NJW 在离群数据挖掘中的应用研究

朱庆生,钟 洵,杨 鹏
ZHU Qing-sheng,ZHONG Xun,YANG Peng

重庆大学 计算机学院,重庆 400030
School of Computer Science,Chongqing University,Chongqing 400030,China
E-mail:brownbearbb@sina.com

ZHU Qing-sheng,ZHONG Xun,YANG Peng.Applied research on NJW for outlier detection.Computer Engineering and Applications,2010,46(7):128-130.

Abstract: Recently,spectral clustering has wide application in data mining.Outlier detection detects and analyzes outliers in order to find information.The NJW of spectral clustering algorithm is applied in data mining combined with outlier factor successfully, and a new outlier detection algorithm based on spectral clustering is proposed.Experimental results demonstrate this algorithm has better efficiency and wide applicability compared with the original outlier detection algorithm which based on clustering.
Key words: NJW;outlier detection;spectral clustering

摘 要: 最近几年,谱聚类思想开始用于数据挖掘领域,并取得了较好的效果;离群数据挖掘是对离群点进行检测,发掘出有用知识。将谱聚类中的 NJW 算法成功应用到离群数据挖掘领域,并结合离群指数的概念,提出了一种适合离群数据挖掘的谱聚类算法。与原有的基于聚类的离群检测算法相比,具有更好的效率和适应性。实验验证了所提算法的有效性和可行性。
关键词: NJW;离群数据挖掘;谱聚类
DOI: 10.3778/j.issn.1002-8331.2010.07.038 **文章编号:** 1002-8331(2010)07-0138-03 **文献标识码:** A **中图分类号:** TP18

1 引言

在数据集中,离群点是指与大量数据相比相对异常孤立的数据模式。在大多数的情况下离群点被视为噪声而遭抛弃,但在实际的应用中一些包含重要信息的数据也是离群点。离群数据挖掘就是从大量的数据中挖掘离群点。近年来,离群检测广泛应用于电信和信用卡欺诈检测,医疗保险,市场分析,气象分析等方面。

近年来基于数据挖掘概念的离群点检测研究取得了一定的进展,最常用的方法有基于聚类,基于统计,基于距离,基于深度,基于密度的算法等[1-2]。基于聚类的离群数据挖掘方法是很常用的一种离群检测方法。由于能在任意形状的样本空间上聚类,且收敛于全局最优解,谱聚类方法逐渐受到广大数据挖掘学者的重视。谱聚类是利用数据的特征向量,将原数据映射到相应的谱特征空间上,再对其进行分析,是建立在谱图理论上的[3-4]。Shi 和 Malik[5]在 2000 年根据谱图理论建立了 2-way 划分 Normailized-Cut(Ncut)的目标函数,设计了用于图像分割的算法。Ng AY,Jordan M I 和 WeissY[6]提出了基于多路切割思想的 NJW 算法,并已经从图像分割领域扩展到文本挖掘[7]和生物信息挖掘[8]等领域中。

分析离群数据可知,离群点可能代表了一种与主体结构特征相异的结构,或者是另一种行为趋势,为了使得不同的聚类结构能检测出不同的离群点,因此引入谱聚类方法显得极为重要:便于从结构特征对数据进行分析,找出离群点与主体数据结构的相异处。该文成功将 NJW 算法引入离群数据挖掘中,并结合了离群指数的概念, 提出了基于 NJW 算法的离群数据挖掘算法,较之传统的基于聚类的方法,该方法对于离群数据的检测更加有效,性能明显提高。实验验证了所提算法的有效性和可行性。

2 基于谱聚类的离群数据挖掘算法

2.1 NJW 算法

谱聚类将数据的聚类问题转化为了图的分割,将每个数据对象当作图中的顶点,由数据点间相似关系建立矩阵,获取该矩阵特征向量构成谱特征空间,并由此完成聚类。目前,谱聚类方法主要分为两大类:(1)基于 2-分迭代划分的思想,其代表算法是 Shi 和 Malik 提出的 SM 算法;(2)基于多路标准切割思想,其代表算法是 NJW 算法。选用 NJW 算法作为离群数据挖掘的基础算法,先将 NJW 算法简单介绍如下:

