

文章编号:1007-6735(2013)02-0193-04

高校信息化部分基础设施建设的实证分析

陈巍巍¹, 张雷¹, 陈世平², 刘秋龄¹

(1. 上海理工大学 管理学院, 上海 200093; 2. 上海理工大学 信息化办公室, 上海 200093)

摘要: 以我国教育信息化建设与应用专题研究报告的部分数据为样本, 针对我国高校信息化基础设施建设的抽样调查情况进行主成分分析、图形聚类分析和系统聚类分析. 结果表明, 我国高校信息化基本设备配置存在着不平衡现象. 东部沿海地区在信息化基础设施建设方面要远远优于中西部地区, 各地区应针对各自不同特点有重点地加以改善.

关键词: 实证分析; 聚类分析; 主成分分析; 信息化

中图分类号: TP 399 **文献标志码:** A

Empirical Analysis on Infrastructure Construction of ICT Using in University

CHEN Weiwei¹, ZHANG Lei¹, CHEN Shiping², LIU Qiuling¹

(1. Business School, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. Informatization Office, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Using the data from the special report of education informatization construction and application in our country, principal component analysis (PCA), graphics cluster analysis and system cluster analysis on the infrastructure construction of information and communication (ICT) were carried out. The study shows that there is a kind of imbalance in the informatization configuration of China. It is much better in the eastern coastal areas than in the central and western regions. Thus in different areas, different emphasis should be put in improving the informatization equipment in school.

Key words: empirical analysis; cluster analysis; principal component analysis (PCA); information and communication technology (ICT)

信息化建设是一个新兴的课题, 高校信息化也是如此. 其建设内容包括基础设施建设、资源建设、应用系统建设、标准规范建设等. 聚类分析作为一种基于相似性的子群划分方法, 可先将调研地区进行分类,

再根据分类情况对各类地区的信息化基础设施建设情况提出建议和指导, 可有效地避免建设的不均衡和盲目性. 按分类对象的不同, 聚类分析可分为 R 型和 Q 型两大类, R 型聚类分析用于指标的分类处理, Q 型

收稿日期: 2012-06-03

基金项目: 上海市重点学科建设资助项目(S30504)

第一作者: 陈巍巍(1977-), 女, 博士研究生. 研究方向: 高校信息化. E-mail: chenww@usst.edu.cn

通讯作者: 张雷(1952-), 男, 教授. 研究方向: 工业工程. E-mail: zhangl@usst.edu.cn

聚类分析则用于样品的分类处理^[1]. 在聚类的过程中,一方面,为度量样本间的相似性会采用如 Euclidean 距离、Manhattan 距离、Minkowski 距离等. 如果将距离的计算过程看成黑盒,那么这一过程就是从多维空间到一维空间的过程;另一方面,距离阈值 d_0 该如何确定也是一件困难的事,且在很大程度上决定了簇类的数量和大小. 从以上两方面来考虑,笔者试图先降维作特征提取,然后再进行比较,得到各个簇类. 比较常用的特征提取方法主要有主成分分析(PCA)^[2]和线性鉴别分析(linear discriminant analysis, LDA)^[3]. PCA 基于这样一种思想:方差最大的方向包含最多的类间信息,而 LDA 则是在最大化类间信息的同时最小化类内信息,并最大化两者之比. 然而,降维必定会导致信息丢失,而且这些具有较高复杂度的方法在动态、多变的环境中并无优势^[4]. 因此,本文根据调研数据情况直接应用 Q 型聚类的方法,同时根据主成分降维聚类的结果对前面的聚类结果加以修正. 以我国部分基础设施建设的数据为研究对象,通过对部分数据的分析和研究,以小见大,试图探索和总结高校整体信息化建设情况的分析方法与手段,并为如何更好地发挥信息化的作用提供理论参考.

