第47卷第4期 中国科学技术大学学报 Vol.47,No.4
2017年4月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Apr. 2017

文章编号:0253-2778(2017)4-0311-09

稀疏谱聚类算法在高维数据上的应用

徐雪丽,赵学靖

(兰州大学数学与统计学院,甘肃兰州 730000)

摘要:提出一种新的稀疏谱聚类算法——基于PAM算法的HSSPAM聚类(high-dimensional sparse spectral clustering based on partitioning around medoids).该算法先用高相关系数过滤及主成分分析降维方法以有效减小甚至消除维度灾难对高维数据处理的影响,再采用Minkowski距离指数变换函数及稀疏化算法来构建分块对角矩阵以重新解释样本之间的相似度;然后构造新颖的拉普拉斯矩阵以实现进一步压缩数据矩阵,进而结合partitioning around medoids(PAM)算法取代传统谱聚类中的K-means算法对特征向量聚类以提高算法的聚类稳定性;最后引入高维基因数据设计了实验,并以不同的聚类评价指标来衡量该研究算法的聚类质量,实验结果表明,新算法能够更精确、更稳定地对基因数据聚类.

关键词:高维数据聚类;稀疏谱聚类算法;降维方法;分块对角矩阵;聚类评价指标

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.04.005

引用格式:徐雪丽,赵学靖. 稀疏谱聚类算法在高维数据上的应用[J]. 中国科学技术大学学报,2017,47(4):311-319.

XU Xueli, ZHAO Xuejing. Application of sparse spectral clustering algorithm in high-dimensional data [J]. Journal of University of Science and Technology of China, 2017,47(4):311-319.

Application of sparse spectral clustering algorithm in high-dimensional data

XU Xueli, ZHAO Xuejing

(*School of Mathematics and Statistics, Lanzhou University, Lanzhou* 730000, *China*)

Abstract: A new sparse spectral clustering algorithm——high-dimensional sparse spectral clustering based on partitioning around medoids (HSSPAM) was proposed, which takes advantage of the sparse similarity matrix in computation as well as the superiority of the PAM algorithm over K-means. To reduce or even eliminate the impact of "dimensionality curse" on high dimensional data processing, the high correlation filter (HCF) and the principal component analysis (PCA) method are also investigated in the algorithm. The proposed method has higher precision and more stable clustering results than the algorithms introduced in this paper for comparison in the real high-dimensional gene data under different clustering evaluation criteria.

Key words: clustering of high-dimensional data; sparse spectral clustering algorithm; dimension-reduction technique; block diagonal matrix; clustering evaluation index

收稿日期:2016-08-28;**修回日期**:2016-12-08

作者简介:徐雪丽,女,1990年生,硕士生,研究方向:数据挖掘. E-mail: xuxl2014@lzu.edu.cn

通讯作者:赵学靖,博士/副教授. E-mail: zhaoxj@lzu.edu.cn

0 引言

高维数据是高新科学技术的重要产物，大数据蕴藏的巨大价值已经逐渐被人们认可，世界俨然进入“大数据”时代.如何提高驾驭庞大数据的能力，洞察分析数据结构空间，提取数据主干信息成为当今社会的热点话题.

近年来，越来越多的学者致力于研究高效处理这些海量数据、及时反馈有意义信息的数据挖掘工具.聚类是识别不同事物的核心工具，一个好的聚类算法不仅能够有效识别结构相似的样本数据点，并将此类数据点聚为一簇，而且能够将相似度较低的样本数据点加以区分便于聚类到不同簇.传统的聚类算法（如 K-means[1]、EM[2]、K-medoids[3]）主要集中于分析球面数据，不适用于识别非球面形状的簇，而实际应用中，出现了大量形状复杂、密度不均衡、凹陷分布的数据集，使得算法极易陷入局部最优从而无法得到最佳聚类划分[4]，不能根据用户的需求有效地分析数据并从中提取有用的信息.谱聚类弥补了传统聚类算法这一缺陷，它属于聚类分析中一个崭新的分支，不用预先对样本空间的整体结构做假设，且能够实现聚类任意形状分布的样本数据，为求解聚类问题提出了新思路，适用于很多实际应用问题，具有很大的应用潜力和科学研究价值.谱聚类算法通过构建样本数据点间的相似结构空间，提取此空间矩阵的有效特征值对应的特征向量，进而利用数据到特征向量空间的投影以实现数据高维复杂结构到低维简单结构的谱映射，最后采用传统聚类算法对特征向量空间聚类.此外，谱聚类对误差数据和噪声的敏感性不强，具有较好的鲁棒性[5].最初仅有图像分割、负载均衡、并行计算以及计算机视觉等领域使用谱聚类算法，随后在数据挖掘、机器学习等领域也开始使用谱聚类算法，而且都能取得较好的聚类效果.传统谱聚类算法多针对低维数据所设计，面对海量高维数据，相应的维度灾难问题随之凸显，同时冗余指标也随之增多，这些因素使得原有谱聚类方法不能有效地完成数据的价值“提纯”，降低了获得的聚类结果标签的价值.

