信息与计算科学专业的主成分模型分析
The Comparative Model Analysis on Education in Speciality of Information & Computing Science

朱 宁,符名培,李 雨
Zhu Ning,Fu Mingpei,Li Yu
(桂林电子工业学院计算科学与数学系,广西桂林 541004)
(Dept. of Comp. Sci. & Math. ,Guilin Institute of Elec. Tech. ,Guilin,Guangxi,541004,China)

摘要:利用比较学和多元统计方法,分析桂林电子工业学院"信息与计算科学"专业1999级和2000级学生的成绩,建立主成分模型,对两级学生的主成分模型进行分析和比较。结果表明,两级学生特点明显不同,其中2000级计算机能力较为突出,但重要的专业基础课却不如1999级,教学过程存在问题;该专业的课程设计和建设还需进行整合,课程开设的合理性有待论证。

关键词:教学管理 主成分模型 比较学 成绩
中图法分类号:G642.3;O212.4

Abstract: The students'scores of two classes in Guilin Institute of Elec. Tech. are analyzed by employing comparison theory and multivariate statistical analysis. Two principal component models for these two classes are established and compared. The analysis shows that the two classes'students are big different in study. One of two classes has higher computer skill but has less knowledge of base courses of the speciality.

Key words: education management, principal component model, comparison theory, scores

1998年教育部颁发了新的专业目录,将"信息与计算科学"列为一个新的数学类专业。桂林电子工业学院(以下简称本院)于1999年首次招收了"信息与计算科学"专业的本科生,首届学生已于2003年毕业,今年第二届学生即将毕业。对这样一个新的专业,在教学过程中不可避免地会出现许多问题,为更好地开办"信息与计算科学"专业,作者从本院教务科收集"信息与计算科学"2000级和1999级前三年的学生成绩,利用比较学和多元统计方法进行分析,得到了一些有用的信息。

1 数据采集

1.1 数据的收集与变量的选取

依据本院教学概览,将1999级和2000级"信息与计算科学"专业的前三年的部分学习成绩分为基础课和专业课两大类,共计21门课程。

基础课(X),包括 x_1:C语言程序设计A; x_2:常微分方程; x_3:复变函数A; x_4:概率与数理统计; x_5:解析几何; x_6:离散数学; x_7:数据结构与算法; x_8:数学分析(Ⅰ,Ⅱ,Ⅲ); x_9:大学英语(Ⅰ,Ⅱ,Ⅲ,Ⅳ); x_{10}:大学物理(Ⅰ,Ⅱ); x_{11}:高等代数(Ⅰ,Ⅱ); x_{12}:面向对象程序设计; x_{13}:数据库原理B; x_{14}:微机原理和应用; x_{15}:电路分析基础。

专业课(Y),包括 y_1:信号与系统; y_2:计算机网络原理; y_3:软件工程; y_4:数学模型A; y_5:数值分析(Ⅰ); y_6:最优化方法(英文教材)。

基础课和专业课的成绩数据来自本院教务科。

1.2 数据的预处理

为了使各门课程之间具有可比性,本文对每一门课程均进行标准化处理,并将同一门课程(如数学分析等)在几个学期中均有成绩的课程按学分绩合并成为一个成绩。

2 主成分模型分析

主成分分析方法是在"用较少的互不相关的新变量来反映原变量所提供的绝大部分信息"的思想下产生的处理高维数据的方法[1]。作者把 $\sum_{i=1}^{m}\lambda_i/\sum_{k=1}^{p}\lambda_k$ 称为前 m 个主成分的累计贡献率,它表明前 m 个主成分综合提供原变量中信息的能力。实

2004-07-23 收稿,2004-08-09 修回。

际应用中，通常选取 $m < p$，使前 m 个主成分的累计贡献率达到较高的比例(70%～80%)。这样用前面 m 个主成分代替原始变量，不但使变量的维数降低，而且也不致于损失原始变量中太多的信息。

