Acta Scientific COMPUTER SCIENCES

Volume 3 Issue 11 November 2021

Review Article

# Evaluation of a Hybrid Principal Component Analysis and Cluster Analysis on Students' Learning Achievement and Motivation on Academic Courses

**Chenyue Hong[1], Zejun Hou[1], Yahui Meng[1]* and Timothy Chen[2]**

*[1]Guangdong University of Petrochem Technol, Sch Sci, Peoples R China*

*[2]Division of Engineering and Applied Science, California Institute of Technology, Pasadena, USA*

***Corresponding Author:** Yahui Meng, Guangdong University of Petrochem Technol, Sch Sci, Peoples R China. **Email:** mengyahui@gdupt.edu.cn; Timothy Chen, Division of Engineering and Applied, **Email:** t13929751005@gmail.com

**Received:** June 02, 2021

**Published:** October 11, 2021



## Abstract

Because the comprehensive evaluation system of College Students' performance is not perfect, most schools take the examination results as the only standard to evaluate students, leading to the contemporary college students can only cope with the examination and lack of independent thinking ability and innovation. In this paper, principal component analysis and cluster analysis are used to evaluate students' performance. Taking the real performance of 39 students in class 1 of information and computing science of Science College of Guangdong University of Petrochemical Technology in 2020-2021-1 semester as the evaluation object, this paper uses MATLAB software combined with two analysis methods to analyze more information about students' performance, so as to evaluate students' performance more scientifically, accurately, fairly and fairly..In this paper, principal component analysis and cluster analysis can help teachers improve their teaching quality and teach students in accordance with their aptitude. Through principal component analysis, from the perspective of professional courses and non professional courses, we can give students a relatively clear direction of employment or postgraduate entrance examination, help students find their strengths, and pursue innovative, unique and personalized development.

**Keywords:** Academic Courses; Principal Component Analysis; Cluster Analysis; MATLAB

## Introduction

The traditional student achievement evaluation system uses the method of summation or weighted average. These two methods are relatively general for the evaluation of students' comprehensive performance, which can not reflect the unique characteristics and independent innovation ability of students, and can not distinguish the advantages and disadvantages of students in each course.

At present, most colleges and universities have no fixed requirements on how to score the scores of professional courses. Most courses are given according to the test scores and usual scores, and according to the proportion. But if the test scores and usual scores are given by teachers, they have strong subjectivity. Most teachers give the usual scores according to the attendance rate of students, the times of being late and absent from class, This evaluation method lacks objectivity and scientificity.

Based on the principal component analysis and cluster analysis, this paper improves the above disadvantages, and promotes college students to improve their independent innovation ability.

## The basic principle of principal component analysis

Principal component analysis is to take a mathematical dimension reduction method, find out a few comprehensive variables to replace the original many variables, so that these comprehensive variables can represent the amount of information of the original variables as much as possible, and they are not related to each other.

Set up $X = (X_1, X_2, \cdots X_p)^T$ Is a p-dimensional random variable X. The expectation matrix and covariance matrix of are known. $E(X) = \mu, Var(X) = \sum$ Consider the following linear transformation: $F_I = a_i^T X = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p \ (i = 1,2, \ldots, p)$ **(1)**

$Var(F_i) = a_i^T \sum a_i$ , $Cov(F_i, F_j) = a_i^T \sum a_j \ (i,j = 1,2, \ldots, p)$ ------ **(2)**

Principal component is a linear combination of the original P $F_1$, $F_2$, ..., $F_p$ variables They are not correlated with each other, and the variance reaches the maximum. From the covariance matrix or correlation coefficient matrix of the original data, the eigenvalues of the matrix can be obtained $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$Of which $\lambda_1$ corresponding $F_1$ And so on $Var(F_1) \geq Var(F_2) \geq \cdots \geq Var(F_p)$The contribution rate of the i-th principal component is calculated. The larger the contribution rate is, the more information the i-th principal component synthesizes the original variables.

### The idea of cluster analysis

In this paper, the k-means algorithm is applied. Firstly, K objects are randomly selected as the initial cluster centers, and then the distance between each object and each seed cluster center is calculated, and each object is assigned to the nearest cluster center. Cluster centers and the objects assigned to them represent a cluster. Once all the objects are assigned, the cluster center of each cluster will be recalculated according to the existing objects in the cluster. This process will be repeated until a termination condition is met.