谱聚类算法自 1992 年由 Hagen 等提出后，逐渐受到学者的关注并不断完善，在最近 20 年里，已提出多种类型的改进谱聚类算法[6-8]，但是这些改进的算法未能有效地解决高维数据聚类问题.本文针对高维样本数据，提出高维版本的稀疏谱聚类算法，该算法的研究建立在低维版本 SSKM 算法①的基础上，以高维数据的降维处理、相似度矩阵的稀疏化及拉普拉斯矩阵（以下简称“拉氏矩阵”）的构建为主线，引入高相关系数过滤及 PCA 投影等降维方法、Minkowski 距离的指数变换及稀疏化算法以及 PAM 聚类算法来实现算法的高效运行及高维数据的有效聚类.

1 理论依据及相关工作

1.1 算法的定义

谱聚类算法往往与图谱相联系.给定一组数据，构建一个无向加权图 $\boldsymbol{G}=\{\boldsymbol{V},\boldsymbol{E}\}$，它的表示形式为一对称矩阵，其中 $\boldsymbol{V}$ 中的元素称为顶点集或点，$\boldsymbol{E}$ 中的元素称为边.

设加权图的顶点集和边集分别为

$$\boldsymbol{V}=(v_1,v_2,\cdots,v_p),\ \boldsymbol{E}=(v_1v_2,\cdots,v_{p-1}v_p).$$

令 a_{ij} 表示 v_i 与 v_j 之间的边，即 $a_{ij}=v_iv_j$，则称矩阵 $\boldsymbol{A}=[a_{ij}]_{p\times p}$ 为图的邻接矩阵.$\boldsymbol{A}$ 是一个对称矩阵，表示待聚类数据点的相似度矩阵，它包含了聚类所需的所有信息，$\boldsymbol{W}=[w_{ij}]_{p\times p}$ 表示相似度矩阵的加权矩阵，w_{ij} 表示连接顶点 i 与 j 的权值[9].该图的拉氏矩阵可定义为：非规范化的拉氏矩阵，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$；规范化的 2 种拉氏矩阵，对称拉氏矩阵：

$$\boldsymbol{L}_{\mathrm{sym}}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$$

非对称随机游走拉氏矩阵：

$$\boldsymbol{L}_{\mathrm{rw}}=\boldsymbol{D}^{-1}\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{W}$$

式中，$\boldsymbol{D}=\{d_1,d_2,\cdots,d_n\}$ 是对角矩阵，称为“度矩阵”，它的对角线元素由邻接矩阵的行加和得到，$d_i=\sum w_{ij}$，$i=1,2,\cdots,n$，$j=1,2,\cdots,n$.拉氏矩阵是半正定矩阵，因此它的所有特征值是实数且是非负的，$0=\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$.

1.2 算法的描述

谱聚类算法最初来源于图论的思想，这是一类新发展的基于图论和图形学的原理来实现聚类的方法，即将数据集合想象为图的组成，进而将数据样本聚类问题转化为图的最优划分问题，通过有效分割子图来进行数据点的合理聚类[9].

标准谱聚类算法的实现主要依赖于数据集中数据样本的相似度矩阵 $\boldsymbol{A}$，加权矩阵 $\boldsymbol{W}$ 及度矩阵 $\boldsymbol{D}$.

