

# 考试数据分析及孤立点检测的谱聚类方法

贾志先

(新疆财经大学 网络与实验教学中心, 新疆 乌鲁木齐 830012)

**摘要:**孤立点(outlier)又称离群点,为位于远离与之相应的随机变量平均值的点。针对考试数据分析和孤立点检测问题,给出了答卷数据的谱聚类算法。利用答卷数据的谱聚类算法对所有考生的答卷数据进行聚类后,根据答卷数据的谱聚类算法中引入的距离矩阵,以每一位被试的答卷数据为参照,构成一个分类图,从而构建一个分类图簇。谱聚类算法在分析考试数据和检测孤立点方面具有明显的优势。实验表明,利用答卷数据的谱聚类算法对考试答卷结果进行聚类和分析,可以从中挖掘出更多有价值的信息。

**关键词:**谱聚类;特征值;特征向量;孤立点检测

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2013)01-0103-04

doi:10.3969/j.issn.1673-629X.2013.01.026

## Spectral Clustering Method for Exam Data Analysis and Outlier Detection

JIA Zhi-xian

(Network and Experimental Teaching Center, Xinjiang University of Finance and Economics,  
Urumqi 830012, China)

**Abstract:** An outlier is defined as a point that lies very far from the mean of the corresponding random variable. For the problem of exam data analysis and outlier detection, give the algorithm of spectral clustering on exam data analysis. After clustering of all the answers of people who take the exam, by using spectral clustering algorithm of examination scripts, according to the introduction of the distance matrix in spectral clustering algorithm, each people who takes the exam can be treated as a reference, to constitute a classification chart, thus build a cluster of classification chart. The algorithm of spectral clustering has obvious advantages in the analysis of exam data and outlier detection. Experiments show that clustering and analysing the results of exam answers by using of spectral clustering, can dig out the more valuable information.

**Key words:** spectral clustering; eigenvalue; eigenvector; outlier detection

## 0 引言

孤立点(outlier)检测可用于发现不具备数据一般特性的数据对象<sup>[1]</sup>。人们已将孤立点检测应用于信用卡欺诈探测、收入极高或极低的客户分区、医疗分析等领域。

到目前为止,孤立点检测的算法有基于统计的孤立点检测、基于距离的孤立点检测、基于偏离的孤立点检测和基于聚类的孤立点检测等<sup>[2~8]</sup>。

可将孤立点检测应用到心理与教育测量中。在心理与教育测量过程中,由于测试数据中孤立点的存在,

有可能产生较大的误差。检测和剔除孤立点,可以提高测验结果的可信度。

文中以MHK(中国少数民族汉语水平等级考试)考试数据为研究对象,探讨基于谱聚类的答卷数据分析以及答卷数据中孤立点的检测方法。

## 1 孤立点

定义:孤立点又称离群点,为位于远离与之相应的随机变量平均值的点<sup>[1,9]</sup>,是与数据的其他部分不同的数据对象。

例如,在考试测试中,个别被试的答卷结果可能与标准答案及大部分被试的答卷结果相差较大。从被试能力的角度来看,这些被试的水平比较差,考试过程中随意填涂了一些答案。这些被试的答卷结果可以看成答卷数据中的孤立点。

收稿日期:2012-05-11;修回日期:2012-08-15

基金项目:全国教育科学规划项目(FFB108172);新疆自治区高校科研计划重点项目(XJEDU2010149)

作者简介:贾志先(1958-),男,山西临猗人,教授,主要研究方向为人工智能、模式识别、算法。

## 2 谱聚类算法

聚类是按照一定的规律和要求对事物进行区分和分类的过程,在这一过程中没有任何关于分类的先验知识,仅靠事物间的相似性作为类属划分的准则<sup>[10,11]</sup>。

与传统的聚类算法相比,谱聚类具有能够在任意形状的样本空间上聚类,且收敛于全局最优解的优点<sup>[12,13]</sup>。

谱聚类算法可简要描述为<sup>[12,14,15]</sup>:

