DOI:10.16644/j.cnki.cn33-1094/tp.2023.11.012

主成分分析和聚类分析在高职学生成绩综合评价中的应用

任东海

(河南经贸职业学院计算机工程学院，河南 郑州 450000)

摘 要：高职学生成绩的合理评价对职业教育的发展具有重要指导意义。本文介绍了主成分分析及聚类分析的基本思想原理、步骤，以及应用过程中需要注意的问题。基于主成分分析对学生成绩建立综合评价方法，构造一个综合评价函数，通过对学生综合评价函数得分的比较，对学生成绩进行排名，并依据所提取的主成分得分对学生进行聚类分析。通过实例分析，证实了主成分分析法及聚类分析在学生成绩综合评价应用中的合理性和必要性，为高职学生成绩管理提供了一种合理且操作简便的方法。

关键词：主成分分析；聚类分析；学生成绩；综合评价

中图分类号：G717 **文献标识码**：A **文章编号**：1006-8228(2023)11-64-05

Comprehensive evaluation of student achievement based on principal component analysis and cluster analysis

Ren Donghai

(*Computer Engineering college, Henan Institute of Economic and Trade, Zhengzhou, Henan 450000, China*)

Abstract: The reasonable evaluation of the achievements of higher vocational students has important guiding significance for the development of vocational education. In this paper, the ideas, procedures and application issues of principal component analysis and cluster analysis are introduced. Based on principal component analysis, a comprehensive evaluation function is constructed. By comparing the scores of students' comprehensive evaluation function, students' scores are ranked, and cluster analysis is carried out on students according to the extracted principal component scores. The rationality and necessity of principal component analysis and cluster analysis in the application of comprehensive evaluation of student achievements are confirmed by the case analysis. It provides a reasonable and simple method for the management of student achievements in higher vocational education.

Key words: principal component analysis; cluster analysis; student achievements; comprehensive evaluation

0 引言

职业院校经常需要对学生成绩进行综合评价，以检测学校管理、教育教学和人才培养的成效，进而做出相应的调整及改进。

目前，学生成绩综合评价应用较多的方法是加权综合评分法或将各种成绩简单的相加，由于加权评分法中的权重设置容易受人为因素的影响，而各项成绩简单的相加又不能体现学生综合的学习能力，这些方法都存在各自的缺点[1]，这里基于主成分分析给出一个简单实用的综合评价方法，并在此基础上进一步对学生成绩进行聚类分析，相对有效的解决了上述评价方法存在的问题，有利于因材施教，提升人才培养质量。

1 数学模型

1.1 主成分分析基本思想

主成分分析是一种数据“降维”分析方法。其基本思想是将原来众多的具有一定相关性的指标，重新组合成一组新的互不相关的综合指标来代替原来指标。通常数学上的处理就是将原来多个指标作线性组合，作为新的综合指标[2]。问题是如何选取相应的

收稿日期：2023-06-27

作者简介：任东海(1988-)，男，河南沁阳人，研究生/硕士，助教，主要研究方向：数据挖掘、统计应用。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

线性组合。

我们考虑比较所有线性组合的方差，方差最大的线性组合（Y_1）称为第一主成分。方差次大的线性组合（Y_2）为第二主成分。并且使 Y_1 与 Y_1 所包含的信不重叠，即 Y_1 与 Y_1 不相关，在数学上就是要求协方差 $Cov(Y_1,Y_2)=0$。依此类推可以构造出多个主成分。

1.2 主成分分析的数学模型

设有 n 个个体元素，每个元素观测 p 项指标（变量）：$X_1,X_2,\cdots,X_P$，原始数据用以下矩阵 X 表示，其中 $X=(X_{1i},X_{2i},\cdots,X_{ni})',i=1,2,\cdots,p$。

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=(X_1,X_2,...,X_p)$$

对 X 的 p 个列向量 $X_1,X_2\cdots X_P$ 作线性组合，得到综合指标向量：

$$\begin{cases} Y_1=a_{11}X_1+a_{21}X_2+...+a_{p1}X_p \\ Y_2=a_{12}X_1+a_{22}X_2+...+a_{p2}X_p \\ \qquad\qquad ..... \\ Y_p=a_{1p}X_1+a_{2p}X_2+...+a_{pp}X_p \end{cases}$$

