Analysis of Students' Academic Achievements Based on Cluster Analysis

Yuncheng Li, Xiao Chen
1075487480@qq.com, 1292389826@qq.com

Shandong Institute of Commerce and Technology Ji'nan, China

Abstract—Cluster analysis is a multivariate statistical analysis method for quantitative classification of multiple samples. In this paper, we use cluster analysis to analyze the correlation of six students' six hundred mark system different course scores, get the relevant courses, and divide the courses with greater similarity into one category. The new course scores after classification are used for cluster analysis, the shortest distance method is used to calculate the shortest distance between two pairs of students in each course, the students' scores are classified, and a classification chart is drawn to obtain the comprehensive ranking of six students' scores.

Keywords-cluster analysis; correlation coefficient; MATLAB; score analysis;

1 INTRODUCTION

In college learning, students' academic achievements are composed of multiple disciplines. According to the characteristics of each discipline, the full score of each discipline is different. Some of these subjects have a full score of 120, some have a full score of 100, some have a full score of 60, and some have a full score of 50. And the assessment of students is some theoretical courses, and some practical courses. The evaluation and analysis of these students' studies is a problem to be solved in this paper. This paper uses cluster analysis to analyze students' strengths and weaknesses in some disciplines, which provides technical and theoretical support for students' teaching and learning in the next step.

2 CLUSTER ANALYSIS

Cluster analysis is a multivariate statistical classification method for quantitative classification of multiple indicators. Similar indicators fall into one category, which makes the research easier [1].

2.1 Index cluster analysis

In the evaluation of samples, in order to avoid the omission of some factors, there will be similar or similar situations when selecting indicators. But too many indicators will bring trouble to the calculation and even disturb important indicators. Therefore, similar indicators can be calculated by indicator similarity analysis, and can be either merged or eliminated in the evaluation, which brings convenience to the evaluation work [2]. The correlation coefficient

ICICA 2022, December 02-04, Chongqing, People's Republic of China

DOI 10.4108/eai.2-12-2022.2327925

calculation or included angle cosine is often used for similarity analysis. This calculation uses computer software, which is fast and accurate, and the results are easy to compare.

2.2 Sample cluster analysis

In the analysis of samples, the sample is often regarded as a sample point, and the distance between samples is used to measure the gap between samples. When calculating the distance, the shortest distance method, the longest distance method, the center of gravity method, the quasi average method, and the sum of squares of deviations are often used. The distance between two pairs of samples is calculated by computer software, and the samples are classified according to the size of the distance. By taking different distance values, the samples are divided into different groups through multi-step classification, and a cluster diagram is drawn to obtain the gap between samples [3].

3 CLUSTER ANALYSIS OF STUDENTS' ACADEMIC ACHIEVEMENTS

Due to the limited space, this paper selects 6 subjects of 6 students (higher mathematics x1, college Chinese x2, ideological and political x3, English x4, labor practice x5, sports x6) for performance analysis. The full scores of the above six courses are x1:120, x2:120, x3:100, x4:100, x5:50, and x6:60. The full score of six courses is shown in Figure 1.

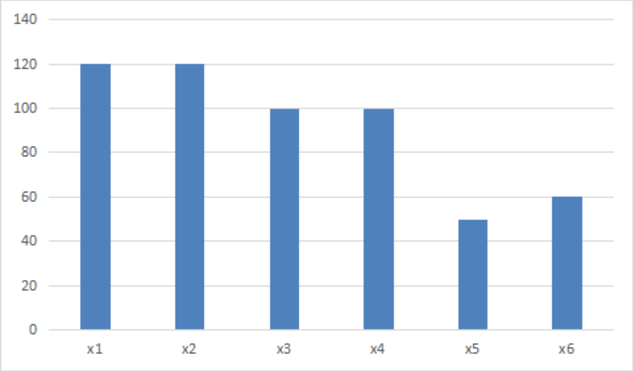

Figure 1. Full score of six courses

The original scores of the six students are shown in Table 1 below.