1 数据来源与构建

2008年2月22日,教育部教育改革发展与战略与政策研究重大课题“教育信息化建设与应用研究”启动工作在教育部展开. 第一工作组是“教育信息化建设与应用状况调研组”,负责调研教育信息化建设与应用现状. 本次调研区域包含:北京市、上海市、广东省、湖南省、湖北省、陕西省、甘肃省、内蒙古自治区、云南省、吉林省 10 个省、市、自治区. 调研范围为这些区域的高校,包括高等职业技术学院^[5].

信息化基础设施是信息化建设的基础条件,也是信息化的“命脉”. 从一个社会组织的内部机构到一个组织的所有机构,从一个地区到一个国家,只有有了一条条畅通无阻的“信息高速公路”,形成了一个信息通信的网络,才有可能实现各种信息化应用. 我国高校信息化基础设施的建设内容主要包含信息化基本设备配置、校园网建设、网络与信息安全建设、高性能计算环境建设四方面内容. 本文由于可用的样本数据有限,主要从信息化基本设备配置数据着手进行研究和分析.

信息化设备包含服务器、个人计算机、交换机、路由器、扫描仪、录像机、投影机等. 调研数据主要从

服务器、个人计算机、多媒体教室的拥有现状三方面考察. 服务器主要从价值方面来考察,统计计入固定资产的服务器总价值及 10 万元以上的服务器的台数. 个人计算机主要从计入固定资产的个人计算机台数(包含笔记本电脑)及学生拥有个人计算机比例两方面考察. 多媒体教室主要是考察配备多媒体设备(如投影仪)的教室占教室总数的比例及多媒体教室的利用率. 综上,得信息化基本设备配置的 6 个指标 $X_1, X_2, \dots, X_6$, 分别表示服务器的总价值/万元、10 万以上服务器的数量/台、个人计算机数量/台、拥有个人计算机数量的比例、多媒体教室配备比例和多媒体教室的利用率. 具体数据是以省和直辖市为单位,求出其所辖学校总体数据的平均值.

2 模型方法概述

2.1 主成分分析

用于数据分析的样品往往包含有多个(间隔)变量,较多的变量会带来分析问题的复杂性. 然而,这些变量彼此之间常常存在着一定程度的、有时甚至是相当高的相关性,这就使含在观测数据中的信息在一定程度上有所重叠. 正是这种变量间信息的重叠,使得变量的降维成为可能,从而使问题的分析得以简化^[6].

主成分分析是设法将原来众多具有一定相关性(比如 P 个指标),重新组合成一组新的互相无关的综合指标来代替原来的指标. 通常数学上的处理就是将原来 P 个指标作线性组合,作为新的综合指标. 最经典的做法就是用 F_1 (选取的第一个线性组合,即第一个综合指标)的方差来表达,即 $\text{var}(F_1)$ 越大,表示 F_1 包含的信息越多. 因此在所有的线性组合中选取的 F_1 应该是方差最大的,故称 F_1 为第一主成分. 如果 F_1 不足以代表原来 P 个指标的信息,再考虑选取 F_2 ,即选第二个线性组合. 为了有效地反映原来信息, F_1 已有的信息就不需要再出现在 F_2 中,用数学语言表达就是要求 $\text{cov}(F_1, F_2) = 0$,则称 F_2 为第二主成分,依此类推可以构造出第三、第四,……,第 P 个主成分.

2.2 系统聚类

聚类分析至今,有许多种不同的聚类方法. 其中应用得最多最成熟的方法为系统聚类法,也是本文将采用的方法. 其思路为首先将每个数据对象各视为一类,根据类与类之间的距离或相似程度将最相似的类加以合并,再计算新类与其它类之间的相似

程度,并选择最相似的类加以合并,这样每合并一次就减少一类,不断继续这一过程,直到所有数据对象合并为一类为止^[7].

定义类与类之间的距离的方法有很多,应用欧几里得距离(Euclidean distance),即

$$d_{ij} = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{1/2}$$

由于纳入分析的各变量方差相差太大时,变异度的差异会影响结果的正确性,所以一定要进行变量的标准化.类的个数可以综合系统聚类的结果和前面主成分分析的图形聚类分析共同确定.