简写成：

$$Y_i=\alpha_{i1}x_1+\alpha_{i2}x_2+\cdots+\alpha_{ip}x_p,i=1,2,\cdots p$$

对线性组合系数 $\alpha_i=(\alpha_{i1},\alpha_{i2},\cdots\alpha_{ip})$ 作如下限制：α_i 为单位向量，$\alpha_i'\alpha_j=1$，且由下列原则决定：

⑴ Y_i 与 $Y_j(i\neq j,i,j=1,\cdots,p)$ 互不相关；

⑵ Y_1 是 $X_1,X_2\cdots X_P$ 的一切线性组合（系数满足上述要求）中方差最大的[3]，Y_2 是与 Y_1 不相关的 $X_1,X_2\cdots X_P$ 一切线性组合中方差最大的，Y_p 是与 $Y_1,Y_2,\cdots,Y_{p-1}$ 都不相关的 $X_1,X_2\cdots X_P$ 一切线性组合中方差最大的。

满足上述要求的综合指标向量 $Y_1,Y_2,\cdots,Y_p$ 就是主成分，每个主成分的方差代表其含有的信息量的大小，原指标相关系数矩阵的特征值 λ_i 就是对应主成分方差，而组合系数 $\alpha_i=(\alpha_{i1},\alpha_{i2},\cdots\alpha_{ip})$ 就是相应特征值 λ_i 所对应的单位特征向量 u_i。每个主成分的方差贡献率为 $\alpha_i=\lambda_i\big/\sum_{j=1}^{p}\lambda_j$，$\alpha_i$ 越大，说明该主成分含有的综合信息量越大。

1.3 学生成绩综合评价的主成分分析具体步骤

设反映学生质量的 p 个指标为 $x_1,x_2,\cdots x_p$，n 个学生的 p 项指标得分构成了原始数据矩阵 $X=\left[x_{ki}\right]_{n\times p}$，其中 x_{ki} 为第 k 个学生的第 i 项指标得分。

⑴ 首先将原始数据进行标准化处理，这样可以消除量纲的影响以及各指标在数量级上的差别。令 $x_{ki}=\dfrac{x_{ki}-\dot{x}_j}{\sqrt{var(x_j)}}k=1,2,\cdots n;j=1,2,\cdots,p$ 其中 $\dot{x}_{kj},var(x_j)$ 分别表示第 j 项指标的均值及方差。

⑵ 建立矩阵 X 的相关系数矩阵 $R=\left[r_{ij}\right]_{p\times p}$ 其中 $r_{ij}=\dfrac{1}{n-1}\sum_{k=1}^{n}x_{ki}x_{kj}$ 为第 i 项指标与第 j 项指标的相关系数。

⑶ 求出相关矩阵 R 的特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\lambda_p>0$ 及对应的特征向量 $u_1,u_2,\cdots u_p$，其中 $u_i=(u_{i1},u_{i2},\cdots,u_{ip}),(i=1,2,\cdots,p)$。于是得到 p 个主成分：

$$Y_i=u_{i1}x_1+u_{i2}x_2+\cdots+u_{ip}x_p,i=1,2,\cdots,p \qquad (1)$$

第 i 主成分 Y_i 方差越大，对总变差的贡献也越大，其方差贡献率为 $\alpha_i=\lambda_i\big/\sum_{j=1}^{p}\lambda_j$，方差贡献率反映了该主成分所综合提取的原始信息占比。特征值 λ_i 对应的特征向量 u_i 的 p 个分量就是第 i 主成分 Y_i 中 p 个标准化变量的系数，他们的绝对值大小和正负号反映了该主成分与相应变量的相关程度和方向[4]。

⑷ 以每个主成分的贡献率为权数，构造综合评价函数：

$$F=\alpha_1Y_1+\alpha_2Y_2+\cdots\alpha_PY_P \qquad (2)$$

综合评价函数值即为学生成绩的综合评价指标。

将每个学生的 p 项指标得分标准化后代入⑴式，算出主成分值，即主成分得分，再由⑵式即可得到其综合评价函数值，即该学生的综合得分。综合得分越高，表明该学生成绩越好。

