Identification of Poverty in College Students Using Campus Dining Consumption Data: An Elman Neural Network Approach

Wenhui Li[a], Yiming Wang[b]

Xi'an Shiyou University, Xi'an, China

[a]2603738851@qq.com, [b]wmnj66@163.com

Abstract. In recent years, China's colleges and universities have launched the "poor college student aid" programme, which aims to provide scholarships for students who are unable to attend school due to family difficulties. Therefore, how to accurately identify students with financial difficulties in colleges and universities is of great significance to the financial aid work of colleges and universities. The purpose of this paper is to study the identification of poor students based on restaurant consumption data, with the aim of providing a more scientific and accurate method for efficient financial aid work, so as to improve the quality and social fairness of higher education. This paper uses data mining theories, such as K-means clustering analysis algorithm and Elman neural network prediction model, combined with the information provided by a university, such as the amount and number of students' consumption throughout the day for three years, to study the method of identifying poor students from the perspective of big data. Based on the final calculation of the F1 values, it can be concluded that all F1 values exceed 98%. This indicates that the model's results are highly accurate and would effectively assist students in need.

Keywords: Principal Components Analysis; Triple classification; Elman; Prediction; Identification of needy students.

1 INTRODUCTION

Poverty alleviation and education support have long been regarded as one of the important tasks of social development, and it is of great significance in improving the inequality of educational opportunities and the living conditions of people in impoverished areas. As the level of higher education in China rises, university education needs not only students with excellent academic performance, but also students with foresight and creativity. For this reason, universities have set up financial assistance programmes for poor students to help those who face the prospect of their studies being compromised for financial reasons.

The Central Committee of the CPC and the State Council attach great importance to financial support for students from poor families, explicitly stating that "we should

M. Yu et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Informatization Education (ICBDIE 2024)*, Advances in Intelligent Systems Research 182,
https://doi.org/10.2991/978-94-6463-417-4_13

strive to run every school well, teach every student well, and not let a single student drop out of school because of his or her family's financial difficulties". In 2007, the State Council Opinions on Establishing and Improving the Policy System of Financial Support for Students with Financial Difficulties in Ordinary Higher and Middle Vocational Schools and Vocational Schools were introduced. With the introduction of the 'Opinion of the State Council on the Establishment and Improvement of the Policy System of Financial Support for Students with Financial Difficulties from Families in Ordinary Higher and Middle Vocational Schools and Vocational Schools', the state has increased its support for students from poor families, which has helped to alleviate the financial difficulties of some students to some extent.

However, the implementation of this programme has not been entirely successful. In the actual work of financial aid management, there are still some problems that need to be considered, mainly the following three points: 1) The independent information comes mainly from the materials provided by the colleges and universities, and there may be some poor students who are reluctant to submit declarations due to psychological factors.2) When assessing the situation, there are judges who don't know the students and they understand the situation only based on the information profiles.3) Due to fluctuations in the number of students each year, there may be cases where the degree of financial difficulty of students' families does not match the level of support. As a result, greater demands have been placed on the issue of accurate classification of funded students.

The purpose of this paper is to use the canteen consumption data of university students at school, use principal component analysis to perform dimensionality reduction on the eigenvalues after data pre-processing, and use the dimensionality reduced features to perform K-means clustering, and finally apply the Elman model to classify and predict poor students. The experimental results indicate that the method achieves satisfactory results in identifying both suspected non-poor students and students with excessive poor consumption characteristics. This data can provide support for financial aid work for poor college students.

2 APPROACH

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical technique that transforms multiple indicators into a set of composite indices, known as principal components, through dimensionality reduction. These components, which are linear combinations of the original variables, are uncorrelated and provide superior performance compared to the original variables.