设数据集 $S=\{x_1,x_2,\cdots,x_n\}\subseteq Rd$,用图 $G=(V,E)$表示。$\boldsymbol{W}=(\omega_{ij})_{N\times N}$ 表示数据对象间的逐对相似度矩阵,通常选用较多的

基金项目: 国家“十一五”科技支撑计划重大项目资助(the National Great Project of Scientific and Technical Supporting Programs Funded by Ministry of Science & Technology of China During the 11th Five-year Plan. No.2007BAH08B04)。
作者简介: 朱庆生(1956-),男,教授,博士生导师,研究方向为数据挖掘、图形图像学等;钟洵(1985-),男,硕士研究生,研究方向为离群数据挖掘。
收稿日期: 2009-02-25 **修回日期:** 2009-04-10

相似度度量为 $w_{ij}=\exp(-\frac{\|x_i-x_j\|^2}{\sigma^2})$，设 $d_i=\sum_{j\in V}\omega_{ij}$，对角矩阵 $\boldsymbol{D}=diag(d_1,d_2,\cdots,d_N)$，则可构建 Laplacian 矩阵：$\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$，此时，矩阵 $\boldsymbol{P}$ 的前 K 个最大的特征值所对应的特征向量便可构成相应的谱特征空间以表示该数据集特征，该谱特征空间设为 $\boldsymbol{V}$，对 $\boldsymbol{V}$ 进行规范行向量得到矩阵 $\boldsymbol{Y}$，最后利用 K-MEANS 或其他算法将 $\boldsymbol{V}$ 的列向量聚成 K 类。

$$N_{cut}(\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_k)=\frac{cut(\boldsymbol{V}_1,\boldsymbol{V}_1^c)}{\sum_{i\in V_1}\sum_j\boldsymbol{W}_{ij}}+\frac{cut(\boldsymbol{V}_2,\boldsymbol{V}_2^c)}{\sum_{i\in V_2}\sum_j\boldsymbol{W}_{ij}}+\cdots+\frac{cut(\boldsymbol{V}_k,\boldsymbol{V}_k^c)}{\sum_{i\in V_k}\sum_j\boldsymbol{W}_{ij}} \tag{1}$$

$$h_{i,j}=\begin{cases}\frac{1}{\sqrt{vol(\boldsymbol{A}_j)}} & \text{if } \boldsymbol{V}_i\in\boldsymbol{A}_j\\ 0 & \text{其他}\end{cases} \tag{2}$$

式(1)表示了判断谱聚类的 MNcut 标准(其中 $\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_k$ 是对图的 K 维划分)，基于多路切割的 NJW 算法通过分析分段常量特征向量(PCE)就能满足 MNcut 最小的标准[9]。式(2)就是通过对特征向量的抽取得出的分段常量特征向量，设 $\boldsymbol{H}$ 是由这 K 个列向量构成的矩阵，则 $\boldsymbol{H}'\boldsymbol{H}=\boldsymbol{I}$，$\boldsymbol{h}_i'\boldsymbol{D}\boldsymbol{h}_i=1$，$\boldsymbol{h}_i'\boldsymbol{L}\boldsymbol{h}_i=cut(\boldsymbol{A}_i,\overline{\boldsymbol{A}_i})/vol(\boldsymbol{A}_i)$，那么最小化 MNcut 最终可转化为最小化 $\boldsymbol{T}'\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}\boldsymbol{T}$ 矩阵的迹，其中 $\boldsymbol{T}'\boldsymbol{T}=\boldsymbol{I}$，最后根据瑞利商定理可知，最小化 $\boldsymbol{T}'\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}\boldsymbol{T}$ 矩阵的迹的问题可转化为求矩阵的分段常量特征向量的问题。