给定一个数据集  $X = \{x_1, x_2, \dots, x_n\}$ ,  $x_i \in \mathbb{R}^N$ 。根据数据集  $X$  建立加权图  $G = (V, E)$ 。其中  $V = \{v_i, i = 1, 2, \dots, n\}$  是顶点的集合,  $E = \{e_{ij}\}$  是连接顶点  $(v_i, v_j)$  的边。图  $G$  中每一个节点  $v_i$  与数据集  $X$  中的  $x_i$  相关。

采用一个相似度准则构造图  $G$  的顶点之间的相似度矩阵 (similarity matrix)  $W$ ,  $W \in \mathbb{R}^{n \times n}$ 。常用的相似度用高斯核函数来表示<sup>[9,16]</sup>,即对于两个对象  $x_i$  和  $x_j$ , 其相似度  $s$  为

$$s(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2) \quad (1)$$

输入:图  $G$  的相似度矩阵  $W$ , 聚类个数  $k$ 。

步骤 1 计算 Laplacian 矩阵  $L$ 。根据相似度矩阵  $W$  建立一个对角矩阵,记为  $D = (d_{ij})_{n \times n}$ ,  $d_{ii} = \sum_{j=1}^n w_{ij}$ , 令 Laplacian 矩阵  $L = D - W$ 。

步骤 2 计算 Laplacian 矩阵  $L$  的特征值和特征向量,选择前  $k$  个特征向量  $u_1, u_2, \dots, u_k$ , 构建以特征向量  $u_1, u_2, \dots, u_k$  为列的矩阵  $U \in \mathbb{R}^{n \times k}$ 。

步骤 3 设  $y_i \in \mathbb{R}^k (i = 1, 2, \dots, n)$  是对应  $U$  的第  $i$  行的向量。用  $k$  均值聚类算法,将向量  $y_i$  聚类到  $C_1, C_2, \dots, C_k$ 。

步骤 4 建立映射

$$x_i \in \mathbb{R}^N \mapsto y_i \in \mathbb{R}^k, i = 1, 2, \dots, n。$$

根据  $x_i$  和  $y_i$  之间的对应关系,利用步骤 3 中矩阵  $U$  的行向量  $y_i$  的聚类结果,确定点  $x_1, x_2, \dots, x_n$  在聚类中的结果。

输出: $x_1, x_2, \dots, x_n$  的聚类结果。

## 3 答卷数据的谱聚类算法

根据答卷数据对象的特点,在谱聚类算法基础上,增加了答卷数据对象的距离矩阵,可得到适合答卷数据分析的谱聚类算法。

### 3.1 答卷数据对象之间的距离

假设研究的答卷数据对象  $x$  有  $p$  个测试点,表示为  $(x_1, x_2, \dots, x_p)$ ,两个被试的答卷数据对象  $u$  和  $v$  之间的距离  $d(u, v)$  可定义为

$$d(u, v) = \sum_{i=1}^p |u_i - v_i| \quad (2)$$

$$\text{其中, } |u_i - v_i| = \begin{cases} 1 & u_i \neq v_i \\ 0 & u_i = v_i \end{cases}, i = 1, 2, \dots, p$$

### 3.2 答卷数据对象之间的相似度

根据答卷数据对象之间的距离,两个被试的答卷数据对象  $u$  和  $v$  之间的相似度  $s(u, v)$  可用下面的核函数

$$s(u, v) = \exp(-d(u, v) / 2\sigma^2) \quad (3)$$

表示。通常情况下,  $\sigma = 1$ 。

### 3.3 答卷数据的谱聚类算法

假定被试的答卷经过扫描仪扫描后,存储在一个数据库中。利用下面的答卷数据的谱聚类算法对答卷数据进行分析。

答卷数据的谱聚类算法为:

步骤 1 读入考试扫描数据库记录。答卷数据对象集合表示为  $X = \{x_1, x_2, \dots, x_n\}$ ,  $x_i \in \mathbb{R}^p$ 。其中,  $x_i$  为第  $i$  个被试的答题信息,  $p$  为试题数。根据 (2) 式中的答卷数据对象之间的距离公式,计算被试的答卷数据对象的距离矩阵  $A$ 。距离矩阵  $A$  是一个  $n$  阶的对称矩阵。

步骤 2 根据 (3) 式中的答卷数据对象之间的相似度公式,计算被试的答卷数据对象的相似度矩阵  $W$ , 矩阵  $W$  的阶数与矩阵  $A$  的阶数相同。

步骤 3 计算扩展邻接矩阵 (scaled adjacency matrix)  $L$ 。根据相似度矩阵  $W$  建立一个对角矩阵,记为  $D = (d_{ij})_{n \times n}$ ,  $d_{ii} = \sum_{j=1}^n w_{ij}$ , 令矩阵  $L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ 。

步骤 4 计算矩阵  $L$  的特征值和特征向量,选择前  $k$  个特征向量  $u_1, u_2, \dots, u_k$ , 构建以特征向量  $u_1, u_2, \dots, u_k$  为列的矩阵  $U \in \mathbb{R}^{n \times k}$ , 并对矩阵  $U$  的每一行进行单位化处理。

步骤 5 设  $y_i \in \mathbb{R}^k (i = 1, 2, \dots, n)$  是对应矩阵  $U$  的第  $i$  行的向量。用  $k$  均值算法,将向量  $y_i (i = 1, 2, \dots, n)$  聚类到  $C_1, C_2, \dots, C_k$ 。

步骤 6 建立映射

$$x_i \in \mathbb{R}^p \mapsto y_i \in \mathbb{R}^k, i = 1, 2, \dots, n。$$

根据  $x_i$  和  $y_i$  之间的对应关系,利用步骤 5 中矩阵  $U$  的行向量  $y_i$  的聚类结果,确定点  $x_1, x_2, \dots, x_n$  在聚类中的结果。

输出: $x_1, x_2, \dots, x_n$  的聚类结果。

在答卷数据的谱聚类算法中,将原谱聚类算法中的 Laplacian 矩阵换成了扩展邻接矩阵。

## 4 实验结果与分析

以 MHK 考试的听力题答题结果,进行答卷数据的谱聚类实验。实验的软件为 Matlab 7.5,实验的数据为 979 名参加 MHK 考试的被试的答卷数据。

在 MHK 考试中,听力题的题型为单选题,设有 40 个小题,每小题有 A、B、C、D 共 4 个备选答案。被试将答题结果填涂在答题卡上。考试结束后,使用扫描仪将答题卡上答题结果扫描到数据库中,答题信息以字符串形式存储在一个字符型字段中,例如,一名被试的听力题的部分答题结果为:

ACBCDCCDACB+CB++CADACACBCBDBAB……

其中,“+”表示答题卡中对应的小题答案被试未填涂。

在实验前,将被试的答题卡的扫描数据库(采用 FoxPro 表)拷贝到 Matlab 工作目录下,并在数据库的第一条记录前插入一条空记录,把标准答案存入该记录的答题信息字段中。

利用 3.3 答卷数据的谱聚类算法对答卷数据对象进行聚类,聚类数  $k=5$ 。

聚类结果如图 1 所示。

图 1 聚类结果

在图 1 中,位于坐标原点,符号为“○”的是标准答案对象;符号为“\*”的是缺考或只回答了个别题的被试的答卷数据对象;符号为“◆”的是大部分被试的答卷数据对象;符号为“◇”、“□”、“+”的是少数被试的答卷数据对象,这些被试填涂的答卷结果与标准答案及大部分被试的答卷结果相差较大。