## Principal component analysis and cluster analysis of students' performance

### Evaluation object and evaluation index

Taking 39 students of class 1 of information and Computing Science in Guangdong University of Petrochemical Technology in 2019 as the object, 10 compulsory courses: Discrete Mathematics in 2020-2021-1 semester were selected $X_1$ Career and development planning of College Students $X_2$, workplace English ($X_3$)College Physics Experiment $X_4$, Java application development $X_5$ College Physical Education (III) $X_6$ College Physics (II) $X_7$ Probability theory and mathematical statistics $X_8$ website design and development $X_9$ Mathematical analysis (II) $X_{10}$. As a variable, the 10 courses of the 39 students make up 39×10 matrix, as the original data matrix.

### Data preprocessing

For the convenience of data analysis, the original 11 digit student number is replaced by a new two digit number. For example, the original student number 19124120101 is recorded as 01; And the students who repeat their grades and transfer their majors will be recorded after the 34th. For example, 17124120233 is recorded as 35;19024020207 is 36, and so on.

For the students who failed in the second make-up examination, their scores were taken as the data objects.

When selecting data, because each student's elective courses have different class hours, credits and subjects, it is difficult to compare the results of elective courses, so these courses are abandoned in data processing.

See table 1 for the original data of students' comprehensive scores; After the data processing, the comprehensive score data of

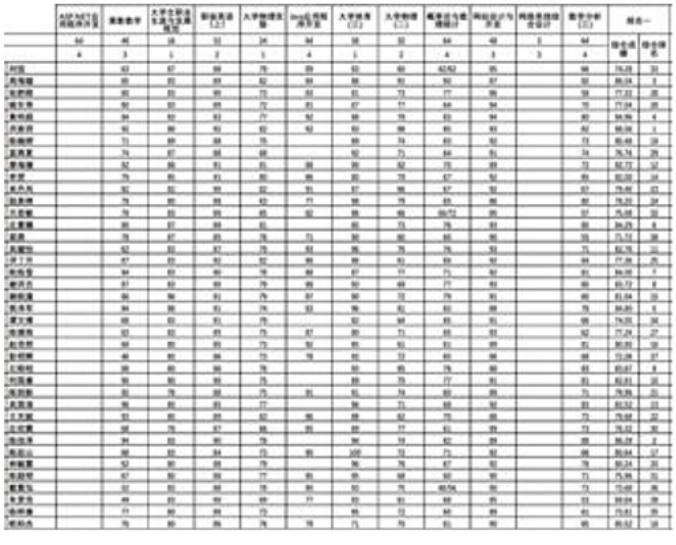

**Table 1:** Comprehensive score data of original students.

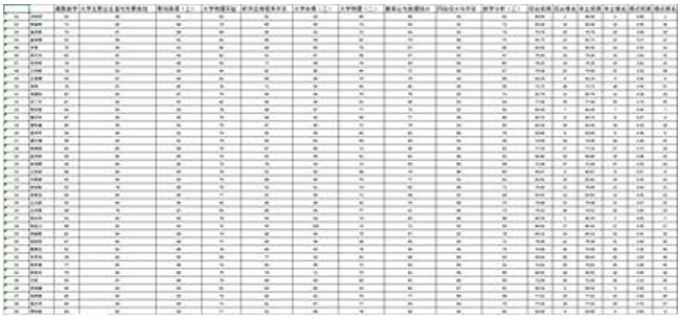

**Table 2:** Comprehensive score data table of students after data processing.

each subject is shown in table 2.

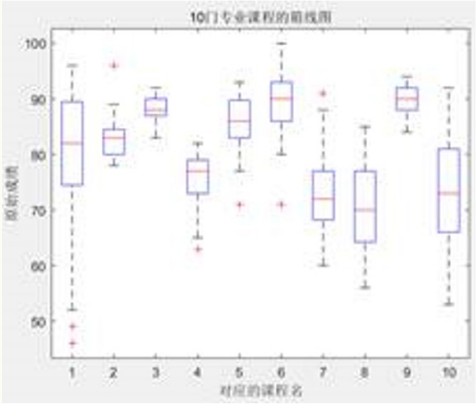


**Figure 1:** Box line diagram of 10 professional courses.

## Statistical analysis of the results of each course

Note: in the x-axis of figure 1, 1 represents discrete mathematics, 2 represents college students' career and development planning, 3 represents Workplace English (I), 4 represents college physics experiment, 5 represents software application development, 6 represents College Physical Education (III), 7 represents College Physics (II), 8 represents probability theory and mathematical statistics, 9 represents website design and development, and 10 represents mathematical analysis (III).