当前 $m(1\leqslant m<p)$ 个主成分的累计贡献率 $\sum_{i=1}^{m}\alpha_i$ 达到 85% 以上时，说明前 m 个主成分已包含了原始变量的绝大部分信息，此时可以只取前 m 个主成分构成综合评价函数，达到了降维目的。

1.4 系统聚类分析的方法步骤

聚类分析是根据相似性，将具有共同特点的元素聚成不同的几类，以帮助我们研究元素间的相似性与差异性。可以用距离来度量相似性，为此需定义元素间的距离，以及不同类与类之间的距离。

初始我们将每个元素看成一类，然后将距离最近的两类合并成为一个新类，再次计算新类与其他类的类间距离，再按最小距离准则并类。每次类的数量减少一个，不断迭代，最终所有的元素都并为一类。可

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

以用谱系聚类图形象地表示聚类过程。

由以上思想方法，我们可以总结出聚类分析的如下步骤。

⑴ 数据归约，目的是为了便于计算比较，然后定义距离，包括度量元素间的距离及类与类间的距离。

⑵ 根据距离定义，计算元素两两之间的距离，得元素间的距离矩阵$D^{(0)}$。

⑶ 并类开始，初始（第1步：$i=1$）n个元素自成一类，类的个数$k=n$，第t类$G_t=\{X_{(t)}\}(t=1,2,\cdots n)$，此时类间的距离就是单个元素间的距离（即$D^{(1)}=D^{(0)}$）。然后对步骤$i=2,\cdots,n$执行步骤⑷。

⑷ 对步骤i得到的距离矩阵$D^{(i-1)}$，找到其中类与类之间距离最小的两个类，将其合并为一个新类。此时类的总数k减少1，即$k=n-i+1$。

⑸ 重新计算新类和其他类的距离[5]，得新的距离矩阵$D^{(i)}$。重复步骤⑷，直到类的总个数为1时停止。

⑹ 画出谱系聚类图，显示聚类过程。

⑺ 根据实际情况，决定最终类别个数及各元素所属类别。

在本文中将采用欧氏距离计算个体元素间的距离，采用最短距离法计算类与类之间的距离。以下是对这两种距离的定义：

二阶闵科夫斯基距离为$d_{ij}(2)=\sqrt{\sum_{t=1}^{m}\left|x_{it}-x_{jt}\right|^2}$，$(i,j=1,2,\cdots,n)$称之为欧氏距离，本文即采用欧式距离计算元素间的距离。

类与类之间的距离定义为两类中相距最近的元素之间的距离，即类D_p和D_q之间的距离D_{pq}定义为$D_{pq}=\min_{i\in G_p,j\in G_q} d_{ij}$。当某步骤类$D_p$和$D_q$合并为$D_r$后，按最短距离法计算新类$D_r$和其他类$D_k$的类间距离。

2 主成分分析和聚类分析在高职学生成绩综合评价中的应用

将主成分分析与聚类分析方法应用于河南经贸职业学院2022-2023学年第一学期某专业班级学生期末各项成绩进行实证研究，对其作出综合评价。数据选取该专业班级16名同学的七门课程：英语、道德与法治、体育、高等数学、计算机基础、Linux基础、Python语言程序设计等作为变量，分别用$x_1,x_2,\cdots x_7$表示，用x_{ij}表示第i个学生在第j门课程上的得分，则$X=(x_{ij})_{16\times 7}$，这样就得到了一个16×7的原始数据阵，如表1所示。

表1 学生各指标取值原始数据

学生编号	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	78	66	65	90	86	81	78
2	87	61	94	74	74	100	70
3	74	72	85	69	61	81	62
4	68	60	85	65	66	61	67
5	72	70	98	60	69	75	60
6	79	72	95	76	65	72	70
7	82	60	83	88	85	81	87
8	84	66	89	80	65	86	86
9	77	65	89	92	87	89	89
10	94	72	91	94	97	100	96
11	77	60	75	93	82	99	84
12	74	64	65	67	68	62	61
13	68	67	77	70	71	81	60
14	80	70	94	71	66	94	66
15	88	76	65	88	85	93	91
16	83	68	86	72	66	81	82