图 1 中的横坐标是被试的答卷数据对象在答题卡的扫描数据库中的记录号,纵坐标是答卷数据对象与标准答案对象的距离。当距离为 0 时,被试的答卷全部正确,当距离为 40 时,被试的答卷全部错误,其中包括缺考。

在答卷数据的谱聚类算法的距离矩阵  $A$  中,存储着被试的答题结果之间的关系,根据谱聚类的结果和距离矩阵  $A$  的每一行或每一列,可以建立一个分类图,因而可以建立一个分类图簇。图 2~图 4 是有代表性的几个分类图。

在图 2 中,是以在答题卡的扫描数据库中的记录号为 134 的被试的答题数据对象作为参照,建立的一

个分类图。在这个图中,可以清楚地看到缺考或只回答了个别题的被试的分布情况。

在图 3 中,是以在答题卡的扫描数据库中的记录号为 400 的被试的答题数据对象作为参照,建立的一个分类图。该对象所在类的符号为“◇”。

图 2 缺考和答题结果很差的答卷数据对象聚类情况

图 3 以聚类符号为“◇”的答卷数据对象为参照的聚类情况

图 4 以聚类符号为“□”的答卷数据对象为参照的聚类情况

在图 4 中,是以在答题卡的扫描数据库中的记录号为 183 的被试的答题数据对象作为参照,建立的一个分类图。该对象所在类的符号为“□”。

图 3 和图 4 中,位于横坐标轴上的这些被试填涂的答卷结果与标准答案及大部分被试的答卷结果相差较大。这些被试的答卷结果就是要检测的孤立点。

5 答卷数据的谱聚类算法存在的问题

在答卷数据的谱聚类算法中,由于需要计算答卷数据对象的距离矩阵和相似度矩阵,以及 Laplacian 矩阵的特征值,其空间复杂度为  $O(n^2)$ ,时间复杂度均为  $O(n^3)$ 。在被试数较大的情况下,会出现内存不足和运行时间过长的问题。

6 结束语

利用答卷数据的谱聚类算法,对所有被试的答题结果进行聚类后,根据距离矩阵,以每一位被试为参照,可以构成一个分类图,从而可以构建一个分类图簇。这为考试数据分析和孤立点检测提供了一种新的分析问题的方法和手段,可以从中挖掘出更多有价值的信息。

答卷数据的谱聚类算法经过修改后,可以应用到其它问题的数据分析和孤立点检测中。

参考文献:

[1] 孙焕良,鲍玉斌,于 戈,等. 一种基于划分的孤立点检测算法[J]. 软件学报,2006,17(5):1009-1016.  
[2] 陈宝国,郑丽英. 基于 Web 日志文件的孤立点检测算法[J]. 计算机与数字工程,2010,38(5):35-37.  
[3] 梁斌梅. 基于层次聚类的孤立点检测方法[J]. 计算机工程与应用,2009,45(32):117-119.  
[4] 刘 欢,吴介军,苏锦旗. 基于分化距离的离群点检测算法

[J]. 计算机应用研究,2010,27(9):3316-3318.  
[5] 廖国琼,李 晶. 基于距离的分布式 RFID 数据流孤立点检测[J]. 计算机研究与发展,2010,47(5):930-939.  
[6] 陆声链,林士敏. 基于距离的孤立点检测及其应用[J]. 计算机与数字工程,2004,32(5):94-97.  
[7] 余建桥,葛继科,李 娅. 一种基于密度偏差抽样的孤立点检测算法[J]. 计算机科学,2004,31(10):206-208.  
[8] 卢辉斌,徐 刚,李 段. 一种基于孤立点检测的入侵检测方法[J]. 微机发展(现更名:计算机技术与发展),2005,15(6):93-95.  
[9] Theodoridis S, Koutroumbas K. Pattern Recognition[M]. 4th ed. [s. l.]:Elsevier Publishers,2009.  
[10] 徐 森,卢志茂,顾国昌. 基于矩阵谱分析的文本聚类集成算法[J]. 模式识别与人工智能,2009,22(5):780-786.  
[11] 孙吉贵,刘 杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.  
[12] 蔡晓妍,戴冠中,杨黎斌. 谱聚类算法综述[J]. 计算机科学,2008,35(7):14-18.  
[13] Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing,2007,17(4):395-416.  
[14] Higham D J, Kibble M. A Unified View of Spectral Clustering[R]. England:University of Strathclyde,2004.  
[15] Tian Z, Li X, Ju Y. Spectral clustering based on matrix perturbation theory[J]. Science in China Series F Information Science,2007,50(1):63-81.  
[16] Taylor J S, Cristianini N. Kernel Methods for Pattern Analysis[M]. Cambridge:Cambridge University Press,2004.

