Application in Grade Early Warning of PCA

Xueli Ren[1, a*] and Yubiao Dai[2, b]

[1] School of Information Engineer Qujing Normal University Qujing, China

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

*The corresponding author

Keywords: Early warning; Principal component analysis; Similarity; Clustering

Abstract. Students are misled by the traditional concepts after enrolling university, these make some students face many problems such as no goal of learning, no power, that causes students to face serious consequences that failed in courses and quit, bring serious influence to student management. In order to improve the management level of the school and help the students to get rid of the predicament, it is necessary to make an early warning. Calculate the similarity between the student and other students, and make sure the nearest neighbors based on similarity to estimate grade. Principal component analysis is an effective method to extract the main features, which is used in the process of grade warning. The method is applied to cluster students, and the result shows that it is feasible.

Introduction

The learning atmosphere of students from colleges is relatively free, and the misleading that the grade is 60 is live in community, these making many students indulge themselves or learn by improper learning methods, so their grades are continue to decline, eventually leading to failing in getting the diploma, or even forced to drop out.It not only brings a severe test of students' management and teaching in colleges, but also has a significant negative impact to students in the future work and learning. The school as a place to train personnel, has the obligation to help these students learn the various courses to avoid the occurrence of dropout, which reflects the humanistic education of students, and improving the student management and teaching management level.The early warning can improve the quality of teaching, motivate students to study hard, and strengthen the study of counseling, therefore, the establishment of a system with early warning is necessary that helps students understand their own situation, strengthen the inspectors and help.

The Related Background

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 [4-7].

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, the set of eigenvectors is orthogonalized and normalized 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.

Clustering. The so-called "Like attracts like, if according to the Birds of a feather flock together". If students are clustered based on the performance for a long time, so that students have more in common in the same clustering, and different clustering are different. The different methods of management can be used for different clustering, which can improve the pertinence of management. Similarity is the basis of clustering; the common methods to count similarity are Euclidean Distance, Cosine, Modify cosine and Pearson correlation [8-10].

Euclidean Distance: if uses rating look as the points in Euclidean space, then the distance in the points is similarity for them. The similarity between user I and user j is computed used Formula 1.

$$sim(i,j)=\frac{1}{1+\sqrt{(\sum_{c\in I_j}(R_{i,c}-R_{jc})^2)}} \tag{1}$$

Where $R_{i,c}$ is the rate of item c by user I; $R_{j,c}$ is the rate of item c by user j.

Cosine: If $\vec{i}$ and $\vec{j}$ are rating vectors by user i and user j, then the similarity between user I and user j is computed used Formula 2.

$$sim(i,j)=\cos(\vec{i},\vec{j})=\frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|\|\vec{j}\|} \tag{2}$$

Modify cosine: As the different user's rating scale does not considered in the cosine similarity, the modified cosine similarity is used to improve the defect by minus the average score of user rating for the project. If I_{ij} is the common item set that are rated by user I and user j, I_i and I_j are separately the rate which is rate by user I and j, then the similarity between user I and user j is computed used Formula 3.

$$sim(i,j)=\frac{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R}_i)(R_{j,c}-\overline{R}_j)}{\sqrt{\sum_{c\in I_i}(R_{i,c}-\overline{R}_i)^2}\sqrt{\sum_{c\in I_j}(R_{j,c}-\overline{R}_j)^2}} \tag{3}$$

Where $R_{i,c}$ is the rate of item c by user I; $\overline{R}_i$ and $\overline{R}_j$ are respectively the average rate for the whole items by user I and user j.

Pearson correlation: If the common item set is Iij which include the items rated by user i and user j,, then $sim(i,j)$ of the Pearson correlation similarity of two users I and j is defined as Formula 4.

$$sim(i,j)=\frac{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R}_i)(R_{j,c}-\overline{R}_j)}{\sqrt{\sum_{c\in I_{ij}}(R_{i,c}-\overline{R}_i)^2}\sqrt{\sum_{c\in I_{ij}}(R_{j,c}-\overline{R}_j)^2}} \tag{4}$$

Where $R_{i,c}$ is the rate of item c by user I; $R_{j,c}$ is the rate of item c by user j. $\overline{R}_i$ and $\overline{R}_j$ are average value of rate for the whole items by both user x and user y.

Grade Early Warning Based on PCA

Clustering can not only reduce the difficulty of student management, but also make student management have the advantages of pertinence and high efficiency; Cosine is an effective algorithm to cluster, so it is used in student management in this paper. In the time, PCA is an effective method to extract feature. The process of system with early warning is shown in Fig.1.