3 实证分析

3.1 分析结果

根据主成分分析的思想,应用 SPSS 16.0 软件作为统计分析工具,首先将原始数据做标准正态变换(Z-Scores),消除数量级和量纲差异的影响^[8],其次建立相关系数矩阵如表 1 所示.

表 1 相关矩阵
Tab.1 Correlation matrix

	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆
X ₁	1.000	0.925	0.766	0.774	0.684	-0.610
X ₂	0.925	1.000	0.811	0.799	0.738	-0.400
X ₃	0.766	0.811	1.000	0.691	0.665	-0.346
X ₄	0.774	0.799	0.691	1.000	0.908	-0.474
X ₅	0.684	0.738	0.665	0.908	1.000	-0.351
X ₆	-0.610	-0.400	-0.346	-0.474	-0.351	1.000

通过数据处理得相关系数矩阵,从矩阵可看出指标间存在较强的相关性.进而对原始数据进行因子分析的可行性检验(KMO 值和巴特利特球形检验).KMO 的值为 0.760>0.6;而巴特利特球形检验的显著性水平小于 0.05,因此拒绝巴特利特球形检验的零假设.以上种种分析结果都表明样本适合主成分分析.进而求得主成分分析的方差累计贡献率如表 2 所示.

表 2 特征值与贡献率^[9]
Tab.2 Eigenvalue and contribution

特征值序号	特征值	贡献率/%	累积贡献率/%
1	4.391	73.187	73.187
2	0.786	13.099	86.287
3	0.474	7.897	94.184
4	0.232	3.871	98.055
5	0.077	1.289	99.344
6	0.039	0.656	100.000

可见得到的前两个因子提取了原始数据的 86.287%的数据信息,因此可以提取前两个因子作为主成分.同时也要注意,第一个主成分提取了 73.187%的数据信息,相对于第二个主成分,它占有绝大多数的比例.而且从第二个主成分的主要数据源 X₆(根据表 3)的原始数据分析,各省市的多媒体教室利用率的数值差异并不大(最小为 86.58%,最大为 94.64%).通过以上分析本文认为,从反映原始数据的角度来说第一主成分是更加重要和需要首要考虑的.

两个主成分与原始变量指标之间的关联程度由因子载荷值来体现,如表 3 所示.从表中可以看出第一主成分在前 5 个指标上具有较大的载荷值,第二主成分在第 6 个指标上具有较大的载荷值.联系实际指标意义可知,第一个主成分主要反映的是信息化设施配置水平,第二个主成分主要反映的是信息化设施的利用情况(这里主要是考察多媒体教室).显然这两个主成分是互为消涨的.

表 3 成分的因子载荷和得分系数表
Tab.3 Component's factor loading and factor score

	因子载荷值		得分的系数值	
	成分 1	成分 2	成分 1	成分 2
X ₁	0.935	-0.135	0.213	-0.172
X ₂	0.931	0.136	0.212	0.173
X ₃	0.850	0.197	0.194	0.251
X ₄	0.918	0.100	0.209	0.127
X ₅	0.864	0.240	0.197	0.306
X ₆	-0.582	0.802	-0.133	1.020

为了更直观的查看考察对象的两个主成分得分情况,采用线性回归的方法得到了成分得分的系数矩阵如表 3 所示.由此可得到主成分得分的表达式为

$$F_1 = 0.213x_1 + 0.212x_2 + 0.194x_3 + 0.209x_4 + 0.197x_5 - 0.133x_6$$
$$F_2 = -0.172x_1 + 0.173x_2 + 0.251x_3 + 0.127x_4 + 0.306x_5 + 1.020x_6$$

将各省、直辖市的标准化后的数据带入上式可计算出它们两个主成分的得分,进而可以画出如图 1(见下页)所示的散点图,由于第一主成分比较重要,故意将横轴(第一主成分)的单位长度适当加大,以方便观察.

从图 1 看到北京的数据在基本设备配置水平方面远远高于其它省份,上海和广东省分居其后.吉林

省和上海市在信息化设施的利用率方面相对较低,说明在现有的教学模式下,信息化设施基本满足使用要求.