2.2　基于 NJW 的离群数据挖掘算法

对带有离群点的数据的描述可分解为离群子特征系统与聚类子特征系统之和。离群点的出现使得聚类子空间发生了移动，不再是由最前面的 K 个特征特征向量所确定，那么，应该将有相同结构特征的离群点构成的离群类考虑在内，假设：$\Delta_1,\Delta_2,\cdots,\Delta_k,C_1,C_2,\cdots,C_n$，分别为数据集 S 中 k 个正常类(有可能包含个别离群点)和 n 个离群类，那么聚类子空间应该由前 $(k+n)$ 个特征向量所确定。用聚类特征系统就可以完整地描述数据集，只需 $(k+n)$ 个相应最大特征向量即可，实际上只需 $(k+n-1)$ 个，因为第一个是冗余的，即用 $(K+n-1)$ 维空间就可以完整地描述带有离群点的聚类。在后面的实验中已经证明这样的方法完全有效，并没有改变数据特征空间的性质。

在将数据集成功进行聚类之后，为了从中成功挖掘出离群数据，需要引入一个指标——离群指数[10]。

定义 1(大聚类和小聚类) 假设 S 为数据集，在用 NJW 算法对 S 成功聚类之后的结果为：$C=\{C_1,C_2,\cdots,C_K\}$，$C_i\cap C_j=\phi$，$C_1\cup C_2\cup\cdots\cup C_k=S$，并且有 $|C_1|\geqslant|C_2|\geqslant\cdots\geqslant|C_k|$，这里需要给出两个参数 α 和 β，定义 b 是大聚类和小聚类的边界处，则有：

$$(|C_1|+|C_2|+\cdots+|C_b|)\geqslant|D|^*\alpha \tag{3}$$

$$\frac{|C_b|}{|C_{b+1}|}\geqslant\beta \tag{4}$$

则有如下定义：

大聚类：$LC=\{C_i|i\leqslant b\}$

小聚类：$SC=\{C_j|j\geqslant b\}$

式(3)的定义表示大聚类中的数据个数不应小于整个数据个数的比例 α，式(4)表示大聚类至少应为小聚类的 β 倍。

定义 2(离群指数) 假设 S 为数据集，$C=\{C_1,C_2,\cdots,C_K\}$ 是其聚类，并有 $|C_1|\geqslant|C_2|\geqslant\cdots\geqslant|C_k|$，$\alpha,\beta,b,LC,SC$ 如定义 1，那么离群指数的定义为：

$$CBLOF(t)=\begin{cases}\frac{1}{|C_j|^*}\min(dis(t,C_j)),t\in C_i,C_i\in SC,C_j\in LC,j=1,2,\cdots,b\\ \frac{1}{|C_j|^*}(dis(t,C_j)),t\in C_i,C_i\in LC\end{cases} \tag{5}$$

式(5)的第 1 式表示小聚类中数据点 t 的离群指数应由 t 与其最邻近的大聚类的距离及其数据对象的个数决定，如果某个小聚类的数据点都是离群点，那这个小聚类则是要找的离群类。第 2 式则表示大聚类中的数据点 t 的离群指数应由 t 与所在类中心的距离和此聚类的数据对象个数决定，这样的定义能方便地找出有别于稳定的大聚类中的离群点。

讨论另一种情况，当类与类之间的界限很模糊时，聚类有时不能达到理想的效果，将一个小聚类和大聚类误聚成一个类，那么离群指数的定义能通过找离群指数最大的方式有效地找出小聚类(可能为离群类)，达到弥补聚类误差的效果。

结合 NJW 算法与离群指数则可设计出基于 NJW 的离群数据挖掘算法，具体算法流程如下：

输入：包含 n 个数据的数据集，预计普通聚类数 k 和离群聚类数 s，参数 α,β,σ

输出：m 个离群数据

步骤 1　将数据集构建成初步的相似性矩阵 $\boldsymbol{S}=(\omega_{ij})_{N\times N}$，其中：

$$\omega_{ij}=\exp(-\frac{\|x_i-x_j\|^2}{\sigma^2})$$

步骤 2　构造 Laplacian 矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$，其中 $\boldsymbol{D}$ 为对角线矩阵 $D_{ii}=\sum_{j=1}^{n}W_{ij}$；

