Assessment and ML-based Prediction for Research Ability of Postgraduates

Hongcheng Liu, Wen Hu, Dong Sun*

University of Electronic Science and Technology of China, No.2006, Xiyuan Avenue, West Hi-tech Zone, Chengdu 611731, China

*sundong@uestc.edu.cn

Abstract. The assessment and prediction of research ability are very important in education. This paper 1) proposes an assessment scheme based on principal component analysis (PCA), which assigns a reasonable score for a postgraduate, and 2) proposes a prediction model based on multiple linear regression, which suggests the research ability of a postgraduate based on indirectly related information. Data from some postgraduates in University of Electronic Science and Technology of China (UESTC) are used to evaluate the performance. Experiments show the effectiveness and reliability of the proposed scheme and model.

Keywords: education, research ability, machine learning (ML), principal component analysis (PCA), multiple linear regression

1 Introduction

Research ability is the ability to find problems, analyze problems, solve problems and carry out specific research work in unknown fields with scientific thinking and appropriate methods [1]. The cultivation of research ability has been paid increasing attention by colleges and universities, and it is also one of the important indicators in the comprehensive strength evaluation of colleges and universities. However, the current methods can't comprehensively assess or predict the research ability of postgraduates.

In recent years, Machine Learning (ML) has been widely used in many fields. More and more scholars try to use ML to assess and predict research ability [1]-[4]. For example, Ying Chen et al. proposed a GA_GBoost model to predict the research ability of college students [3]. Wei Wu et al. achieved a relatively objective and accurate assessment of scientific research potential based on Learning Vector Quantization (LVQ) neural network [4]. However, few works separate assessment and prediction, and the introduction of deep learning leads to a lack of interpretability.

This paper insists that 1) assessment should be based on the directly related research outcomes, while the prediction should be based on indirectly related information, and 2) where a simple model works well, there is no need to introduce any heavy model. Therefore, this paper uses the principal component analysis (PCA) [5] to construct the assessment scheme and uses multiple linear regression [6] to obtain a prediction model.

X. Ding et al. (eds.), *Proceedings of the 2023 4th International Conference on Big Data and Social Sciences (ICBDSS 2023)*, Atlantis Highlights in Social Sciences, Education and Humanities 12,
https://doi.org/10.2991/978-94-6463-276-7_44

The structure of this paper is introduced as follows. Section 2 introduces the key technics used in this paper, i.e., PCA and multiple linear regression. Section 3 discusses the assessment based on PCA and the outcomes. Section 4 discusses the prediction based on multiple linear regression. Section 5 carries experiments to show the performance. Section 6 concludes this work and proposes future expectations.

2 Preliminaries

2.1 Principal Component Analysis

Principal component analysis (PCA) is a kind of linear dimension reduction algorithm, which makes data in a higher dimension projected into a space with a lower dimension while preserving information as much as possible. It can be derived from two perspectives equivalently, i.e., variance maximization, and reconstruction error minimization. Let's take the first perspective for example and show the derivation [5].

For a set of data $\{\boldsymbol{x}_n\}$, n=1, 2, …, N, where $\boldsymbol{x}_n$ is a variable in D-dimension Euclidean space, the aim is to project them into an M-dimension space ($M<D$) and maximize the variance of the projected data. Consider M=1 first and denote its direction as unit vector $\boldsymbol{u}_1$. Then the mean of the projected data can be denoted as $\boldsymbol{u}_1^T\overline{\boldsymbol{x}}$ where $\overline{\boldsymbol{x}}$ is the mean of original data with the form of (1).

$$\overline{\boldsymbol{x}} = \sum_{n=1}^{N} \boldsymbol{x}_n / N \tag{1}$$

And the variance of the projected data can be calculated as (2), where $\boldsymbol{S}$ is the covariance matrix with the form of (3).

$$\sum_{n=1}^{N} (\boldsymbol{u}_1^T\boldsymbol{x}_n - \boldsymbol{u}_1^T\overline{\boldsymbol{x}})^2 / N = \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 \tag{2}$$

$$\boldsymbol{S} = \sum_{n=1}^{N} (\boldsymbol{x}_n - \overline{\boldsymbol{x}})(\boldsymbol{x}_n - \overline{\boldsymbol{x}})^T / N \tag{3}$$