调用SPSS，由相关阵出发进行主成分分析，输出结果如表2所示。

表2 主成分分析结果

Eigenvalues of the Correlation Matrix

Eigenvalue	Difference	Proportion	Cumulative
3.69031741	2.30097840	0.5272	0.5272
1.38933902	0.46229086	0.1985	0.7257
0.92704816	0.51746815	0.1324	0.8581
0.40958001	0.07218588	0.0585	0.9166
0.33739414	0.14956593	0.0482	0.9648
0.18782821	0.12933515	0.0268	0.9916
0.05849306	0.0084	1.0000	

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
x1	0.419253	0.340229	0.034975	0.152231	−.635131	−.460045	0.262757
x2	0.055904	0.514879	0.797267	0.046816	0.287391	0.106212	−.008496
x3	−.072182	0.653586	−.570735	0.299914	0.376646	0.010437	0.099788
x4	0.481331	−.220192	−.016388	0.059571	0.203117	0.417803	0.707250
x5	0.446031	−.245126	0.006308	0.030504	0.558488	−.631991	−.169343
x6	0.396969	0.285054	−.187550	−.802842	−.009786	0.179676	−.221472
x7	0.477162	−.046287	−.043810	0.485439	−.137445	0.413175	−.585661

由表2输出结果的特征值累积贡献率可知：前三个主成分的累积贡献率已达85%以上，因此只需用前三个主成分就能很好地概括这组数据，前三个主成分与标准化变量的关系为：

$$Y_1=0.4192x_1+0.0559x_2-0.0722x_3+0.4813x_4+0.4460x_5+0.3970x_6+0.4772x_7$$

$$Y_2=0.3402x_1+0.5149x_2+0.6536x_3-0.2202x_4-0.2451x_5+0.2851x_6-0.0463x_7$$

$$Y_3=0.0350x_1+0.7973x_2-0.5707x_3-0.0164x_4+0.0063x_5-0.1876x_6-0.0438x_7$$

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

综合评价函数：

$F = 0.5272Y_1 + 0.1985Y_2 + 0.1324Y_3$

代入数据计算学生各主成分得分如下表3所示，由每位学生的主成分得分代入综合评价函数，可以得到综合得分，按照综合得分大小排序，就得到每位学生的名次，由于排名是通过统计分析计算得到，其排名结果将减少主观因素，反映某些实际情况。

表3 学生各主成分得分及综合得分排名

学生	第1主成分		第2主成分		第3主成分		综合得分	
编号	得分	排名	得分	排名	得分	排名	得分	排名
1	1.0048	6	−1.7909	16	0.8394	3	0.2854	7
2	0.4562	7	0.8965	5	−1.6623	16	0.1984	8
3	−1.8025	13	0.8600	6	0.8196	4	−0.6711	12
4	−2.7114	16	−1.1833	12	−0.8307	13	−1.7744	16
5	−2.3537	15	1.1922	3	−0.0643	8	−1.0128	14
6	−1.0945	11	1.2232	2	−0.4299	10	−0.3911	11
7	1.3167	5	−1.1282	13	−1.0602	15	0.3298	5
8	0.4138	8	0.6750	8	−0.4738	11	0.2894	6
9	1.6334	4	−0.4383	10	−0.7294	12	0.6776	3
10	3.8215	1	1.1742	4	0.1791	6	2.2715	1
11	1.6376	3	−1.4359	14	−0.9470	14	0.4529	4
12	−2.2149	14	−1.7071	15	0.8755	2	−1.3906	15
13	−1.7768	12	−0.6510	11	0.4149	5	−1.0110	13
14	−0.6621	10	1.6099	1	−0.1502	9	−0.0494	10
15	2.5097	2	0.0112	9	2.2598	1	1.6246	2
16	−0.2213	9	0.6889	7	0.0997	7	0.0333	9