To accurately represent the original p variables, it is necessary to identify additional principal components beyond the first. This includes determining the second, third, and fourth principal components. The second principal component should not duplicate information captured by the first principal component. This is achieved by ensuring that the covariance between the two principal components is zero, resulting in orthogonal

directions. The specific method for determining each principal component is described below:

Let Z_i represent the i th principal component,i =1,2,⋯,p, which can be set as Equation 1:

$$\begin{cases} Z_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p \\ Z_2 = c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Z_p = c_{p1}X_1 + c_{p2}X_2 + \cdots + c_{pp}X_p \end{cases} \tag{1}$$

For each value of i, there exists Equation 2:

$$c_{i1}^2 + c_{i2}^2 + \cdots + c_{ip}^2 = 1 \tag{2}$$

This This study analysed three years of restaurant consumption data collected from university students. The data were processed using Excel software to identify six key indicators, and the dataset was fully analysed using these indicators. Principal component analysis was then conducted using SPSS 26.0.[1]. PCA is a long-standing issue with a vast literature available. In this text, we only provide a brief overview (for a comprehensive statistical perspective, see[2]).

2.2 K-Means Clustering

Unlabeled data is a common occurrence and organizing it appropriately is becoming an essential procedure in the era of big data. Clustering analysis is a crucial data mining technique that involves organizing a group of unlabeled data into multiple semantically consistent clusters. This is achieved by ensuring high intra-cluster similarity and low inter-cluster similarity. Clustering analysis is a useful technique for extracting valuable information from unlabeled data. It has been applied in various fields, including agriculture, finance, image processing, and medicine. K-means clustering via Lloyd's method is currently the most popular among the many clustering algorithms. This simple yet powerful approach is considered one of the top-10 algorithms in the field of data mining.[3]

For a given data set containing n d-dimensional data points, where $x_i \in R^d$ and the number of data subsets to be generated is K, the K-means clustering algorithm organises the data objects into K divisions$C = \{c_k, i = 1,2,\cdots,K\}$. Each partition represents a classC_K, and each classC_K has a particular centre μ_i. Euclidean distance is chosen as the criterion for judging similarity and distance, and the distance from each point within the class to the cluster centre μ_i is calculated as the sum of the squared distances and the sum of squared distances is calculated as in Equation 3:

$$J(C_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \tag{3}$$

For each class, the clustering objective is to minimise the total distance sum of squares.

$$J(C) = \sum_{k=1}^{K} J(c_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} d_{ki} \|x_i - \mu_k\|^2 \quad (4)$$

According to least squares and Lagrange's principle, the clustering centre should be taken to be in the category that reduces the total sum of squared distances. This is because the total sum of squared distances tends to decrease as the number of categories increases in the K-means clustering algorithm (J(C) = 0 when K = n). Therefore, the total sum of squared distances can only be minimized for a definite number of categories K.[4]

The study stratified the entire sample into three categories: economically disadvantaged students, economically challenged students, and students with no economic difficulties, resulting in a value of k = 3.

2.3 Elman Neural Network

2.3.1 Introduction to Elman Neural Networks.

The Elman neural network (Figure 1) is a type of recurrent neural network that is an improvement on the BP neural network. It was first proposed by J.L. Elman and has the ability to adapt to time-varying and memory functions. It is suitable for a variety of regression and classification tasks, making it a powerful and widely used neural network model.

In order to classify data using the Elman model, the weights and bias values must be learned through a training process. The training data includes input sequences and their corresponding labels. The model's output is compared to the true labels to calculate a loss function, which is then used to update the model's parameters using a back propagation algorithm. This process is repeated until the model converges, meaning the loss function reaches its minimum value.[5]

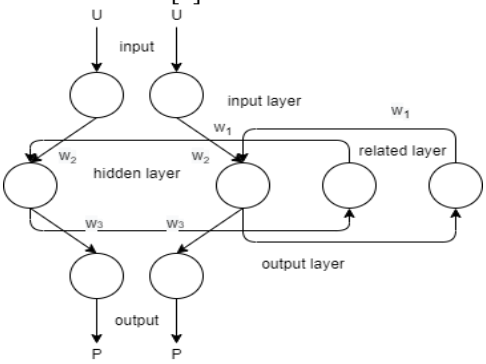

Fig. 1. Elman neural network.

