ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 73
7th International Conference on Education, Management, Information and Computer Science (ICEMC 2017)

Application in the Teaching of Principal Component Analysis

Xueli Ren[1, a*] and Yubiao Dai [1, b]

[1] School of Computer Science and Engineer, Qujing Normal University, Yunnan 655011, China

[a]abiaodai@163.coml, [b]oliveleave@126.com

Keywords: Principal component analysis; Teaching; Knowledge points; Classroom

Abstract. The classroom is the main ways for students to obtain knowledge, there are a lot of curriculum knowledge points and complicated, so it is difficult for students that are required to master all the knowledge. In order to improve the learning effect, the knowledge points should be distinguished to the primary and secondary of the knowledge points effectively, and then the important points of knowledge learning are strengthened to improve the quality of teaching. In this paper, the course of computer foundation as an example, the examination results of students are analyzed by principal component analysis method to know the main influencing factors of the courses, which lays a solid foundation to better carry out the teaching of this course.

Introduction

With the further development of quality education in China, the requirement has been put forward that students are trained having innovative spirit and practical ability. Classroom teaching plays an important role in cultivating students' innovative spirit and practical ability, and it is bound to be the most important place for students to study and practice[1,2]. Teachers, students and teaching content are three elements of the teaching process. They are independent and restrict each other, and form a complete system of practical activities. The teacher and the students is mainly responsible for teaching activities, no teachers, teaching activities can't be carried out, the students can't get effective guidance; no students, teaching activities will lose the object, not random; teaching content, teaching activities will become no rice, passive water, and good development goals the teaching purpose, again good, also can't be achieved[1-3]. Therefore, teaching is a social practice system that composes of the above three basic elements. It is an organic combination of the three basic elements. The change of each factor itself will inevitably lead to the change of the teaching system. Appropriate teaching content not only makes students easy to understand and accept, but also stimulate students' interest in learning and lighten the burden on students, so as to achieve better teaching results. Computer foundation is a complex course, where teaching content is various, if all the knowledge points are treated identically, then it will not only increase the burden of students, but also makes students lose interest. Therefore, it should be reasonable to choose teaching content, explain the important content in detail, so as to achieve a better learning effect in the teaching process.

Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that is used to analyze the interrelationships among a large number of variables and to explain these variables in terms of a smaller number of variables, called principal components, with a minimum loss of information[3-7].

To find the axes of the ellipsoid, we must first subtract the mean of each variable from the dataset to center the data on the origin. Then, we compute the covariance matrix of the data, and calculate the eigenvalues and corresponding eigenvectors of this covariance matrix. Then, we must ensure the orthogonal set of eigenvectors, and normalize each to become unit vectors. Once this is done, each of the mutually orthogonal, unit eigenvectors can be interpreted as an axis of the ellipsoid fitted to the data. The proportion of the variance that each eigenvector represents can be calculated by dividing the eigenvalue corresponding to that eigenvector by the sum of all eigenvalues.

The procedure of calculating principal component analysis is listed as follows:

Calculate Correlation Coefficient Matrix. The correlation coefficient matrix is listed in formula (1)

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1P} \\ r_{21} & r_{22} & \cdots & r_{2P} \\ M & M & M & M \\ r_{P1} & r_{P2} & \cdots & r_{PP} \end{bmatrix} \quad (1)$$

In that, r_{ij} （i，j=1，2，…，p) is correlation coefficient of variables x_i and x_j, which is computed by formula (2)

$$r_{ij} = \frac{\sum_{k-1}^{n} (x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})}{\sqrt{\sum_{k-1}^{n} (x_{ki} - \overline{x_i})^2 \sum_{k-1}^{n} (x_{kj} - \overline{x_j})^2}} \quad (2)$$

As R is a real symmetric matrix, that the relation of rij=rji is satisfy; the element of upper triangular or lower triangular elements is simply calculated.

Compute Eigen values and Eigenvectors. Firstly, the Eigen values λ_i（i=1，2，…，p) is solved based on the characteristic equation |λI-R|=0，then arrange them in order of size from large to small, that is $\lambda_1 \geqslant \lambda_2 \geqslant \cdots, \geqslant \lambda_p \geqslant 0$;Finnally, the characteristic vector e_i（i=1，2，…，p) is calculated respectively corresponding to the characteristic value λ_i.

Calculate Principal Component Contribution Rate and Cumulative Contribution Rate. The contribution rate ci of the principal component zi is calculated by formula (3),the cumulative contribution rate cci is calculated by formula (4).

$$c_i = r_i / \sum_{k-1}^{p} \gamma_k (i = 1,2, \cdots, p) \quad (3)$$

$$cc_i = \sum_{k-1}^{m} \gamma_k \Big/ \sum_{k-1}^{p} \gamma_k \quad (4)$$

Generally, these eigenvalues $\lambda_1, \lambda_2, \cdots \lambda_m$ are extracted, when the cumulative contribution rate reaches 85-95%, which are corresponding to the first, second… The m（m≤p) principal components.