将 1999 级和 2000 级两个年级的 21 门课程进行主成分分析，利用 sas 软件[2]可得到主成分的相关数据(见表 1)。

表 1　1999 级主成分分析数据

主成分 (z_i)	特征值 $(\hat{\lambda}_i)$	差　异	贡献率 $(\hat{\lambda}_i/\sum_{k=1}^{p}\hat{\lambda}_k)$	累积贡献率 $(\sum_{i=1}^{m}\hat{\lambda}_i/\sum_{k=1}^{p}\hat{\lambda}_k)$
1	12.8830626	11.3468009	0.6135	0.6135
2	1.5362617	0.5163739	0.0732	0.6866
3	1.0198878	0.1063115	0.0486	0.7352
4	0.9135763	0.2691074	0.0435	0.7787
5	0.6444689	0.0467672	0.0307	0.8094

由表 1 可知，前五个主成分的累计贡献率已达到 81%，且第 1 主成分的贡献率就已达到了 61.35%！第 1 主成分对第 2 主成分的散点图如图 1 所示。从图 1 中不难发现有 3 个“奇异点”，这 3 个点对应 3 个同学。将本年级的所有学生的第 1 主成分排序，发现它们在排序中载荷最低。经了解，这三个学生均已退学，他们在学习过程中的有些考试成绩为 0 值。为使分析问题具有连续性，我们将这 3 个学生不作为进一步研究的对象。1999 级也做了类似的处理，再次作主成分分析(见表 2)。

表 2　2000 级主成分分析数据

主成分 (z_i)	特征值 $(\hat{\lambda}_i)$	差　异	贡献率 $(\hat{\lambda}_i/\sum_{k=1}^{p}\hat{\lambda}_k)$	累积贡献率 $(\sum_{i=1}^{m}\hat{\lambda}_i/\sum_{k=1}^{p}\hat{\lambda}_k)$
1	8.04179492	6.24561981	0.3829	0.3829
2	1.79617511	0.45514273	0.0855	0.4685
3	1.34103238	0.20752885	0.0639	0.5323
4	1.13350353	0.14022057	0.0540	0.5863
5	0.99328296	0.08648085	0.0473	0.6336
6	0.90680211	0.08336840	0.0432	0.6768
7	0.82343371	0.03149244	0.0392	0.7160

从表 2 可以看出，只要选取前 7 个主成分便可使累计贡献率达到 71.6%，为此，再次做第 1 次主成分与第 2 主成分成分的散点图(图 2)。

从图 2 可知第 1 主成分对第 2 主成分的散点图的基本聚在一起，且不再有奇异点。

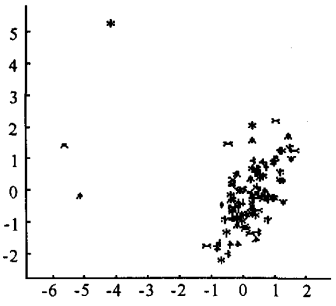


图 1　1999 级散点

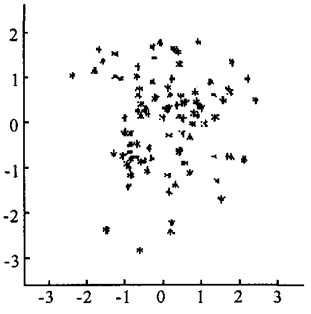


图 2　2000 级散点

2000 级主成分函数为：

$y_1 = 0.07x_1 + 0.27x_2 + 0.22x_3 + 0.28x_4 + 0.19x_5 + 0.23x_6 + 0.28x_7 + 0.3x_8 + 0.22x_9 + 0.22x_{10} + 0.3x_{11} + 0.18x_{12} + 0.24x_{13} + 0.19x_{14} + 0.24x_{15} + 0.22x_{16} + 0.14x_{17} + 0.17x_{18} + 0.12x_{19} + 0.14x_{20} + 0.2x_{21}$；