Taking the above figure as an example, the expression for the neural network's nonlinear state space is as follows:

$$y(k) = g(\omega^3 x(k)) \quad (5)$$

$$x(k) = f(\omega^1 x_c(k) + \omega^2(\mu(k-1))) \quad (6)$$

$$x_c(k) = x(k-1) \tag{7}$$

The equation above shows that y is the output node vector with m dimensions, x is the intermediate layer node unit vector with n dimensions, u is the input vector with r dimensions, x_c is the feedback state vector with n dimensions, ω^3 is the connection weight between the intermediate layer and the output layer, and ω^2 is the connection weight between the input layer and the intermediate layer. The take-up layer-to-intermediate layer-to-connection weights are represented by ω^1. The transfer function of the output neuron, which is a linear combination of the outputs of the intermediate layer, is represented by 'g(*)'. The transfer function of the intermediate layer neuron, is represented by 'f(*)'.[6] Other long and short-term memory and adaptive boosting prediction models can be found in[7].

2.3.2 Elman Neural Network Workflow.

The Elman neural network workflow comprises several steps. Firstly, data is imported and the weights and thresholds of the network are initialised. Next, the sample is divided into training and test sets. The number and hierarchy of neuron nodes are determined based on the number of nodes in the input and output layers. Suitable transfer functions and training algorithms are selected to achieve nonlinear mapping of the network. Input data preprocessing involves normalisation or standardisation to avoid accuracy problems caused by extreme feature values. During training, parameters such as learning rate and maximum number of iterations are set. The backpropagation algorithm is used for error backpropagation and weight updates. In the prediction stage, error analysis is performed to compare Elman's predicted values with the actual values. An inverse normalisation operation is applied to obtain the correct output. Finally, the prediction results and error tables are printed. [8]

2.4 Precision-Recall Curve

The PR curve illustrates the relationship between accuracy and recall by displaying the change in accuracy rate as the recall rate varies. This method is useful for evaluating the performance of classification models in situations where the data is imbalanced or the positive classes are of greater importance. The precision rate, also known as the positive predictive value, is denoted by P and is defined as the probability that all predicted positive samples are actually positive. The recall rate, also known as the check-all rate, is denoted by R and refers to the true labeling, which is the probability of correctly categorizing actual positive samples as positive. Table 1 shows the mixing line matrix. The variables l, m, and n represent the number of true samples in categories 1, 2, and 3, respectively. The variables r, s, and t represent the number of predicted samples in the three categories, respectively. The variable w represents the total number of samples. The variables a, b, and c represent the number of correctly categorized samples, while d, f, g, e, i, and h represent the number of incorrectly categorized samples.[9]

Table 1. Confusion matrix for the three-classification problem.

		Classification results			
		Category 1	Category 2	Category3	total
truth tag	Category 1	a	d	f	l
	Category 2	g	b	e	m
	Category 3	i	h	c	n
	total	r	s	t	w

The calculation of its P and R values is shown in Equation 8:

$$P = \frac{a}{r}, R = \frac{a}{l} P = \frac{a}{r}, R = \frac{a}{l} \tag{8}$$

For the triple classification problem, a macro-averaging approach was used to calculate as shown in Equation 9:

$$P_{macro} = \frac{1}{3}\sum_{i=1}^{3} P_i, R_{macro} = \frac{1}{3}\sum_{i=1}^{3} R_i \; P_{macro} = \tfrac{1}{3}\textstyle\sum_{i=1}^{3} P_i \, , R_{macro} = \tfrac{1}{3}\sum_{i=1}^{3} R_i \tag{9}$$

where P_i and R_i represent the precision rate and recall rate of i, respectively. On this basis, the researchers also proposed the F-score method[10], i.e Equation 10:

$$F - score = (1 + \beta^2)\frac{PR}{\beta^2 P + R} \tag{10}$$

3 EXPERIMENTS SETTINGS

3.1 Data pre-processing and feature selection

The pre-processing of non-associative features and screening for missing values and outliers is crucial in reducing resource consumption and improving model accuracy in the presence of a large dataset. This is the first step in the empirical study.

The first step in data processing involves identifying and addressing missing values, duplicate entries, irrelevant data, and outliers.

