as Random Forest (RF), Bayesian network (BN), Kalman filtering, and BP neural network, generally have poor prediction effects in various data categories, and may require further optimization and adjustment. According to the F-Measure coefficient of various academic prediction models in different data categories, it can be seen that among all data categories, the prediction effects of the model in this paper are generally good, and the F-Measure coefficient shows a high level in all data categories, indicating that the model in this paper provides relatively accurate academic predictions in different data categories. The RF model performs well in some data categories, but its prediction effects still lag behind that of the model in this paper. BN, NBM, logistic regression, Kalman filtering, BP neural network, and grey prediction model generally have poor prediction effects in various data categories. According to the RMSE of various academic prediction models in different data categories, it can be seen that among all data categories, the model in this paper generally has good prediction effects, and RMSE shows a low level in all data categories, indicating that the model in this paper provides accurate academic predictions in different data categories. Models, such as RF, BN, NBM, logistic regression, Kalman filtering, and BP neural networks, have significant prediction errors in various data categories, and their early warning effects are relatively moderate.

6 Conclusion

This paper studied the academic early warning model of college students based on big data analysis. After elaborating on the key points of collecting the academic early warning model data based on big data analysis, this paper explained the reasons of calculating the Pearson correlation coefficient of the collected big data. This paper constructed an academic early warning model of college students based on deep self-coding network, provided the construction process, and explained its working principle. After optimizing the model parameters, this paper analyzed the model reconstruction error based on sliding window statistical method, and further improved the prediction ability and generalization performance of evaluating the deep self-coding network model, thus obtaining higher academic early warning accuracy. This paper offered the distribution diagram of academic early warning scores of college students participating in the experiment, and compared the contributions of corresponding features of collected data variables to the academic early warning model. Then this paper drew the ROC curve, which verified the effectiveness of applying deep self-coding network to the academic early warning model. In addition, this paper offered early warning classification of students output by the model constructed in this paper, the statistical description of students with high early warning, and analysis results. Finally, this paper compared the academic early warning effects of different models in different data categories, and verified the performance advantages of the model constructed in this paper.

7 References

[1] Retnawati, H. (2020). Diagnosis of learning difficulties in mathematics for students resolving problems related to material in the pythagorean theorem for 8th grade students in SMP 1 Todanan and SMP Muhammadiyah 9 Todanan, academic year 2018/2019. Journal of Physics: Conference Series, 1581: 012026. https://doi.org/10.1088/1742-6596/1581/1/012026

[2] Xiao, G., Chen, X. (2015). English academic writing difficulties of engineering students at the tertiary level in China. World Transactions on Engineering and Technology Education, 13(3): 259–263.

[3] Dou, H., Liu, Y. (2022). Optimization and management system for academic early warning of college students. In 2022 IEEE 2nd International Conference on Educational Technology (ICET), Beijing, China, pp. 125–129. https://doi.org/10.1109/ICET55642.2022.9944408

[4] Duong, H.T.H., Tran, L.T.M., To, H.Q., Van Nguyen, K. (2022). Academic performance warning system based on data driven for higher education. Neural Computing and Applications, 35: 5819–5837. https://doi.org/10.1007/s00521-022-07997-6

[5] Luo, J. (2020). Study on academic early warning system for information engineering vocational students. Journal of Physics: Conference Series, 1682: 012017. https://doi.org/10.1088/1742-6596/1682/1/012017

[6] Zhang, J., You, C., Huang, J., Li, S., Wen, Y. (2020). Research on application of frequent pattern growth algorithm in academic early warning. In Proceedings of the 2020 8th International Conference on Information and Education Technology, Okayama, Japan, pp. 116–121. https://doi.org/10.1145/3395245.3395247

[7] Zhu, J., Kang, Y., Zhu, R., et al. (2019). College academic achievement early warning prediction based on decision tree model. In Big Data: 7th CCF Conference, BigData 2019, Wuhan, China, pp. 351–365. https://doi.org/10.1007/978-981-15-1899-7_25

[8] Yang, P., Yang, M. (2019). Research on the management model of university students academic early warning based on big data analysis. In 2019 International Conference on Communications, Information System and Computer Engineering (CISCE), Haikou, China, pp. 639–642. https://doi.org/10.1109/CISCE.2019.00148

[9] Santoso, L.W. (2018). Early warning system for academic using data mining. In 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, pp. 1–4. https://doi.org/10.1109/ICACCAF.2018.8776788

[10] Yin, M., Zhao, J., Sun, S. (2016). Key course selection for academic early warning based on Gaussian processes. In Intelligent Data Engineering and Automated Learning–IDEAL 2016: 17th International Conference, Yangzhou, China, pp. 240–247. https://doi.org/10.1007/978-3-319-46257-8_26

[11] Dai, J.R., Li, M.Y., Li, W.W., Lu, Z., Zhang, Z.G. (2014). Setting of academic warning based on multivariate copula functions. Applied Mechanics and Materials, 571: 156–163. https://doi.org/10.4028/www.scientific.net/AMM.571-572.156

[12] Dai, J.R., Li, M.Y., Li, W.W., Xia, T., Zhang, Z.G. (2014). Application of Monte Carlo simulation in college and university academic warning. Advanced Materials Research, 955: 1817–1824. https://doi.org/10.4028/www.scientific.net/AMR.955-959.1817

[13] Xu, Y. (2022). Research on academic early warning and assistance under artificial intelligence vision: current situation, trend and application. In 2022 3rd International Conference on Education, Knowledge and Information Management (ICEKIM), Harbin, China, pp. 268–273. https://doi.org/10.1109/ICEKIM55072.2022.00067

[14] Dong, J., Liu, X., Wang, Z., Chen, L., Wang, F., Tang, J. (2022). Research on Academic Early Warning Model Based on Improved SVM Algorithm. In 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dalian, China, pp. 1186–1191. https://doi.org/10.1109/TOCS56154.2022.10016199

[15] Chen, S. (2022). Improved fuzzy algorithm for college students' academic early warning. Mathematical Problems in Engineering, 2022: 5764800. https://doi.org/10.1155/2022/5764800

8 Author

Kun Wang, is a graduate from Shandong University, holding a Master's degree. He currently works at Zibo Vocational Institute, and his main research areas include student affairs management and ideological and political education.

Article submitted 2023-02-05. Resubmitted 2023-04-27. Final acceptance 2023-05-05. Final version published as submitted by the authors.