Calculate Principal Component Loading. The loading of the principal component is calculated by formula (5).

$$p(z_k, x_i) = \sqrt{\gamma_k} e_{ki} (i, k = 1,2, \cdots, p) \quad (5)$$

Analyze Main Factors of Curriculum Based on PCA

The classroom is the main way for students to obtain knowledge, teaching content is the soul, which is the main factor influencing the teaching effect, therefore, the content of teaching should be pay close attention in order to enhance students' learning effect. Principal component analysis (PCA) is a kind of statistical analysis method that turns the original variables into a few comprehensive indexes, and it can find the main factors that influence the teaching effect. The specific calculation process is listed as following:

Data Preprocessing. A grade table is constructed where courses are columns and students are rows, courses are labeled from T1, students are labeled from S1.The missing value in the table should be processed. The techniques of missing value imputation are: list wise deletion, mean imputation and some types of hot-deck imputation [8-10]. The listwise deletion is used to deal with missing value in the paper.

Analysis Based on PCA. The main factors that affect the examination result is used to by principal component analysis in the paper based on the examination related information of university computer, which provides the basis for the later course teaching. The procedures refer to the following code:

```
[Z,mu,sigma]=zscore(originalData);
[coeff,score,latent] = princomp(Z);
% Selection dimension
latent=100*latent/sum(latent);
A=length(latent);
percent_threshold=95;
% Percentage threshold, set the number of principal components reserved;
percents=0;                    % Cumulative percentage
for n=1:A
   percents=percents+latent(n);
   if percents>percent_threshold
      break;
   end
end
coeff=coeff(:,1:n);
% The coefficient matrix for the cumulative effect rate of principal components
score=score(:,1:n);
% Principal component that achieves principal component cumulative impact rate requirements
```

The Experiment

As the examination of the university computer foundation is done using the computer, the relevant information of the examination is accurate and perfect. In this paper, the calculation process of principal component analysis is illustrated by taking the computer basic examination information of a university in 2016 as an example. Part of the data for preprocessing and missing value processing is shown in Table 1:

Table 1 the part of the data

	T1	T2	T3	T4	T5	T6
S1	11	8	6	10	13	7
S2	14	8	10	13	15	13
S3	14	8	9	13	8	12
S4	17	8	6	13	14	12
S5	15	10	10	15	18	15
S6	14	10	10	16	16	14
S7	12	8	6	16	10	13
S8	8	8	4	16	15	11
S9	15	10	5	17	18	15
S10	14	10	10	17	9	14
S11	9	8	8	17	10	11
S12	7	2	6	17	9	13
S13	14	8	6	17	15	14
S14	11	10	10	18	16	13
S15	17	8	8	18	16	12

Calculate the correlation coefficients of the processed matrix, and some of the results are shown as follows:

0.2756	0.6454	0.6962	0.0437	-0.1420	0.0292
0.3841	0.3578	-0.6116	0.2333	-0.5267	0.1366
0.4433	0.2978	-0.2814	-0.0816	0.7822	-0.1308
0.3918	-0.4107	0.1978	0.7495	0.0377	-0.2749
0.4694	-0.3684	0.1432	-0.2074	0.0314	0.7612
0.4541	-0.2498	0.0508	-0.5765	-0.2969	-0.5553

The eigenvalues are calculated according to the correlation coefficient matrix, and the contribution rates and cumulative contributions of each principal component are shown in table 2:

Table 2 The contribution rates and cumulative contributions of each principal component

Principal component	eigenvalue	Contribution rate（%）	Cumulative contribution rate（%）
1	35.3631	47.40	47.40
2	11.6194	15.57	62.97
3	10.7098	14.36	77.33
4	9.0711	12.15	89.48
5	4.9181	6.60	96.08
6	2.9284	3.92	100.00