Table 1 Score Matrix of Six Courses for Six Students

Student	X1	X2	X3	X4	X5	X6
y1	98	102	80	89	45	42
y2	106	90	86	78	42	40
y3	88	80	70	76	35	38

y4	110	96	92	85	48	45
y5	90	112	82	82	50	48
y6	115	96	91	90	36	39

If the above six courses are ranked according to the simple addition of scores, they are: y4, y6, y5, y1, y2, y3, as shown in Figure 2.

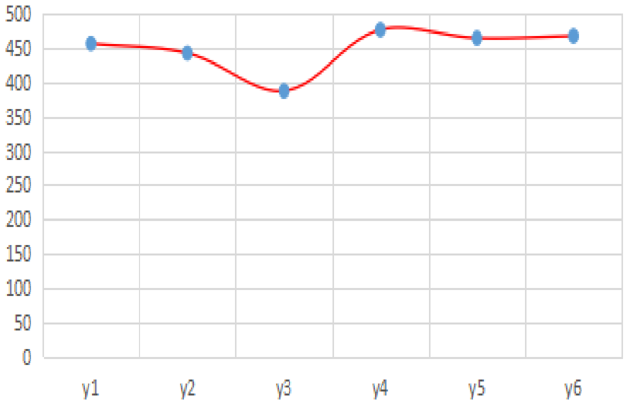

Figure 2. Original score chart of six students

Because the percentile system of each subject is different, simple summation ranking is not scientific. It is easy to see that these six courses have certain relevance when inspecting students, such as the relevance between theoretical courses and practical courses. Moreover, simple addition is easy to lead to the problem that large numbers eat small ones [4].

3.1 Analysis of discipline similarity

In this paper, six courses are regarded as evaluation indicators, and now we use correlation coefficients to analyze the similarity of the six indicators. First, standardize the data, and the processing formula is:

$$\overline{x_{ij}} = \frac{x_{ij}-\overline{x_j}}{s_j}$$

Where, $\overline{x_j}$, s_j are the mean and standard deviation of each column of matrix $(x_{ij})_{6\times 6}$.

Then calculate the correlation coefficient, the formula is

$$r_{jk}=\frac{\sum_{i=1}^{6}(x_{ij}-\overline{x_j})(x_{ik}-\overline{x_k})}{\left[\sum_{i=1}^{6}(x_{ij}-\overline{x_j})^2\sum_{i=1}^{6}(x_{ik}-\overline{x_k})^2\right]^{\frac{1}{2}}}$$

In the above formula, $|r_{jk}|\leq 1, r_{jk}=r_{jk}$. The closer $|r_{jk}|$ is to 1, the more relevant x_j is to x_k ,The closer $|r_{jk}|$ is to 0, the weaker the similarity between x_j and x_k .

The correlation coefficient matrix of the above six courses is calculated by using MATLAB from the above formula, as shown in Table 2.

Table 2 Correlation coefficient matrix of six course scores

Course	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
X1	1	0.4829	0.3526	0.3695	0.2691	0.2036
X2	0.4859	1	0.8653	0.9326	0.3568	0.3696
X3	0.3526	0.8653	1	0.6238	0.5829	0.4869
X4	0.3695	0.9326	0.6238	1	0.3876	0.2698
X5	0.2691	0.3568	0.5829	0.3876	1	0.9592
X6	0.2036	0.3696	0.4869	0.2698	0.9592	1

It can be seen from the above correlation coefficient matrix that language courses such as college Chinese and English have a strong correlation, labor practice and sports have a strong correlation, and advanced mathematics and other courses have no correlation. Therefore, we can consider the courses with strong correlation as a cluster analysis [5]. The longest distance method can be used to cluster the correlation coefficient matrix. The calculation formula is:

$$d_{jk}=1-|r_{jk}|$$

r_{jk} is the correlation coefficient, the smaller d_{jk} can be classified into one category, and the larger d_{jk} can not be classified into one category.