步骤 3　求 $\boldsymbol{L}$ 的 $(k+n)$ 个最大特征值对应的特征向量 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{k+n}$ 构造矩阵 $\boldsymbol{V}=[v_1,v_2,\cdots,v_{k+s}]\in\boldsymbol{R}^{n\times(k+n)}$，其中 $\boldsymbol{v}_i$ 为列向量；

步骤 4　规范化 $\boldsymbol{V}$ 的行向量，得到矩阵 $\boldsymbol{Y}$，其中 $Y_{ij}=V_{ij}/(\sum_j V_{ij}^2)^{\frac{1}{2}}$；

步骤 5　将 $\boldsymbol{y}$ 的每一行看成是空间 $\boldsymbol{R}^{k+s}$ 内的一点，使用 K 均值或其他算法将其聚为 $(k+s)$ 类；

步骤 6　根据参数 α,β 及大，小聚类的定义将 $(k+s)$ 类分为大聚类集合 C_j，小聚类集合 C_i。

步骤 7　计算聚类中每个数据 t 的离群指数 CBLOF，其中：

$$CBLOF(t)=\begin{cases}\frac{1}{|C_j|^*}\min(dis(t,C_j)),t\in C_i,C_i\in SC,C_j\in LC,j=1,2,\cdots,b\\ \frac{1}{|C_j|^*}(dis(t,C_j)),t\in C_i,C_i\in LC\end{cases}$$

步骤 8 将数据点按离群指数的大小进行排序，根据输入参数 m,输出前 m 个离群指数最大的数据点,即为要求的 m 个离群数据点。

3 实验

为了验证新算法的可行性,在实验过程中将新算法用于人工数据集,真实数据集,并与利用 *K*NN 和基于 *K*-MEANS 的离群数据挖掘算法进行比较。实验证明,新算法对于高维数据具有良好的效率。

人工数据集的聚类效果如图 1 所示,其中 A_1,A_2,A_3,A_4是 4 个不同大小的正常聚类，在其周围分布了 5 个离群数据,理想状况是能在聚类时将 5 个离群点聚进离其最近的聚类当中,C_1是唯一的离群类。

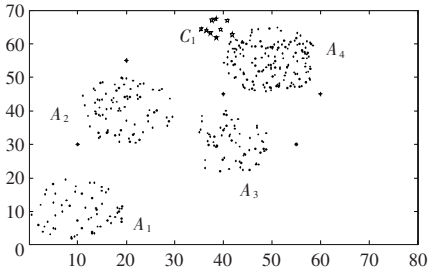


图 1 人工数据集

图 2(a)表示未加入离群点和离群类的特征空间,图 2(b)则表示加入了离群类与离群点的特征空间,从图 2(a)可以看到,第 2,3…k(k=4)个向量完整地描述了数据集,并能够成功地通过这些分段常量特征向量(PCE)表示聚类,也证明了第一个分段常量特征向量是冗余的。图 2(b)则证明了按文中算法的思想，加入离群点后不会影响特征空间对数据集的描述,离群点能被稳定地聚入离其最近的大聚类中。表 1 则表示了聚类成功后离群指数的情况。

通过对离群指数的计算,离群点与离群类中的离群指数远大于正常点的离群指数,从而可以方便地找出有别于稳定的大聚类中的离群点以及离群类。

对于人工数据集,*K*-MEANS 算法能够较好地进行聚类,并通过离群指数挖掘出离群数据，在时间效率上,*K*-MEANS 具有一定的优势。但是,在对高维,稀疏,并且样本空间不规则的数据集的实验中,该算法则比利用 *K*-MEANS 挖掘离群数据拥有更好的效果和性能。