$y_2 = 0.07x_1 - 0.08x_2 - 0.03x_3 - 0.04x_4 - 0.3x_5 - 0.13x_6 + 0.1x_7 - 0.25x_8 - 0.04x_9 - 0.34x_{10} - 0.2x_{11} + 0.2x_{12} - 0.06x_{13} + 0.5x_{14} + 0.01x_{15} + 0.3x_{16} + 0.4x_{17} + 0.2x_{18} + 0.17x_{19} + 0.1x_{20} - 0.05x_{21}$；

$y_3 = 0.51x_1 - 0.04x_2 - 0.34x_3 - 0.02x_4 - 0.01x_5 - 0.08x_6 + 0.01x_7 - 0.01x_8 + 0.11x_9 + 0.09x_{10} + 0.03x_{11} + 0.04x_{12} + 0.01x_{13} + 0.06x_{14} + 0.2x_{15} + 0.1x_{16} - 0.33x_{17} - 0.33x_{18} + 0.52x_{19} + 0.07x_{20} - 0.2x_{21}$；

$y_4 = 0.58x_1 + 0.03x_2 - 0.14x_3 - 0.1x_4 - 0.06x_5 - 0.08x_6 - 0.07x_7 - 0.04x_8 - 0.18x_9 + 0.03x_{10} + 0.01x_{11} + 0.32x_{12} + 0.18x_{13} - 0.06x_{14} + 0.15x_{15} - 0.18x_{16} - 0.01x_{17} + 0.34x_{18} - 0.35x_{19} - 0.29x_{20} + 0.26x_{21}$；

$y_5 = 0.03x_1 + 0.13x_2 - 0.002x_3 - 0.14x_4 - 0.12x_5 - 0.4x_6 + 0.01x_7 - 0.1x_8 + 0.38x_9 + 0.006x_{10} + 0.02x_{11} - 0.08x_{12} - 0.06x_{13} - 0.1x_{14} - 0.1x_{15} - 0.1x_{16} - 0.24x_{17} + 0.31x_{18} - 0.03x_{19} + 0.6x_{20} + 0.26x_{21}$；

$y_6 = 0.06x_1 - 0.05x_2 - 0.004x_3 - 0.24x_4 + 0.28x_5 + 0.31x_6 - 0.22x_7 + 0.1x_8 + 0.11x_9 - 0.22x_{10} + 0.04x_{11} + 0.56x_{12} - 0.22x_{13} - 0.02x_{14}$

$-0.13x_{15}-0.1x_{16}+0.12x_{17}-0.29x_{18}-0.14x_{19}+0.36x_{20}+0.007x_{21}$;

$y_7=0.31x_1-0.1x_2-0.03x_3+0.05x_4-0.15x_5-0.02x_6+0.26x_7+0.1x_8-0.01x_9+0.25x_{10}+0.1x_{11}-0.11x_{12}+0.24x_{13}-0.05x_{14}-0.44x_{15}-0.06x_{16}+0.3x_{17}-0.08x_{18}-0.28x_{19}+0.26x_{20}-0.46x_{21}$.

从第1主成分来看,各分量的载荷虽然有些差异,但并不明显,且均为正载荷。说明第1主成分能较均匀地提供各个原始变量的信息,这里可称其为"综合能力因子"。

从第2主成分中,发现 x_{14},x_{16},x_{17} 具有较高的正载荷,而 x_5,x_{10} 则相对地具有较高的负载荷。所以称此为"计算机硬件能力因子"。

第3主成分中,x_1 和 x_{19} 具有极高的正载荷(0.51),所以称此因子为"编程能力因子"。观察第4主成分,发现,x_1,x_{12} 和 x_{18} 具有较高的正载荷,尤其以 x_1(C语言)为最高,而 x_{12},即面相对象程序设计,恰恰就是 x_1 的后继课程。这样可以称此因子为"软件开发能力因子"。

观察第5主成分,发现 x_9,x_{18} 和 x_{20} 有较高的正载荷,这几门课程的共同特性就是,要求同学们具有较强的记忆能力。而在这一主成分中恰恰有 x_6(离散数学)为较大的负载荷。所以称其为"记忆能力因子"。