结果分析：第一主成分中x_1,x_4,x_5,x_6,x_7五个变量的系数比较大且都为正数，说明第一主成分主要由这五个变量决定，且与这五个变量的取值成正相关，从这五个变量分别所代表的课程（英语、高等数学、计算机基础、Linux基础、Python语言程序设计）来看，他们相对于其他两门课程（道德与法治、体育），是专业核心课，传统学分占比也较大，是反映高职学生技能水平和决定一位学生成绩的主要方面，得分较高，名次就较高，这与目前实际情况也是相符的。

之后，将每个学生的前三个主成分得分作为指标，数据输入SPSS系统进行聚类分析后结果为如图1所示。

依据图1结果可将全部学生分为四类，具体如表4所示。

从系统聚类结果可以看出，第一类学生的第一主成分得分都较高，说明他们专业核心课程学习成绩不错，但道德与法治、体育等课程成绩一般，有偏科现象，说明这类学生自身重视专业技能学习，不重视公共课程的学习；第二类学生大部分课程成绩中等，偶有课程成绩偏高或偏低，而且这类学生数量占比最大，属于学习积极性不高的一类。第三类学生各科成绩均较差，以后应注意培养学习兴趣，注重基础知识学习。第四类学生属于尖子生，各科成绩均较好，学习主动性高。

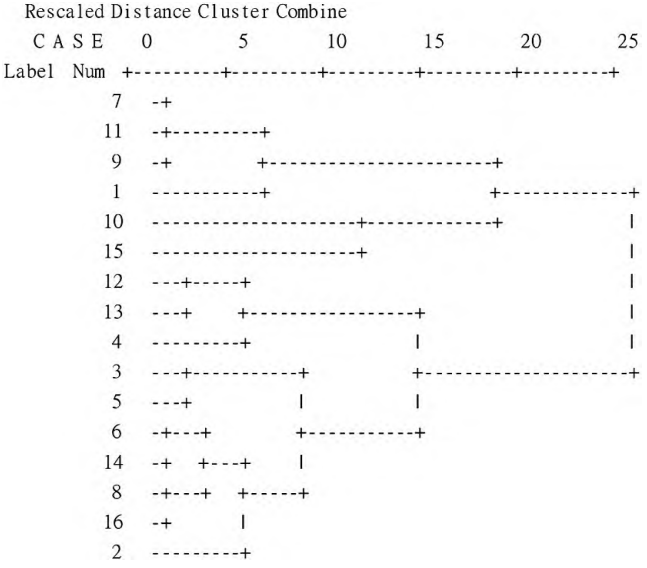

图1 谱系聚类图

表4 聚类结果

类别	学生个数	学生编号
1	4	9，11，7，1
2	7	8，2，3，16，14，6，5
3	3	13，12，4
4	2	10，15

3 结果启示

主成分分析法是一种多指标降维分析方法，当评价指标较多时，可以在保留绝大部分信息的情况下用少数几个综合指标代替原指标进行分析；在综合评价函数中，各主成分的权数是其贡献率，它反映了该主成分包含原始数据的信息量占总信息量的比重，这样确定权数是相对客观、合理的。通过对主成分得分进行聚类分析，能反映学生总体学习情况，各类型之间同学的发展特点及优劣势所在，为学生发展、教师教学、学校管理都提供了一些有价值的参考与启示[6]。

学生方面，学生应根据自身所属类别，提升自我认知，明确未来方向，如尖子生和核心课程成绩优异的学生下一步可考虑就业或继续升本深造；成绩一般的学生应找准方向，结合自身兴趣，重点学习提升一两门专业技能水平，为将来保底就业打下基础。

教师方面，需重视生源多样化的情况，应结合学生成绩评价结果去了解学生背景，尊重学生个体差

（下转第70页）

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表1 每日参与提交数据的人数

日期	参与人数
2022/1/7	80630
2022/1/8	101503
2022/1/9	91477
2022/1/10	107134
2022/1/11	153880

取S(0)=5 $\times 10^6$,I(0)=80630(为参考数据第一天的值),α_0 =1.2 $\times 10^{-6}$,β=1.4 $\times 10^{-7}$时,可以得出模型结果如图2所示。