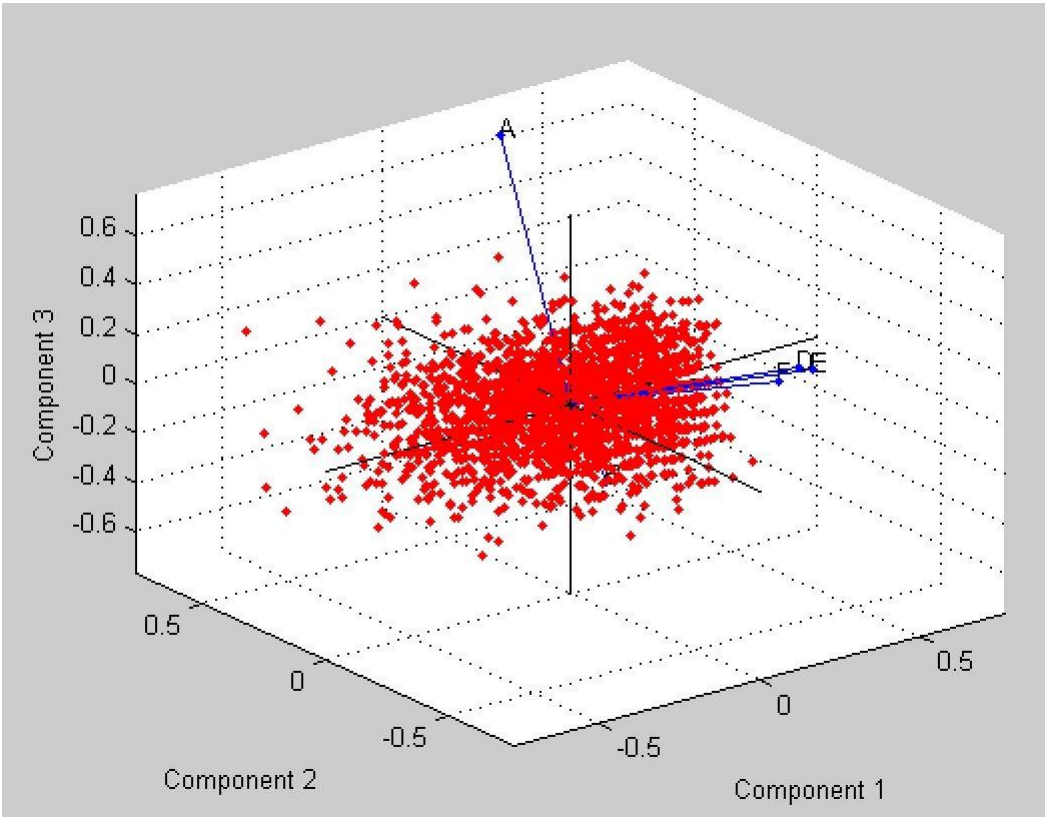

Figure 1. Score distribution

Summary

The teaching content is an important factor in teaching, is the key to influence the teaching effect of the teaching content, the teaching effect is reasonable assurance; principal component analysis is a statistical analysis method, this paper analyzes the use of it to the university computer based test information, so as to determine the main factors affecting the test scores of College Fundamentals of computers, for better to provide a solid foundation for the study of this course..

References

[1] Lin Haiming. Improvement of teaching content of factor analysis [J]. Statistics and Decision,2009.23
[2] Ma Jun,Wang Wei. Quality education and the choice of physical education contents in Higher Medical Colleges. Journal of Guizhou Educational Institute, 2006.02
[3] Lin Haiming,Du Zifang. Pay attention to the problem of comprehensive evaluation in principal component analysis. Statistical research,2013/08
[4] Moritz von Stosch; Cristiana Rodrigues de Azevedo. A principal components method constrained by elementary flux modes: analysis of flux data sets.BMC Bioinformatics,2016-12-15
[5] Kok Chooi Tan, Hwee San Lim, Mohd Zubir Mat Jafri. Prediction of column ozone concentrations using multiple regression analysis and principal component analysis. Atmospheric Pollution Research, 2016-01-09
[6] Principal Component Analysis[EB/OL]. http://baike.baidu.com/link?url=7pCfNrbtulgENxYz2-dgsRLt_8cz3mdzZC82mSomDNAGXhDTU9ldE9CNyy-oeJvQeXcxfj0n3GbHOgmbqec_FzS9ud7IIVCcwtcR7fZX7MRh6fUjVoTnPhnZjGnt_NXv4FB8bBxHzYTebFyIjPz2bq,2016.11
[7] Principal Components Analysis[EB/OL]. https://my.oschina.net/gujianhan/blog/225241,2017.2
[8] Shichao Zhang, Jilian Zhang. Missing Value Imputation Based on Data Clustering[J]. Springer Berlin Heidelberg,2008
[9] Sen Wu , XiaoDong Feng. Missing Data Imputation Approach Based on Incomplete Data Clustering[J]. Chinese Journal of computer,2012.35 (8) :1727-1729
[10] X Feng ， S Wu ， Y Liu. Imputing Missing Values for Mixed Numeric and Categorical Attributes Based on Incomplete Data Hierarchical Clustering[J]. Springer, 2011 :414-424
[11] Cleve Moler. experiment With Matlab[M].Beihang University Press,2013.12
[12] John H.Mathews, Kurtis D.Fink. Numerical Methods Using Matlab, Fourth Edition[M]. Prentice Hall/Pearson, 2005.7