第7主成分,x_1,x_7 和 x_{17} 具有较高的正载荷,称其为"计算机网络系统集成能力因子"。

从表3可以看出,前4个主成分已经达到70.14%。

表3 1999级主成分分析数据

主成分 (z_i)	特征值 ($\hat{\lambda}_i$)	差 异	贡献率 ($\hat{\lambda}_i/\sum_{k=1}^{p}\hat{\lambda}_k$)	累积贡献率 ($\sum_{i=1}^{m}\hat{\lambda}_i/\sum_{k=1}^{p}\hat{\lambda}_k$)
1	11.1261578	9.6988778	0.5298	0.5298
2	1.4272800	0.3154464	0.0680	0.5978
3	1.1118336	0.0473031	0.0529	0.6507
4	1.0645305	0.2689548	0.0507	0.7014

1999级主成分函数:

$y_1=0.17x_1+0.24x_2+0.24x_3+0.24x_4+0.19x_5+0.22x_6+0.23x_7+0.26x_8+0.23x_9+0.23x_{10}+0.23x_{11}+0.21x_{12}+0.21x_{13}+0.25x_{14}+0.2x_{15}+0.22x_{16}+0.16x_{17}+0.2x_{18}+0.11x_{19}+0.23x_{20}+0.25x_{21}$;

$y_2=0.36x_1-0.15x_2-0.13x_3+0.07x_4-0.05x_5+0.2x_6+0.002x_7+0.03x_8+0.01x_9+0.39x_{10}+0.37x_{11}-0.1x_{12}-0.04x_{13}-0.29x_{14}+0.3x_{15}-0.1x_{16}-0.3x_{17}-0.4x_{18}+0.13x_{19}-0.2x_{20}-0.03x_{21}$;

$y_3=0.12x_1+0.001x_2+0.04x_3-0.02x_4-0.38x_5-0.18x_6+0.14x_7-0.09x_8+0.11x_9-0.003x_{10}-0.002x_{11}-0.27x_{12}-0.15x_{13}-0.12x_{14}-0.08x_{15}+0.04x_{16}+0.41x_{17}-0.04x_{18}+0.67x_{19}+0.17x_{20}+0.01x_{21}$;

$y_4=-0.13x_1+0.06x_2-0.07x_3-0.28x_4+0.29x_5-0.0001x_6-0.08x_7-0.19x_8-0.19x_9+0.28x_{10}+0.29x_{11}+0.11x_{12}-0.54x_{13}+0.08x_{14}+0.23x_{15}-0.05x_{16}+0.26x_{17}+0.31x_{18}+0.04x_{19}-0.008x_{20}-0.2x_{21}$.

1999级的第1主成分类似于2000级的第1主成分,其第1主成分的个分量均具有相近的正载荷。说明第1主成分较为均匀的提供了各原始变量的信息,所以亦称其为"综合能力因子"。

经过观察发现,在第2主成分中,x_1,x_{10} 和 x_{11} 等基础课程的变量具有较高的正载荷,而诸如 x_{17} 和 x_{18} 这样的专业课变量却具有较高的负载荷,所以称其为"基础能力因子"。

第3主成分中的 x_{17} 和 x_{19} 正载荷较大,而 x_{17},即计算机网络,同学们只是懂得组建网络的方法等能力,而 x_{19},即数学建模,恰恰是锻炼同学们的建立模型的一种能力。从这一个角度上讲,可称其为"网络设计能力因子"。

第4主成分有2个较大载荷的变量,即 x_{13} 和 x_{18}. 不同的是,x_{13} 即数据库,具有较高的负载荷,而软件工程却具有较高的正载荷。这样,数据库和软件工程可使同学学会很多的方法和能力,由于它们是一个为正,一个为负,这里称其为"记忆能力因子"。

3 分析与讨论

将1999级和2000级同学的学习成绩的主成分作比较,可得到表4。从表4不难看出,1999级数学基础比2000级显著,而2000级的计算机能力和数学应用能力比1999级要突出。2000级的编程能力和软件开发能力较为突出,这一点在2000级的计算机二级一次通过率、计算机三级、四级、中级程序员