Thus, the constrained optimization problem can be denoted as (4), and can be transformed into an unconstrained optimization problem as (5) further.

$$\arg\max \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1,\ s.t.\ \boldsymbol{u}_1^T\boldsymbol{u}_1 = 1 \tag{4}$$

$$\arg\max \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 + \lambda_1(1 - \boldsymbol{u}_1^T\boldsymbol{u}_1) \tag{5}$$

Calculate the deviation of the expression, let it be 0, and solve this equation. Then, it can be noted that $\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1$ holds, which means $\boldsymbol{u}_1$ is an eigenvector of $\boldsymbol{S}$ and $\lambda_1 = \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1$ is the corresponding eigenvalue (also vairance). This eigenvector is called the first principal component.

Likely, additional components can be defined: choose new directions within the directions orthogonal with considered directions, and maximize the variance of the projected data. Then, the optimal *M*-dimension space can be constructed by *M* eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_M$ of ***S***, corresponding with *M* max eigenvalue $\lambda_1, \lambda_2, \ldots, \lambda_M$.

2.2 Multiple Linear Regression

Multiple linear regression means predicting the independent variable *y* as $\hat{y}$ with the linear combination of multiple dependent variables *m* and an intercept $\hat{\theta}_0$, as (6).

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 m_1 + \ldots + \hat{\theta}_k m_k \tag{6}$$

The mission is to find appropriate parameters $\boldsymbol{\theta}$ in the expression according to observed data ***X*** and try to minimize the prediction error, which can be summarized into an optimization problem as (7).

$$\arg\min_{\hat{\theta}} \sum_{n=1}^{N} [y_n - (\hat{\theta}_0 + \hat{\theta}_1 m_1^{(n)} + \ldots + \hat{\theta}_k m_k^{(n)})]^2 \tag{7}$$

Through derivation [6], the parameter can be calculated as (8).

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{8}$$

3 Research Ability Assessment

3.1 Index Introduction

Assessment means assigning a score for a postgraduate which reflects his/her research ability. In order to avoid subjectivity, several objective indexes an assessment framework based on the indexes are to be put forward and. The indexes include x_1, x_2, x_3, and x_4, where x_1 stands for one's publication, x_2 stands for the number one's patents obtained, x_3 stands for one's competition awards obtained, and x_4 stands for one's innovation projects finished. And x_1, shown in (9), is a weighted sum of 4 sub-indexes, i.e., $x_{1,1}$, $x_{1,2}$, $x_{1,3}$, $x_{1,4}$, standing for the number of papers published in Chinese Academy of Sciences ranking (CAS) Q1, CAS Q2, CAS Q3/Q4, and other journals or conferences [7].

$$x_1 = 0.4x_{1,1} + 0.3x_{1,2} + 0.2x_{1,3} + 0.1x_{1,4} \tag{9}$$

3.2 Assessment Based on Principal Component Analysis

Let's denote the data matrix *A* as where every row represents a postgraduate with four indexes, and $(A)_{ij}=a_{ij}$. The detailed form of *A* is shown in Appendix A. There are a total of N=39 postgraduates in our dataset.

Firstly, data process should be conducted, i.e., handling the indexes positively and non-differently. Since all indexes are positively correlated with one's research ability, only normalization with (10) is needed, where $s_j = \sqrt{\sum_{n=1}^{N}(a_{ij} - \mu_j)^2 / (N-1)}$ and $\mu_j = \sum_{n=1}^{N} a_{ij} / N$. The matrix after normalization is denoted as $\tilde{A}$ where $(\tilde{A})_{ij} = \tilde{a}_{ij}$.

$$\tilde{a}_{ij} = (a_{ij} - \mu_j) / s_j \tag{10}$$

Secondly, calculate the correlation matrix $\boldsymbol{R}$ using (11), where $(\boldsymbol{R})_{ij}=r_{ij}=r_{ji}$ is the correlation coefficient between index i and j.