Screening was conducted to remove illogical time data and non-representative data. Secondly, the three-year overall consumption of students corresponding to the serial number (x_1), the total number of effective consumption(x_2), the average amount of money spent per consumption (x_3), the extreme deviation of effective consumption(x_4), the median(x_5), and the gender(x_6)were calculated using Excel's SUM, AVERAGE, COUNTIF, and SUMIF functions. These values were then used for subsequent descriptive analysis, principal component analysis, and cluster analysis using SPSS 26.0.

4 EXPERIMENTAL ANALYSIS

4.1 A test of data standardization and the usefulness of factor analysis

The raw data was processed using SPSS 26.0 software and transformed into Z-scores. The suitability of factor analysis was evaluated through the Kaiser-Meyer-Olkin (KMO) test and Bartlett's sphere test. [11]The KMO value was 0.762 and the Sig value was 0.000, indicating a correlation between the indicators. These results suggest that the factor analysis method is appropriate for the data.

4.2 Categorising Student Poverty Levels

4.2.1. Feature Dimensionality Reduction.

Table 2 presents the process of deriving the principal components from the correlation matrix:

Table 2. Correlation Matrix.

Initial eigenvalue			Extract the sum of squares and load		
amount	Variance%	Accumu-late%	amount	Variance%	Accumu-late%
2.590	43.242	43.242	2.590	43.242	43.24
1.843	30.776	74.018	1.843	30.776	74.018
0.888	14.828	88.846			
0.581	9.694	98.540			
0.063	1.045	99.586			
0.025	0.414	100.000			

Table 2 shows that the first two principal components explain 74.018% of the total variance. This indicates that these two components can represent 74.018% of the consumption information of the original six student samples. The extracted principal components have a certain level of certainty for the subsequent cluster analysis. Therefore, two principal components, y_1 and y_2, were extracted.

A linear combination of y_1 and y_2 is obtained based on the principal component coefficients.

$$y_1 = 0.387 \times x_1 + 0.100 \times x_2 + 0.541 \times x_3 + 0.346 \times x_4 + 0.543 \times x_5 + 0.346 \times x_6 \tag{11}$$

$$y_1 = 0.387 \times x_1 + 0.100 \times x_2 + 0.541 \times x_3 + 0.346 \times x_4 + 0.543 \times x_5 + 0.346 \times x_6 \tag{12}$$

Determine the composite score, y:

$$y = 0.43242 \times y_1 + 0.30776 \times y_2 \tag{13}$$

4.2.2. Process of K-means clustering analysis.

The two principal component scores were used as variables for cluster analysis. The cluster analysis algorithm with iterative solution using Euclidean distance as a classification index was employed. The clustering results are presented in Table 3.

Table 3. Clustering Results.

Class	amount
1	1805
2	2154
3	1454

The table above displays an effective sample size of 5413, with one sample missing. Category 2 comprises 2154 students, all of whom have a composite score of less than 0, identifying them as exceptionally poor. Category 3 includes 1454 students, who are identified as moderately poor based on their composite ranking score. Category 1 consists of 1805 students, who are identified as not poor due to their high principal components y_1 and y_2, as well as their composite score.

4.3 Prediction Models for Classification

The Elman model classification prediction of the sample set is done using MATLAB. First, 5415 samples were divided into training and test sets in a ratio of 7:3. Then the data was normalised and the model was built using the newelm function. The training parameters were set to a maximum number of iterations of 2000, a target error of 1e-5 and a learning rate of 0.01.

Figure 2 displays the training set prediction results, while Figure 3 shows the test set prediction results.

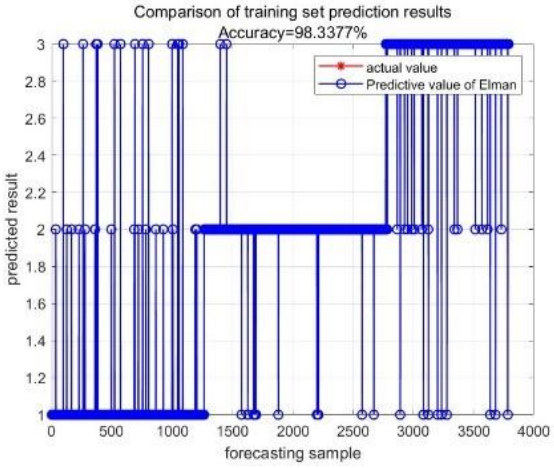

Fig. 2. Training set prediction results.