相似度的计算通常采用高斯核函数：

① 徐雪丽，赵学靖. 稀疏谱聚类方法及应用[J]. 兰州大学学报（自然科学版），2016 录用.

$$\boldsymbol{A}=(a_{ij})=\begin{cases}\exp\left(\dfrac{-\|x_i-x_j\|^2}{2\sigma^2}\right), & i\neq j\\ 0, & \text{其他}\end{cases}$$

式中，$i,j=1,2,\cdots,n$.

几乎所有谱聚类的运行都是建立在某种相似性度量假设基础之上，完成相似度矩阵的构建之后，再稀疏化相似度矩阵生成相应的邻接矩阵及构建拉氏矩阵.稀疏化算法通常有 ε-近邻加权、K-最近邻加权及全连通加权三种方式.然后对拉氏矩阵进行奇异值分解，得到拉氏矩阵的特征值和特征向量，最后采用 K-means 算法对特征向量空 间中的特征向量进行聚类.

2　相关工作及稀疏谱聚类算法

谱聚类的首要关键是相似度矩阵的计算.相似度通常依据数据点之间的距离得到，通过将距离进行某种简单的变换生成相似度，计算两组数据 $x=(x_1,x_2,\cdots,x_n)$，$y=(y_1,y_2,\cdots,y_n)$ 的相似度 $a(x,y)$.首先计算数据点间的距离 $d(x,y)$，具体的距离算法由传统距离函数提供.选择一个合适的距离度量，使该度量能够正确刻画数据点间的内蕴结构.

本文算法中设计了新颖的相似度测量方式，解释为 Minkowski 距离的指数变换，距离转换为相似度的计算通过下式进行：

$$a(x,y)=\exp\left(-\left(\sum_{l=1}^{p}|x_{il}-x_{jl}|^q\right)^{\frac{\lambda}{q}}\right).$$

实值参数 $\lambda,q=1,2,\cdots,n$，其中参数 λ 控制相似度值的大小.该度量是一种组合函数，即距离的指数变换函数[10] 与 Minkowski 距离的合成.选择 Minkowski 距离的优势在于，不同的参数代表着不同的距离函数，扩大了数据类型的适用范围.

谱聚类算法旨在构建有效的拉氏矩阵，单纯地依赖已有拉氏矩阵有一定局限性.本文设计了一种压缩式拉氏矩阵

$$\boldsymbol{L}_{sp}=\boldsymbol{D}^{-1}\boldsymbol{L}\boldsymbol{D}^{-1}=\boldsymbol{D}^{-1/2}\boldsymbol{L}_{sym}\boldsymbol{D}^{-1/2},$$

其计算形式简单明了，只是稍微改造对称形式的拉氏矩阵 $\boldsymbol{L}_{sym}$，但将其运用于谱聚类算法中，效果良好.

2.1　稀疏化算法

针对大样本高维数据，为降低算法的复杂度及运行时间，数据矩阵稀疏化是必不可少的步骤，本文采取的稀疏化算法不同于传统谱聚类.

稀疏矩阵转换算法：预先设定截止阈值 α，输入数据矩阵，函数自动计算出稠密的相似度矩阵，同时识别阈值 α；然后将小于或者等于 α 的非对角元素及所有对角元素以 0 值存储，与此同时，保留大于 α 的元素不变；最后计算机自动展示一个稀疏的相似度矩阵.

稀疏相似度矩阵通过行列排序等初等变换，再经过局部调整得到分块对角矩阵[11].例如，一个稀疏相似度矩阵 $\boldsymbol{A}$，对 $\boldsymbol{A}$ 矩阵的行和列进行相同的排序调整，保证矩阵的对称性，把经过排序的稀疏矩阵按主对角线进行分块处理，再局部调整即可得到分块对角 $\boldsymbol{B}$ 矩阵.

$$\boldsymbol{A}=\begin{pmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \cdots & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nn}\end{pmatrix}\rightarrow$$

$$\boldsymbol{B}=\begin{pmatrix}\boldsymbol{A}_1 & & & \\ & \boldsymbol{A}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{A}_m\end{pmatrix}.$$

在聚类分析中，$\boldsymbol{B}$ 矩阵很容易被算法聚为 m 个类别，且以 $\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_m$ 为类集群.