It can be seen from figure 1 that the box length of discrete mathematics, probability theory and mathematical statistics, and mathematical analysis (III) is relatively long, which indicates that the students' scores of these three courses are scattered, that is to say, the comprehensive scores of these three courses can check the students' learning situation to a certain extent. For the four courses of college physics experiment, College Physics (II), probability theory and mathematical statistics, and mathematical analysis (III), the average score and upper quartile are lower than 80, which means that the four courses are relatively difficult for most students.

## Cluster analysis of student achievement

This paper uses k-means algorithm. Considering the scale of the sample, the number of clusters is determined to be in the range of 5-13, and then the sum of squares of the total error of each cluster is calculated. The calculation results show that when the number of clusters is 8, the sum of squares of the total error of the corresponding cluster is the smallest, so the final number of clusters is 8. The clustering results are shown in table 3.

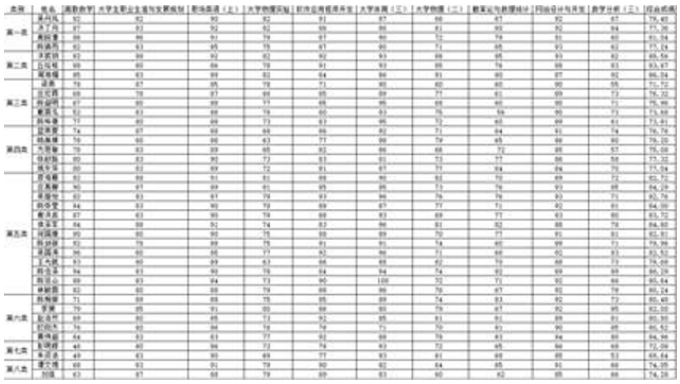

**Table 3:** K-means algorithm results.

It can be seen from table 3 that there is a large gap in the comprehensive data of students among different types, while there is a small gap within the class. The second kind of students' comprehensive score is higher. The first kind of students are weak in probability theory and mathematical statistics, College Physics (II), discrete mathematics and mathematical analysis (III).The third and fourth categories of students' comprehensive performance ranking are relatively low, and there is a serious phenomenon of partial subjects. The fifth category is the one with the largest number of students. There is an extreme phenomenon in mathematical analysis (III) of students in this category. The difference between the highest score and the lowest score is 17 points. However, the comprehensive scores of students in the fifth category are all middle and upper, and the scores of all subjects are about the average. The sixth kind of students' comprehensive scores are lower than middle, but there is no obvious bias phenomenon. The seventh and eighth categories of students' comprehensive performance ranking are relatively low, and for most of the courses are relatively weak.

## Principal component analysis of student achievement

This paper uses MATLAB software to analyze the scores of 10 courses of 39 students in information 19-1 class.

It can be seen from the data in Table 4 that most of the students have differences in the ranking of principal component scores and

| 编号 | 主成分综合得分 | 主成分得分排名 | 综合成绩 | 综合成绩排名 |
|---|---|---|---|---|
| 01 | 3.8379 | 1 | 89.56 | 1 |
| 02 | 0.7862 | 17 | 80.48 | 19 |
| 03 | -0.6826 | 26 | 76.76 | 29 |
| 04 | 1.5475 | 8 | 82.72 | 12 |
| 05 | 1.3912 | 11 | 82.00 | 14 |
| 06 | 0.5438 | 18 | 79.40 | 23 |
| 07 | -1.8925 | 35 | 78.20 | 24 |
| 08 | -2.5374 | 36 | 75.08 | 32 |
| 09 | 2.0235 | 4 | 84.29 | 6 |
| 10 | -2.6741 | 37 | 71.72 | 38 |
| 11 | 1.5438 | 9 | 82.76 | 11 |
| 12 | -0.1728 | 22 | 77.36 | 25 |
| 13 | 1.7073 | 5 | 84.00 | 7 |
| 14 | 1.5689 | 7 | 83.72 | 8 |
| 15 | 1.2524 | 14 | 81.04 | 15 |
| 16 | 1.374 | 12 | 84.80 | 5 |
| 17 | -0.4681 | 25 | 74.05 | 34 |
| 18 | -0.3814 | 24 | 77.24 | 27 |
| 19 | -0.217 | 23 | 80.80 | 16 |
| 20 | -3.0819 | 38 | 72.08 | 37 |
| 21 | 1.534 | 10 | 83.67 | 9 |
| 22 | 1.0635 | 15 | 82.81 | 10 |
| 23 | -0.079 | 20 | 79.96 | 21 |
| 24 | 1.3195 | 13 | 82.52 | 13 |
| 25 | -1.2478 | 29 | 79.68 | 22 |
| 26 | -1.7281 | 33 | 76.32 | 30 |
| 27 | 1.6617 | 6 | 86.29 | 2 |
| 28 | 0.2378 | 19 | 80.64 | 17 |
| 29 | 0.8283 | 16 | 80.24 | 20 |
| 30 | -1.2404 | 28 | 75.96 | 31 |
| 31 | -1.6951 | 32 | 73.68 | 36 |
| 32 | -4.0791 | 39 | 69.64 | 39 |
| 33 | -1.6839 | 31 | 73.81 | 35 |
| 34 | -0.1074 | 21 | 80.52 | 18 |
| 35 | -1.7913 | 34 | 74.28 | 33 |
| 36 | 2.1238 | 2 | 86.04 | 3 |
| 37 | -1.1035 | 27 | 77.32 | 26 |
| 38 | -1.5542 | 30 | 77.04 | 28 |
| 39 | 2.0725 | 3 | 84.96 | 4 |