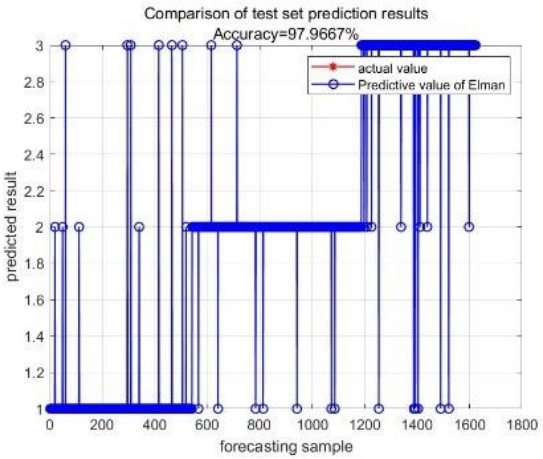

Fig. 3. Test set prediction results.

As illustrated in the aforementioned figure, the blue region depicts the predicted classification outcomes derived from the Elman model, and the red region denotes the actual classification results. Notably, there is a remarkable level of overlap between the two regions. Moreover, by evaluating the accuracy rate, we ascertain that this particular outcome exhibits an impressive accuracy of 97%.

Figures 4 and 5 present the confusion matrices for the training and test sets, respectively, demonstrating the performance of the Elman neural network model.

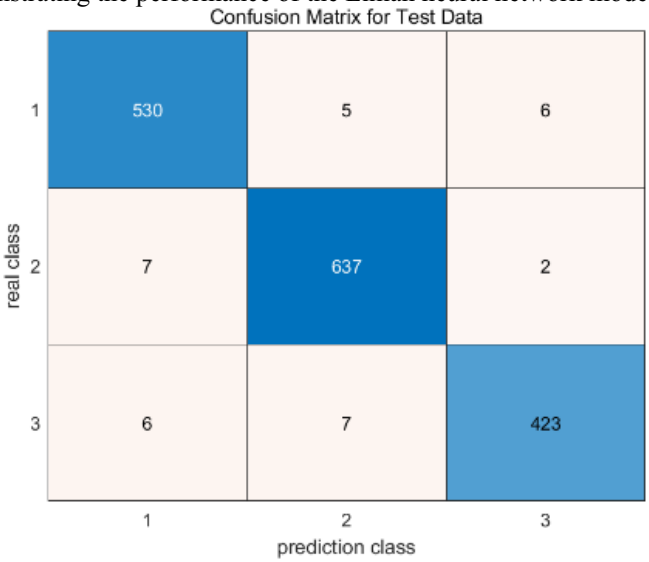

Fig. 4. Training set confusion matrix.

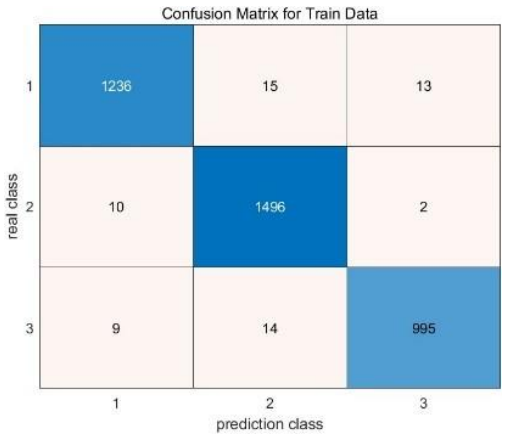

Fig. 5. Test set confusion matrix.

From the two confusion matrices above, the individual indicators of the confusion matrix in Table 4 were calculated.

Table 4. Confusion matrix evaluation indicators.

Evaluation Indicator	Accuracy%	P/R%	F-score%
Training set	97.6	98.4/98.2	98.4
testing set	97.2	98.7/98.6	98.6

The evaluation metric P represents the accuracy of identifying poor students by measuring the proportion of correctly predicted poor students. A higher value of P indicates a better ability of this model to identify negative samples. The experimental results demonstrate a high level of checking accuracy, affirming the high accuracy of the established data model in predicting the degree of poverty among college students.