The MATLAB program for the above calculation is as follows:

```
Clc,clear

a=textread(‘matlab.mat’);

d=1-abs(a);

d=tril(d);
```

```
b=nonzeros(d);
b=b’;
z=linkage(b,’complete’);
y=cluster(z,’maxclust’4,)
for i=1:4
  tm=find(t=i);
       tm=reshape(tm,1,length(tm));
       fprintf(s\n,i,int2str(tm));
end
```

According to the results of the program calculation, labor practice and sports were first brought together, and college Chinese and English were brought together. In this way, six courses could be grouped into four categories, {Advanced Mathematics X1}, {College Chinese and English X2}, {Ideological and Political X3}, {Labor Practice and Sports X4}. Take the average value of the courses with strong correlation to synthesize a new course, and then standardize the scores to make the dimensions of each course consistent, as shown in Table 3.

Table 3 Scores of the last six students in clustering and standardization

Student	X1	X2	X3	X4
y1	0.1614	0.1775	0.1597	0.1712
y2	0.1746	0.1561	0.1717	0.1614
y3	0.1450	0.1454	0.1397	0.1438
y4	0.1812	0.1683	0.1836	0.183
y5	0.1483	0.1793	0.1637	0.1929
y6	0.1895	0.1733	0.1816	0.1477

3.2 Cluster analysis of students' achievements

First, use the absolute distance to measure the distance between each course, and use the shortest distance method to measure the distance between classes [6]. The formula is:

$$d(y_i, y_j) = \sum_{k=1}^{4} \left| x_{ik} - x_{jk} \right|,$$

Use MATLAB to calculate the distance matrix of four courses, as shown in Table 4.

Table 4 Distance Matrix of Student Scores

Student	y1	y2	y3	y4	y5	y6
y1	0	0.0564	0.0959	0.0647	0.0406	0.0777
y2	0.0564	0	0.0899	0.0523	0.0890	0.0557
y3	0.0959	0.0899	0	0.1422	0.1103	0.1182
y4	0.0647	0.0523	0.1422	0	0.0737	0.0506
y5	0.0406	0.0890	0.1103	0.0737	0	0.1103
y6	0.0777	0.0557	0.1182	0.0506	0.1103	0

According to the distance matrix, students' scores are divided into different categories according to the shortest distance.

The first step is that all students form their own category, and the platform height is 0; Step 2, take the platform height as 0.0406, y1 and y5 to form a category, which is recorded as c7. At this time, the classification is {c7, y2, y3, y4, y6}. Step 3, take the platform height as 0.0506, y4 and y6 to form a category, which is recorded as c8. At this time, the classification is {c7, c8, y2, y3}. Step 4, take the platform height as 0.0523, c8, y2 to form a category, which is recorded as c9. At this time, the classification is {c7, c9, y3}. Step 5, take the platform height as 0.0564, c7 and c9 are classified into one category, which is recorded as c10. At this time, the classification is {c10, y3}. Step 6, the platform height is 0.0899, and c10 and y3 are classified into one category, which is recorded as c11. At this time, the classification is {c11}. In this way, all students have been classified. See Table 5 for the number of hierarchical clusters and Figure 3 for the clustering diagram.

Table 5 Number of hierarchical clusters

New classes	Classes of the same platform	Platform level height
C7	y1 and y5	0.0406
C8	y4 and y6	0.0506
C9	y2 and C8	0.0523
C10	C7 and C9	0.0564
C11	C10 and Y3	0.0899