**Table 4:** Principal component analysis comprehensive ranking and comprehensive performance ranking.

comprehensive scores. The main reason may be that the contribution rate of the first principal component after principal component analysis accounts for 30.44%, which mainly reflects the students' learning ability for professional courses. Some students may have a higher score in the examination, but their ability to learn professional courses is weak, so the ranking of principal component score is lower than that of comprehensive score. Some students have strong learning ability for professional courses, but they have difficulty in learning other subjects, so the ranking of principal component analysis is higher than that of comprehensive score [1-5].

## Conclusion

- Some students may have a higher score in the examination, but their ability to learn professional courses is weak, so the ranking of principal component score is lower than that of comprehensive score. Some students have strong learning ability for professional courses, but they have difficulty in learning other subjects, so the ranking of principal component analysis is higher than that of comprehensive score.
- Through the principal component analysis and cluster analysis of these two methods to analyze the comprehensive performance of students, students can better find the subjects they are good at and interested in. Teachers can also focus on the direction of training students, find the uniqueness of students, and improve their special ability in the future.

## Implications

- For some students with serious partial subjects and weak key subjects, they should communicate with teachers in time, and teachers should pay more attention to these students.
- Teachers can find and solve problems from the principal component analysis and cluster analysis of students' performance, constantly improve their teaching quality and teach students in accordance with their aptitude.
- For college students' career and development planning, college students' innovation and entrepreneurship, college physics and other courses, the credits and class hours of these courses are very small compared with professional courses. The school should appropriately increase the proportion of these courses to cultivate students with both ability and political integrity.

## Acknowledgements

The authors are grateful for the research grants given to Yahui Meng from the Provincial key platforms and major scientific research projects of universities in Guangdong Province, Peoples R China under Grant No. 2017GXJK116, and Guangdong Provincial Higher Education Association Laboratory Management Professional Committee Foundation, Peoples R China under Grant No. GDJ2016048, and the research grants given to Chenyue Hong and Zejun Hou from the College Students' Innovation and Entrepreneurship Projects under Grant No. 73321162.

## Bibliography

1. Fan Shuangnan., *et al*. "Comprehensive evaluation of students' usual performance in computer foundation course based on principal component analysis". *Light Industry and Technology* 34.10 (2018): 175-176.
2. Li Zhenliang and Zhang Zhenqi. "Comprehensive evaluation of students' performance in Independent Colleges Based on principal component analysis". *Science and Technology Horizon* 18(2018): 107-109.
3. Shang Lihua. "Application of fuzzy clustering analysis algorithm in comprehensive evaluation of student performance". *Computer Knowledge and Technology* 10.5 (2014): 1024-1026.
4. Qi Zonghui. "Application of principal component cluster analysis in comprehensive evaluation of students' performance". *Journal of Taiyuan City Polytechnic* 8 (2016): 182-183.
5. Liu Lu., *et al*. "Application of principal component cluster analysis in comprehensive evaluation of student achievement". *Journal of Liaoning University of Technology* (NATURAL SCIENCE EDITION) 32.3 (2012): 200-204.

**Volume 3 Issue 11 November 2021**