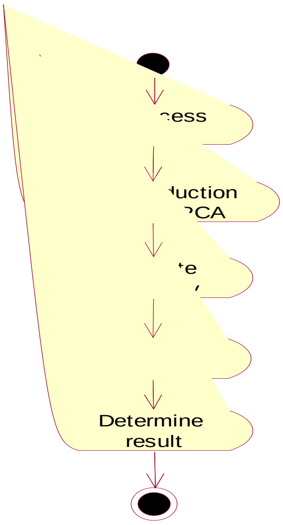

Figure 1. The process of system with early warning

Preprocess data. A grade table is constructed where courses are columns and students are rows, courses are labeled from c1, students are labeled from s1.Two steps are taken to preprocess grades. Firstly, as there are two data types of grades that are numerical and non-numerical, these grades of non-numerical are convert to numeric. Then missing value in the table should be processed. The techniques of missing value imputation are: listwise deletion, mean imputation and some types of hot-deck imputation [11]. The mean imputation is used to deal with missing value in the paper.

Feature Reduction Based on PCA. PCA is a statistical analysis method that deals with the main contradiction; it can resolve the main influencing factors from multiple things, revealing the essence of things, to simplify the complex problem. The goal of PCA is to find r (r<n) new variables, so that they reflect the main characteristics of things, the size of the original data matrix compression.

Compute Similarity. In this step, similarity is computed between the target student and other students using Cosine methods which are described in the foregoing [9-11].

Estimate Grade. A grade estimated is calculated for the target student using similarity calculated in previous step. The steps are as following: Firstly, the k-nearest students are chosen based on similarity. Then the weighted sum is employed to compute estimation whose value is computed as the sum of the metrics' values given by the other students. Each value is weighted by the corresponding the similarity.

Determine Result. The rank of grade is determined based on the estimated grade values.

Example

Some grades of students in a class specialized in software engineering are taken for an example. The grade table is pretreated by the method in step 1. A part of results are shown in Table 1.

Table 1. the data preprocessed

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14
s1	83	75	84	55	68	82	88	77	93	60	62	62	67	74
s2	89	72	78	53	73	75	82	74	91	61	35	63	56	53
s3	83	70	82	63	57	71	89	75	90	51	39	39	61	60
s4	74	67	83	41	95	88	87	86	93	14	72	69	66	76
s5	80	65	83	60	86	77	71	77	89	61	78	78	68	62
s6	94	67	80	66	77	71	92	71	93	61	67	85	65	87
s7	92	68	80	56	61	84	79	71	95	61	65	46	60	61
s8	83	81	83	67	89	80	91	73	94	66	75	63	75	71
s9	80	83	84	56	77	62	74	74	92	57	62	66	70	66
s10	88	72	83	73	71	81	93	79	94	65	64	64	67	61

Feature are selected using PCA ,and the result is shown in Fig. 2.

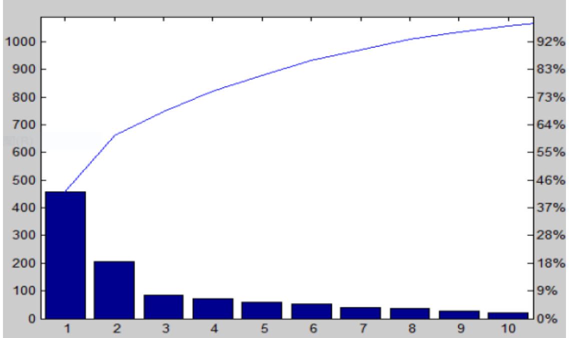

Figure 2. The result of PCA

The set of {C1, C2, C3, C4, C5，C6, C7, C8, C9, C10} is selected to form the main feature set. Similarity is computed based on cosine, 10 nearest neighbors are selected to estimate grade and rank.

Conclusions

The grade early warning is discussed in this paper, and the students are clustered based on similarity, the PCA is used to feature reduction in order to improve the computational efficiency.

References

[1] Zhang yan, Li aiqiu. Reformation on Presentiment System of University Students Score[J].Journal of Shenyang Normal University(Natural Science).2010.4:225

[2] Zhang Wei. Design of early warning system for college students based on Data Mining. Science & Technology Information .2013.6

[3] Zhao Ziyun, Chen Mingxuan. Research on the design of student achievement warning system in the information environment [J]. China In

[4] Dimensionality Reduction [EB/OL]. http://blog.csdn.net/abcjennifer/article/details/8002329.2016-10-20

[5] The mathematical theory of PCA [EB/OL]. http://www.360doc.com/content/13/1124/02/9482_331688889.shtml.2016.11

[6] J Yang, D Zhang, AF Frangi, JY Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. [J]. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. 2004, 26(1):131-7

[7] K Yan, R Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors [J]. IEEE Computer Society，2004.7：506-510.

[8] Goldberg, D., Nichols, D., Oki, B.M., and Terry, D.: Using Collaborative Filtering to Weave an Information Tapestry. Comm. of the ACM, vol.35, no.12, pp.61-70 (1992)

[9] Benjamin Marlin. Collaborative Filtering: A Machine Learning Perspective,21-23

[10] Weike Pan, Evan W. Xiang, Nathan N. Liu and Qiang Yang. Transfer Learning in Collaborative Filtering for Sparsity Reduction.in AAAI10:4-5

[11] Xueli Ren, Yubiao Dai. Application in Effort Estimation of Collaborative Filtering,iscid,2013:331