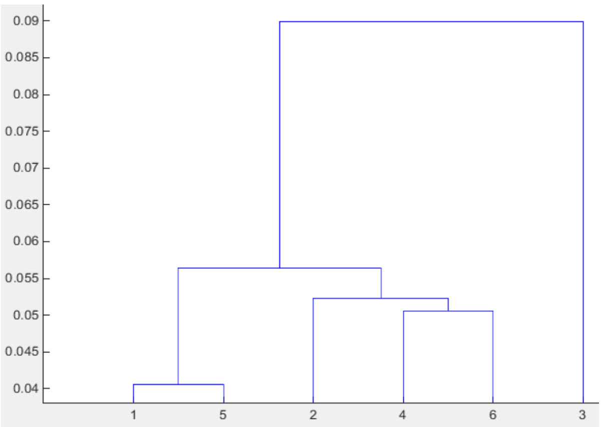

Figure 3. Classification of academic achievements of six students

It can be seen from the above figure that y1 and y5 have the same scores, and y4 and y6 have the same scores, while y4 and y6 have little difference with y2, only y3 has a big difference with other students. By analyzing the original scores, the reason for this classification is that each subject of y3 is quite different from other students, so it is divided into a separate category. The result of this cluster analysis is obviously different from that of weighted average ranking. Cluster analysis is more objective in grouping, while weighted ranking pays more attention to sorting. The MATLAB program for the above calculation is as follows:

```
Clc,clear

a=a=textread('matlab.mat');

y=pdist(a,'cityblock');

yc=squareform(y)

Z=linkage(y)

Dendrogram(z)
```

3.3 Comparison and analysis of cluster analysis results and weighted ranking results

The ranking obtained by cluster analysis is different from the simple weighted ranking, and the results are shown in Table 6.

Table 6 Cluster analysis and weighted ranking comparison

Cluster analysis ranking	**y6 and y4, y5 and y1, y2, y3**
Weighted Sum Ranking	**y4, y6, y5, y1, y2, y3**

It can be seen from the above table that through cluster analysis, y5 and y1 have the same scores, and y6 and y4 have the same scores. Combining their original scores, it can be seen that y6 has much lower scores than y4 in labor sports, but higher scores in higher mathematics. In cluster analysis, the labor practice category belongs to similar subjects. In simple addition, the labor sports category covers the higher mathematics scores, and the labor sports category scores are clustered together for comparison, The performance of y6 naturally rose, which reflected the necessity of merging similar indicators, and also proved the scientificity of cluster analysis method. But it also reflects from another aspect that y6 students should strengthen labor practice and physical training. Y5 and y1, y6 and y4 are grouped together in the comparison, which shows that their overall academic level is not much different, and the simple addition ranking has severely hit the students who rank lower in the same category.

4 CONCLUSION

Clustering analysis method is widely used in engineering, which is an effective method to analyze multi index and multi-objective elements, but rarely used in education. This paper attempts to use cluster analysis to analyze the scores of six students, so as to make some students with better subjects stand out. For example, if higher mathematics and labor practice are simply added together, the students with good higher mathematics may be ranked lower, while the use of cluster analysis also makes the students with partial subjects rank lower. Cluster analysis makes all subjects more balanced, especially the students with outstanding subjects rank higher, At the same time, theoretical courses and practical courses are classified to make the evaluation indicators more scientific [7]. From the above analysis, it can be seen that cluster analysis can show strong advantages in the analysis of students' academic achievements, and this method can be used to analyze all subjects of a large number of students in future teaching.

REFERENCES

[1] Huang H.Y., Liu X.M., Sun H.Y., Yang Z.C. (2019) Application of clustering analysis algorithm in uncertain decision-making. J. Computer Science, (S1):593-597.

[2] Zeng Z.L. (2020) Research on financial risk screening of listed companies based on clustering algorithm.D.Jiangxi University of Finance and Economics.

[3] Duan G.Q., Liu F., Zou C.S. (2019) Application of improved K-center clustering algorithm in performance evaluation. J. Information Technology, (03):40-43.

[4] Xing T., Yong Y.,Hou J., Wu Y., Wu D. (2022) Comprehensive evaluation of water quality based on principal component analysis and hierarchical cluster analysis. J. Sichuan Environment ,41 (04): 131-139

[5] Zhang Z.W., Yang H.L., Fu J., Xie W.J., Li F.M. (2022) Comprehensive evaluation of maize varieties based on principal component analysis and cluster analysis. J. Agricultural Science and Technology Newsletter, (06): 30-35.

[6] Tan Y., Wang G.Q., Chen G., Wu J.Z., Zhou A.M. (2022) Comprehensive evaluation of asparagus quality from different habitats based on principal component analysis and cluster analysis. J. Modern Food Technology ,38 (07):316-326.

[7] Li W.M., Wu X.D., Hu W.C., Liu Q.Y., Chen L.L. (2022) Comprehensive evaluation of different pepper varieties based on principal component analysis and cluster analysis. J. Northern Gardening, (09):55-60.