$$r_{ij} = \sum_{n=1}^{N} \tilde{a}_{ni}\tilde{a}_{nj} / (N-1), i, j = 1, 2, 3, 4 \tag{11}$$

Thirdly, calculate the eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4$ and corresponding normalized eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4$ where $\boldsymbol{u}_j = \left[u_{1j}, u_{2j}, u_{3j}, u_{4j}\right]^T$. And 4 new indexes, i.e., the first to fourth principal components, can be generated from old ones using (12).

$$\begin{aligned} y_1 &= u_{11}\tilde{x}_1 + u_{21}\tilde{x}_2 + u_{31}\tilde{x}_3 + u_{41}\tilde{x}_4 \\ y_2 &= u_{12}\tilde{x}_1 + u_{22}\tilde{x}_2 + u_{32}\tilde{x}_3 + u_{42}\tilde{x}_4 \\ y_3 &= u_{13}\tilde{x}_1 + u_{23}\tilde{x}_2 + u_{33}\tilde{x}_3 + u_{43}\tilde{x}_4 \\ y_4 &= u_{14}\tilde{x}_1 + u_{24}\tilde{x}_2 + u_{34}\tilde{x}_3 + u_{44}\tilde{x}_4 \end{aligned} \tag{12}$$

Fourthly, calculate the information contribution rate of each eigenvalue α_j as (13) and the cumulative contribution rate β_p as (14).

$$\alpha_j = \lambda_j / \sum_{j=1}^{N} \lambda_j \tag{13}$$

$$\beta_p = \sum_{j=1}^{p} \lambda_j / \sum_{j=1}^{N} \lambda_j = \sum_{j=1}^{p} \alpha_j \tag{14}$$

Choose a specific p which makes β_p very close to 1, and use the first p principal components to construct the assessment model with (15). Z is the assessment [8].

$$Z = \sum_{j=1}^{p} \alpha_j y_j \tag{15}$$

3.3 Outcomes

By applying spectral decomposition, the eigenvalues and eigenvectors can be calculated, and the results of principal components analysis is shown in Table 1.

Table 1. The results of principal component analysis.

No.	Eigenvalue	Contribution rate	Cumulative contribution rate
1	2.1006	52.52%	52.52%
2	0.9521	23.80%	76.32%
3	0.5873	14.68%	91.00%
4	0.3601	9.00%	100.00%

Since the first 3 components have contained more than 90% original information, we choose the first 3 components to construct the assessment model. The parameters of (12) and (15) are all calculated, which are shown in Appendix B.

And the distribution of the research ability according to the assessment scheme is shown in Fig. 1. It is noted that the distribution of research ability is more like long-tailed distribution [9], rather than normal distribution. It means few postgraduates are outstanding in research while most of them are ordinary.

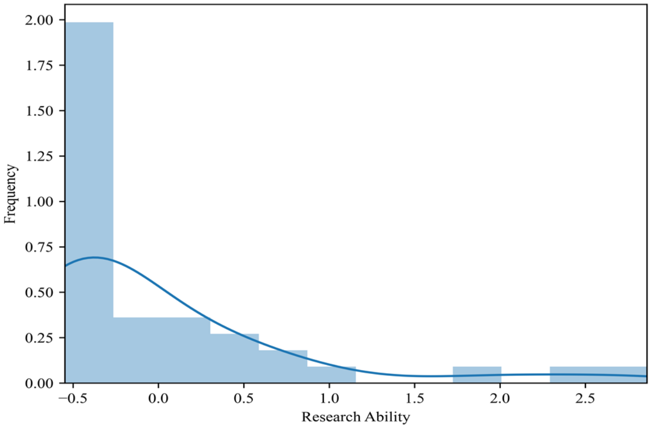

Fig. 1. The histogram and the kernal density estimation of the research ability of 39 postgraduates.

4 Research Ability Prediction

4.1 Indexes Introduction

There are 14 indexes that may be related to a postgraduate's research ability and can be used to predict it. The details and illustration are shown in Table 2.

Table 2. The illustration of indexes for forecasting a postgraduate's research ability.