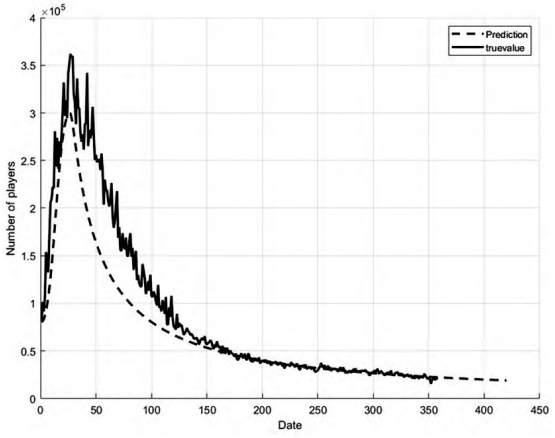

图2 模型预测数据实际数据对比图

改进后的SIR模型生成的预测数据与真实数据的误差均值为13.37%,特别是在网络信息传播的后期(从第160天以后直至结束),这一误差降低到5.46%,较为精准地对网络信息的感染人数进行了预测。

3 结束语

本文针对网络信息是如何进行传播的这一问题,提出了采用SIR传染病动力学模型进行分析研究的方法,并且根据网络信息传播的特性对SIR模型进行了一定的改良。得到的改进后的SIR模型可以较好的预测网络信息的传播情况。该模型可以在谣言舆情的控制、网络通信系统信道建设等方向发挥作用。下一步,作者准备考虑不同类型的网络信息是如何对"感染因子"产生影响的,从而得出更加准确地对网络信息的传播做出预测。

参考文献(References):

[1] 霍英,何碧姚,丘志敏,等.基于社会网络分析的网络信息传播规律研究[J].电脑知识与技术,2023,19(7):88-91.
[2] 崔增乐,钱晓东.区块链社交网络信息传播模型的优化研究[J].计算机工程与应用,2021,57(7):59-69.
[3] 张凌,罗曼曼,朱礼军,等.情感视角下的社交网络信息传播分析研究[J].图书情报知识,2019,188(2):111-119.
[4] 范伟,刘勇.基于时空Transformer的社交网络信息传播预测[J].计算机研究与发展,2022,59(8):1757-1769.
[5] 傅熙雯.在线社交网络信息传播建模及转发预测研究[J].科技与创新,2022,208(16):71-73.
[6] 方劲皓,钱晓东.改进的SCIR模型中社交网络信息传播研究[J].计算机工程与应用,2020,56(19):105-113.
[7] 赵剑华,万克文.基于信息传播模型—SIR传染病模型的社交网络舆情传播动力学模型研究[J].情报科学,2017,35(12):34-38.
[8] 李可嘉,王义康.改进SIR模型在社交网络信息传播中的应用[J].电子科技,2013,26(8):168-171.
[9] 宋英华,夏亚琼,方丹辉,等.基于改进SIR的正负面网络舆情信息传播模型与仿真[J].计算机应用研究,2021,38(11):3376-3381.

(上接第67页)

异,分类分层指导,因材施教,激发学生学习兴趣,帮助学生找到未来发展方向。

学校方面,应跟踪学生差异化的特点,可以实施分类培养,如就业方向、升学方向、创业方向等。对于就业方向,强化核心技能训练,创新实践教学体系,紧贴岗位实际需要;对于升学方向,加强专业课与公共通识课的整合衔接,全面提升学生科学素养;对于创业方向,加强创业指导,鼓励学生参加创新创业大赛,挖掘学生创新创业潜质,全面提升人才培养质量。

参考文献(References):

[1] 张琼.因子分析在学生成绩综合评价中的应用[J].惠州学院学报(自然科学版),2010,30(3):40-45.
[2] 尹芳.小麦RVA曲线拟合模型的建立及应用[D].扬州:扬州大学,2013.
[3] 高惠璇编著.应用多元统计分析[M].北京:北京大学出版社,2005.
[4] 张翎.主成分分析法在高校学生质量评价中的应用[J].云南民族学院学报(自然科学版),2001(1):284-286.
[5] 刘月河,王亚彬.基于主成分聚类分析的器材保管员考评问题研究[J].物流技术,2020,39(12):137-140.
[6] 孙文兴,王俊英,蒋励.因子分析法在高职院校学生成绩评价中的应用研究[J].昆明冶金高等专科学校学报,2019,35(1):101-106.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net