表4 1999级和2000级主成分分析

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分	第6主成分	第7主成分
1999级	综合因子	基础能力因子	网络设计能力因子	记忆能力因子			
2000级	综合因子	计算机硬件	编程能力因子	软件开发能力因子	记忆能力因子	空间逻辑能力因子	计算机网络系统集成能力因子

(下转第246页)

5 结束语

数据仓库作为数据库领域发展起来的一种新型的技术，它在数据的管理和使用上与传统数据库有着本质的不同，数据仓库是大量集成化数据的集合，是多种技术的综合体。OLAP提供给数据仓库系统一种高灵活性、高性能地存取、浏览和分析数据的手段，它的分析结果可以为数据挖掘提供挖掘依据，而数据挖掘又可以拓展OLAP分析的深度，可以发现OLAP不能发现的更为复杂的信息，因此，将OLAP与数据挖掘相结合将会在数据仓库中发挥更好的效用，这也是OLAP发展的一个新的方向。

参考文献：

1 Inmon W H. What is a data warehouse? http://www.billinmon.com,2000.

2 马宏鹏，赵 新，李 明，等. 数据仓库原型系统设计. 计算机工程与应用，2000，11：109～111.

3 彭木根. 数据仓库技术与实现. 北京：电子工业出版社，2002.6.

4 柳 莺，赵艳红，钱 旭，等. 数据仓库技术研究和应用探讨. 计算机应用，2001，2(2)：46～48.

5 Inmon W H. A Data Warehouse/OLAP Framework for Web Usage Mining and Business Intelligence Reporting. http://www.billinmon.com,2002.

6 张 旭，董有田. OLAP多维数据分析与应用研究. 黑龙江科技学院学报，2002，9(3)：16～19.

7 Inmon W H. OLAP and Data Warehouse. http://www.billinmon.com,2002.

8 Inmon W H[美]著. 构建数据仓库. 王志海，林有芳，等译. 北京：机械工业出版社，2003.3.

9 戴超凡，邓 苏，黄宏斌，等. DSS中数据管理新技术研究. 计算机工程与应用，2000，12：21～24.

（责任编辑：黎贞崇）

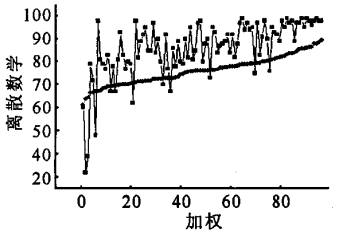
（上接第242页）

以及高级程序员的通过率可以看出，其中高级程序员的通过率占全院的40%左右。从这个比率看，2000级的软件开发能力和编程能力非常突出。

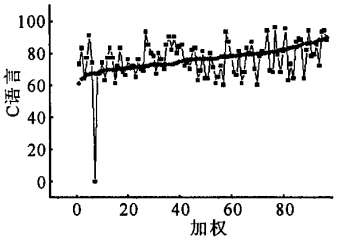
2000级和1999级的一些重要专业基础课程未能出现在前几个主成分中，为此，针对离散数学和C语言两门专业基础课进行分析，以加权主成分排序作为自变量，专业课成绩作为因变量作散点图（图3）.

从图3可以看出，两课程的成绩不能很好地区分按加权主成分排序的学生的学习能力，这也说明，两课程的成绩不能有效地区分学生的学习能力。

通过表3的分析，可认为我系新办专业的课程设计和建设还需要进行整合，课程的开设的合理性



(a)



(b)

图3 散点图

(a)离散数学；(b)C语言

还有待论证，课程成绩中反映出的问题需要认真分析、解决。

参考文献：

1 范金城，梅长林. 数据分析. 北京：科学出版社，2002.

2 樊 欣，邵谦谦. SAS8.X经济统计. 北京：北京希望电子出版社，2003.

3 朱 宁，符名培，方进一. 教学研究中的主成分模型. 桂林：桂林电子工业学院学报，2004，(2)：97～99.

（责任编辑：黎贞崇）