为了在真实数据集上将算法与 *K*NN、基于 *K*-MEANS 的离群数据算法进行对比,引进离群误差率指标(Outlier Error Rate,OER):

$$OER=\frac{SO-RSO}{DO-RSO} \tag{6}$$

其中,*SO* 表示通过该算法找出离群点的数量,*RSO* 表示通过该算法找出的正确离群点的数量,*DO* 则表示数据集的数据个数。

表 2 列出了实验采用的基本数据集的信息。

表 2 数据信息

数据集	数据个数	属性数	聚类数(包含小聚类)	离群数据个数
气候检测	259	9	5	21
癌症基因	1 041	35	17	57
保险公司	7 992	86	11	261

由图 3 可以看出，基于 NJW 的离群数据挖掘算法在真实数据集的应用上相比 *K*NN 和基于 *K*-MEANS 的离群数据挖掘算法离群误差率更小,其中参数 α=0.8,β=4,σ=5。特别对于保险公司的数据集,由于数据高维、稀疏,*K*NN 和基于 *K*-MEANS 的离群数据挖掘算法的离群误差率更高,而该算法是对数据结构特征进行分析,数据维数与个数的增加并不会对数据的整体结构造成影响,因此,离群误差率较低,在 0.2 以下。这有效地证明了对于真实的数据集，基于 NJW 的离群数据挖掘算法具有更广泛的适用性和更高的性能。

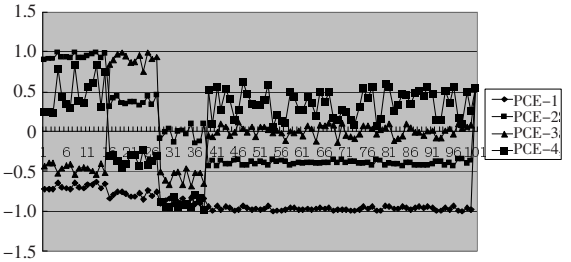


图 2(a) 未加入离群数据的分段常量特征向量(PCE)

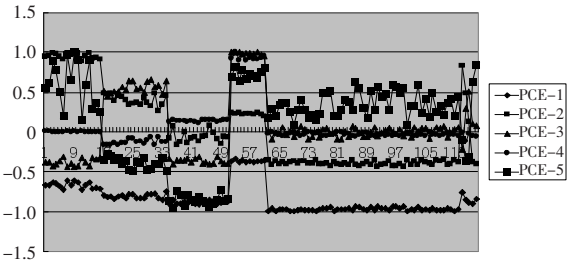
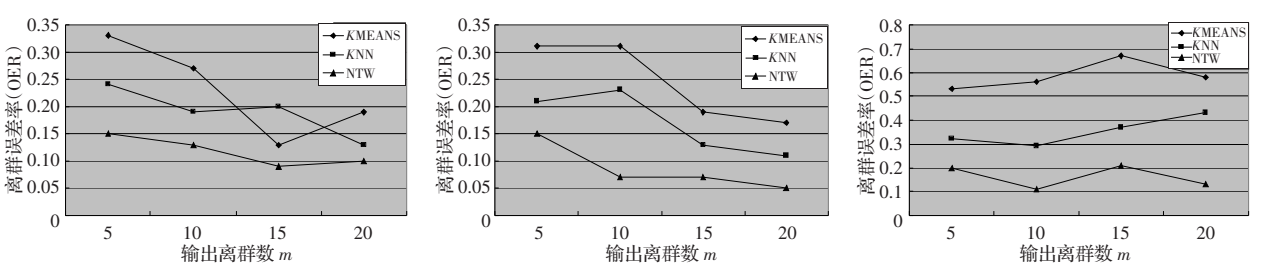


图 2(b) 加入离群数据的分段常量特征向量(PCE)

表 1 聚类中的离群指数范围

	聚类 1	聚类 2	聚类 3	聚类 4	离群类	离群点
离群指数	0.018 6~0.237 2	0.003 0~0.000 9	0.018 6~0.057 0	0.007~0.023	0.096~0.130	0.17~0.33



(a)气候检测数据 (b)癌症基因数据 (c)保险公司数据

图 3 离群误差率对比

(下转 212 页)