(上接第 102 页)

连续若干感知时刻内信道占用状况不发生变化,为了提高频谱感知的实时性,文中提出了一种基于 OMP 的改进的重建算法 MOMP。MOMP 算法利用前一感知时刻获得信道占用信息,大大降低了 OMP 重建确定频谱位置时的计算量。仿真结果表明,MOMP 算法能有效地降低重建算法的耗时,并且达到和 OMP 方法基本一致的重建效果,提高了频谱感知的实时性。

参考文献:

[1] Mitola I J. Cognitive radio: An integrated agent architecture for software defined radio[D]. Sweden: KTH Royal Institute of Technology Stockholm,2000.  
[2] 胡 波,傅丰林,陈 东,等. 认知无线电系统中关键技术研究[J]. 电子元器件应用,2008(6):70-73.  
[3] 高欢芹,宋荣芳. 自适应 OFDM 系统中基于压缩感知的反馈压缩新方法[J]. 南京邮电大学学报(自然科学版),2010,30(3):16-19.  
[4] Donho D L. Compressed sensing[J]. IEEE Transactions on Information Theory,2006,52(4):1289-1306.  
[5] Tian Z, Giannakis G B. Compressed sensing for wideband cog-

nition on Acoustics, Speech and Signal Processing. [s. l.]: [s. n.],2007:1357-1360.  
[6] 戴琼海,付长军,季向阳. 压缩感知研究[J]. 计算机学报,2011,34(3):425-434.  
[7] Troop J, Gilbert A C. Signal recovery from partial information via orthogonal matching pursuit[J]. IEEE Transactions on Information Theory,2007,53(12):4655-4666.  
[8] 王璐瑜,朱 琦. 基于 DSCS 的宽带频谱感知新算法[J]. 信号处理,2011,27(6):813-819.  
[9] 石 磊,周 正,唐 亮. 认知无线网络中的压缩协作频谱感知[J]. 北京邮电大学学报,2011,34(5):76-79.  
[10] 顾 斌,杨 震,胡海峰. 基于压缩感知信道能量观测的协作频谱感知算法[J]. 电子信息学报,2012,34(1):14-19.  
[11] Kirolos S, Laska J, Wakin M, et al. Analog-to-information conversion via random demodulation[C]//Proceedings of IEEE Dallas/CAS Workshop on Design, Applications, Integration and Software. [s. l.]: [s. n.],2006:71-74.  
[12] Wang Yue, Tian Z, Feng Chunyan. A two-step compressed spectrum sensing scheme for wideband cognitive radio[C]//Proceedings of IEEE Global Telecommunications Conference. [s. l.]: [s. n.],2010:1-5.

# 考试数据分析及孤立点检测的谱聚类方法

作者: [贾志先](#)  
作者单位: [新疆财经大学 网络与实验教学中心, 新疆 乌鲁木齐 830012](#)  
刊名: [计算机技术与发展](#)  
英文刊名: [Computer Technology and Development](#)  
年, 卷(期): 2013(1)

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_wjtz201301028.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjtz201301028.aspx)