Index	Illustration	Index	Illustration
m_1	Program type, 0 for academic postgraduate, 1 for professional one, 2 for bachlor-straight-to-PhD and 3 for others	m_8	Help gained from seniors in the same program, 0 for little, 1 for some, and 2 for much
m_2	Admission type, 0 for postgraduate qualification exam, 1 for recommended for admission, and 2 for others	m_9	Important research projects involved
m_3	Rank, a fraction showing one's ability among one's class or program	m_{10}	English ability, mapping CET-4 to [0,0.5] linearly and CET-6 to [0.5,1] linearly
m_4	Grade, 0, 1, 2 for grade 1, 2, 3 of postgraduate and 3, 4, 5 for grade 1, 2, 3 of PhD, etc.	m_{11}	Number of important social practices
m_5	Tutor's research ability, presented by his/her h-index [10]	m_{12}	Max duration of important social practices
m_6	The number of students guided by the same tutor	m_{13}	Hours for research per week
m_7	The help gained from the seniors in the same program, 0 for little, 1 for some, and 2 for much	m_{14}	Academic writing training received, 0 for none, 1 for part of, and 2 for fully trained

4.2 Prediction Based on Multiple Linear Regression

According to (8), the multiple linear regression model can be constructed. Details are shown in Appendix C.

5 Experiments

5.1 Data Description and Evaluation Metrics

There are 39 samples in the dataset, separated randomly with a ratio of 7:3 into training set and testing set. To evaluate the performance comprehensively, 3 different metrics, i.e., Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE), are consider [11], shown as (16)-(18).

$$\text{MAE} = \sum_{n=1}^{N} | y_n - \hat{y}_n | / N \tag{16}$$

$$\text{MAPE} = \sum_{n=1}^{N} | y_n - \hat{y}_n | / | y_n | / N \tag{17}$$

$$\text{RMSE} = \sqrt{\sum_{n=1}^{N} | y_n - \hat{y}_n |^2 / N} \tag{18}$$

5.2 Comparison

To assess the performance of multiple linear regression, we compare it with three other regression methods, i.e., support vector regressor (SVR) [12], random forest regressor (RFR) [13], and deep neural network (DNN) [14], whose complexity increase. The results are shown in Table 3. It can be concluded that multiple linear regression is far better than others in all evaluation metrics.

Table 3. Comparison of different models for research ability prediction.

Metrics / Model	SVR	RFR	DNN	**Multiple regression**
MAE	0.400	0.409	0.498	**0.295**
MAPE	0.900	2.338	0.916	**0.669**
RMSE	0.528	0.485	0.610	**0.396**

And the visualization of prediction result is shown in Fig. 2.

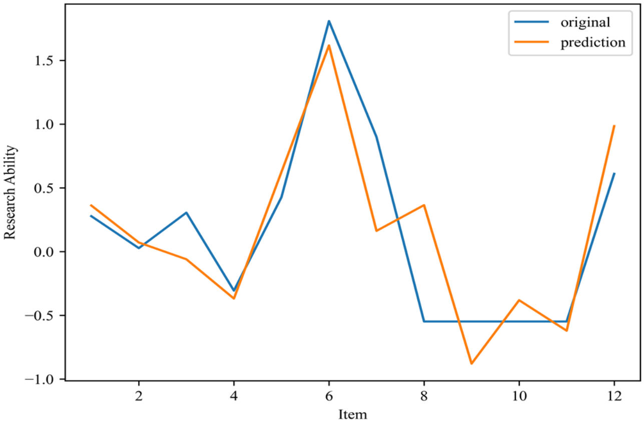

Fig. 2. The result of prediction using multiple linear regression.

6 Conclusion

In this paper, an objective assessment scheme based on PCA is proposed, which suggests that the research ability of postgraduates in UESTC may follow long-tailed distribution. By comparison and analysis with SVR, RFR and DNN, the multiple linear regression model has the best fitting effect in this case, and this model is the most accurate to predict the research ability of postgraduates and the closest to the real situation. Based on this, the prediction model based on multiple linear regression is proposed. Though simplest, it has the best performance, which suggests that non-linear methods may not apply to linear problems well. The experiments illustrate the effectiveness as well as the reliability of the scheme and model.