The full check rate (R) indicates the proportion of correctly predicted samples, with higher values indicating greater recognition ability of positive samples.[12] The model's full detection rate was 98%.

In order to assess the accuracy of the model more accurately, the concept of F1 value is proposed,[13] according to the formula in equation 5, it is concluded that all the values of F1 are greater than 98%, it can be seen that the results of the model are ideal, and it can be a good way to help all the poor students who need help.

5 CONCLUSIONS

This paper addresses the current challenges in identifying economically disadvantaged students and proposes a method that combines data mining techniques with poor student identification and prediction. In the data pre-processing stage, feature dimensionality reduction by principal component analysis is applied to reduce the number of fea-

tures, effectively reducing noise, redundancy and overfitting classifiers. The Elman algorithm is used to train the k-means clustering results to obtain the classification prediction model. And the practicality of the experiment is analysed by the accuracy and the calculation of the testing accuracy and the full testing rate.

REFERENCES

1. Nafis Faizi, Yasir Alvi, Chapter 2 - Data management and SPSS environment**For datasets, please refer to companion site: https://www.elsevier.com/books-and-journals/book-companion/9780443185502, Editor(s): Nafis Faizi, Yasir Alvi, Biostatistics Manual for Health Research, Academic Press, 2023, Pages 17-43, ISBN 9780443185502.
2. Jing Zhang, Qian Liu. A study on variability analysis of threshold selection for principal component analysis[J]. Data Acquisition and Processing,2022,37(05):1012-1017. DOI: 10. 16337/j.1004-9037.2022.05.006.
3. Haowen Zhang, Jing Li, Junru Zhang, Yabo Dong, Speeding up k-means clustering in high dimensions by pruning unnecessary distance computations, Knowledge-Based Systems, Volume 284, 2024, 111262, ISSN 0950-7051
4. WANG Sen, LIU Chen, XING Shuaijie. A review of research on K-means clustering algorithm[J]. Journal of East China Jiaotong University,2022,39(05): 119-126. DOI: 10. 16749/ j.cnki.jecjtu.20220914.001.
5. Feng Jiang, Qiannan Zhu, Jiawei Yang, Guici Chen, Tianhai Tian, Clustering-based interval prediction of electric load using multi-objective pathfinder algorithm and Elman neural network, Applied Soft Computing, Volume 129, 2022, 109602, ISSN 1568-4946,
6. Shou Y, Meng T, Ai W, Xie C, Liu H, Wang Y. Object Detection in Medical Images Based on Hierarchical Transformer and Mask Mechanism. Comput Intell Neurosci. 2022 Aug 4;2022: 5863782. doi: 10.1155/2022/5863782. PMID: 35965770; PMCID: PMC9371842.
7. R. Ying, Y. Shou and C. Liu, "Prediction Model of Dow Jones Index Based on LSTM-Adaboost," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021, pp. 808-812, doi: 10.1109/ CISCE52179.2021.9445928.
8. Beckermann B, Goreinov S A, Tyrtyshnikov E E. Some remarks on the Elman estimate for GMRES[J]. SIAM journal on Matrix Analysis and Applications, 2005, 27(3): 772-778.
9. Qin Feng, Huang Jun, Cheng Zekai, et al. Research on accuracy evaluation method of multi-label classifiers[J]. Computer Technology and Development, 2010, 20(01): 46-49.
10. Zhang Kaifang, Su Huayou, Dou Yong. A new method for accuracy evaluation of multi-classification tasks based on confusion matrix[J]. Computer Engineering and Science, 2021, 43(11): 1910-1919.
11. Hill B D. The sequential Kaiser-Meyer-Olkin procedure as an alternative for determining the number of factors in common-factor analysis: A Monte Carlo simulation[M]. Oklahoma State University, 2011.
12. Lu Guiming, Zhang Yuan, Zhou Zhimin. Research on prediction of poor students classification based on machine learning[J]. Computer Applications and Software, 2019, 36(01): 316-319.
13. Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis[J]. Neurocomputing, 2022, 501: 629-639.