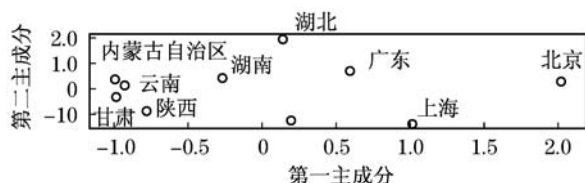


图1 主成分散点图

Fig.1 Main components scatter chart

3.2 聚类分析结果

应用 SPSS 软件的系统聚类方法,得到图 2 的聚类结果.

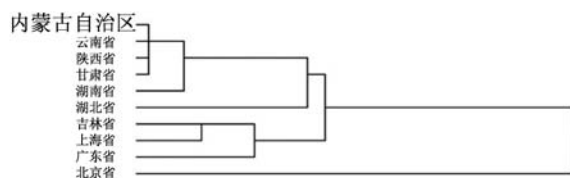


图2 系统聚类图

Fig.2 System cluster chart

结合图 1 的情况考察对象划分为 4 类比较合理.第一类为第一主成分较低的省市,分别为内蒙古自治区、云南省、甘肃省、陕西省和湖南省,这些省份应该首要考虑提高辖内高校的信息化基本设备的配置水平,其次要考虑转变教学模式,灵活运用多媒体等信息化手段来改善教学效果.第二类为第二主成分比较高的湖北省,由于信息化基础设施的利用率相对较高,说明该省内高校信息化设备的利用处于饱和状态,所以进一步提高信息化设施的配置是当务之急.第三类为第一主成分相对较高,第二主成分相对较低的吉林省、广东省和上海市,它们所辖的高校现有信息化基本设备配置基本满足要求.但是从图 1 中注意到吉林省和上海市的信息化基础设施的利用率偏低,应进一步关注其高校的教学方式,及利用信息化先进技术的情况.最后一类为北京市,其所属高校无论在信息化基本设备配置的水平上,还是在信息化基础设施的利用上都表现较好.

4 结论和对策

通过上述分析,得出我国高校的信息化基本设备配置方面存在着不平衡表现的结论.东部沿海地区在信息化基础设施建设方面要远远优于中西部地区,这不利于我国高等教育的均衡发展.国家应当考虑从多方面扶持内陆贫困地区的高等教育信息化基本设备配置建设.从调研的 10 个省市来看,北京市的信息化基础设施建设最好,内蒙古自治区、云南省、甘肃省、陕西省和湖南省的基础设施建设还有待改善.从信息化基础设施的利用率来看,吉林省和上海市的利用率较低,可以进一步考察其在教学方式方法上是否有待改善.

参考文献:

- [1] 杨小平. 统计分析方法与 SPSS 应用教程[M]. 北京:清华大学出版社,2008:227-228.
- [2] Jolliffe I T. Pricipal component analysis [M]. New York:Springer-Verlag,1986.
- [3] Wang X C, Paliwal K K. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition[J]. Pattern Recognition, 2003,36(10):2429-2439.
- [4] 陈清华,李林锦,翁正秋. 基于新聚类算法的推荐系统的研究与实现[J]. 电脑知识与技术,2010,6(6):1523-1525.
- [5] “教育信息化建设与应用研究”课题组. 我国教育信息化建设与应用专题研究报告[M]. 北京:高等教育出版社,2010.
- [6] 王学民. 统计分析方法及应用[M]. 上海:上海财经大学出版社,2010.
- [7] 唐敏,陈道平. 基于因子分析和聚类分析的重庆市经济发展状况研究[J]. 现代商贸工业,2008(9):130-131.
- [8] 张文彤. SPSS11 统计分析教程[M]. 北京:北京希望电子出版社,2002.
- [9] 柯兵,钱省三. 聚类分析和因子分析在股票研究中的应用[J]. 上海理工大学学报,2002,24(4):371-373.

(编辑:金虹)