This paper lays a foundation for future research and references. For example, researchers can use this assessment scheme to test what kind of measures help improve one's research ability most; university can consider what kind of student have the best research ability according to the prediction model; etc.

Data Availability

The codes and data are available at https://github.com/swulhcx/Research-Ability-Assessment-and-Prediction.

Acknowledgement

We appreciate Jierui Zhang for his support and discussion, and the help offered by our colleagues.

References

1. Etkina, Eugenia, et al. "Scientific abilities and their assessment." *Physical Review special topics-physics education research* 2.2 (2006): 020103.
2. Kruit, Patricia M., et al. "Assessing students' ability in performing scientific inquiry: instruments for measuring science skills in primary education." *Research in Science & Technological Education* 36.4 (2018): 413-439.
3. Yin Chen, Xin yang, and Daohe Sun. "Prediction Based on GA_XGBoost Model Research on College Students' Scientific Research Ability." *Mathematics in practice and theory* 51.6 (2021): 318-328.
4. Wei Wu, Nong Wu. "Evaluation of Research Potential of Architecture Graduate Students Based on LVQ Neural Network" *Urban Architecture* 18.7 (2021):32-34+100.
5. Bishop, Christopher M., and Nasser M. Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. No. 4. New York: springer, 2006.
6. Zhou, Zhi-Hua. Machine learning. Springer Nature, 2021.
7. Xu, F., et al. "Ranking academic impact of world national research institutes—by the Chinese Academy of Sciences." *Research Evaluation* 22.5 (2013): 337-350.

8. Shi, Shijie, et al. "Comprehensive evaluation of 17 qualities of 84 types of rice based on principal component analysis." *Foods* 10.11 (2021): 2883.
9. Downey, Allen B. "Evidence for long-tailed distributions in the internet." *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*. 2001.
10. Hirsch, Jorge E. "Does the h index have predictive power?" *Proceedings of the National Academy of Sciences* 104.49 (2007): 19193-19198.
11. Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." *PeerJ Computer Science* 7 (2021): e623.
12. Smola, Alex J., and Bernhard Schölkopf. "A tutorial on support vector regression." *Statistics and computing* 14 (2004): 199-222.
13. Segal, Mark R. "Machine learning benchmarks and random forest regression." (2004).
14. Miikkulainen, Risto, et al. "Evolving deep neural networks." *Artificial intelligence in the age of neural networks and brain computing*. Academic Press, 2019. 293-312.

Appendix

A. The Detailed Form of Matrix *A*

$$A = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(N)} & x_2^{(N)} & x_3^{(N)} & x_4^{(N)} \end{bmatrix} \tag{19}$$

B. The Research Ability Assessment Model of Postgraduates in UESTC

$$\begin{aligned} y_1 &= 0.2304\tilde{x}_1 + 0.5746\tilde{x}_2 + 0.5915\tilde{x}_3 + 0.5165\tilde{x}_4 \\ y_2 &= 0.9614\tilde{x}_1 - 0.2250\tilde{x}_2 - 0.0186\tilde{x}_3 - 0.1572\tilde{x}_4 \\ y_3 &= 0.0384\tilde{x}_1 - 0.3937\tilde{x}_2 - 0.3674\tilde{x}_3 + 0.8417\tilde{x}_4 \end{aligned} \tag{20}$$

$$Z = 0.5252y_1 + 0.2380y_2 + 0.1468y_3 \tag{21}$$

C. The Details of the Constructed Multiple Linear Regression Model

$$\begin{aligned} \hat{y}_n = {} & 0.4149 + 0.7484m_1^{(n)} - 1.0230m_2^{(n)} - 0.0137m_3^{(n)} + 0.8136m_4^{(n)} \\ & + 0.1542m_5^{(n)} - 0.9444m_6^{(n)} - 0.0051m_7^{(n)} + 0.8136m_8^{(n)} + 0.4251m_9^{(n)} \\ & - 0.5050m_{10}^{(n)} - 0.5567m_{11}^{(n)} + 1.0366m_{12}^{(n)} - 0.0163m_{13}^{(n)} - 0.1722m_{14}^{(n)} \end{aligned} \tag{22}$$