2.2　截止阈值的选取

截止阈值的选取直接影响算法的运行效率及聚类结果.截断相似度较小的数据值，即在相似度矩阵中压缩为 0 值，该选取准则建立在相同类内部相似度尽量极大化、不同类之间相似度尽量极小化的原则之上，压缩步骤描述如下：

(Ⅰ)寻求初始概率值.将相似度矩阵的全部数据值升序排列后，寻求走势趋于平缓时的“拐点”，计算拐点及之前数据点的个数所占总数据点个数的比例，这个比例定义为初始概率值 P_0.

拐点计算方法：

①计算最大最小相似值数据点所在的直线 L；

②寻求其余相似度索引值点到直线 L 的最大距离，这个距离对应的数据点即为拐点.

(Ⅱ)生成截止阈值.输入初始 P_0 值，计算对应的分位数，这个值定义为算法预设的初步截止阈值 α，压缩相似度矩阵中小于或等于阈值的数据为0，保留其余数据不变，得到稀疏矩阵.稀疏矩阵必须满足：稀疏矩阵的每一行至少有一个数据不为 0，若出现矩阵行数据全为 0 的情况，则需减小初始概率 P_0 的取值至 P 值，直到矩阵行数据全为 0 的情况消失

为止，P 值为算法聚类所输入的最终概率值.

选取截止阈值的操作过程以常见的 Soybean 数据集为例.图 1 表示 Soybean 数据集的稠密相似度矩阵数据值与相似度值索引指标的散点图，从图中可以看到，当 $P_0=0.97$ 时，相似度值出现较明显的跳跃，随后表现出靠近 1 值的平稳状态，故此时的 P_0 值为初步预定的取值；与此同时计算的相似度矩阵出现了某些行值全部取 0 的情况，因此减小概率值至 $P=0.85$，以消除相似度矩阵某些行值全为 0 的情况，则这个值即是最终的概率值.

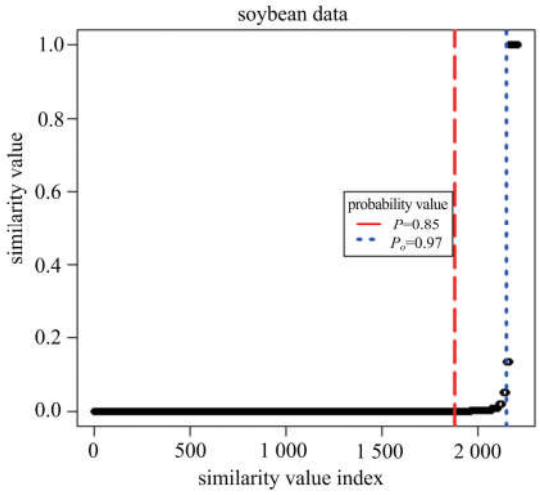

图 1 相似度值与相似度值索引指标

Fig1 Similarity value v.s. similarity value index

2.3 L_{sp} 与 L_{sym} 拉氏矩阵

本文算法中设计了 L_{sp} 形式的拉氏压缩矩阵，引入对称拉氏矩阵作比较，用一个简单的例子来描述压缩矩阵的效果，并以 Leukemia(HU6800)数据集为例进行了验证.

对于某些特殊的数据集，对称拉氏矩阵并不适用，这种情况很可能导致谱算法聚类效果不佳，为了使谱算法能够适用更多的数据集，本文采用压缩拉氏矩阵.图 2 展示了算法在 Leukemia(HU6800)数据集上，压缩矩阵及对称矩阵的最小特征值对应的特征向量的散点，左图表示压缩矩阵的最小特征值对应的特征向量，右图表示对称拉氏矩阵的最小特征值对应的特征向量，同类群体采用相同符号标识，观察左右图，左图中数据点较明显的聚为 3 个类组，而右图没有体现出较好的聚类群体.

2.4 稀疏谱聚类算法

HSSPAM 算法首先对原数据 $\boldsymbol{X}\in R^{n\times p}$ 进行规范化、相关系数及 PCA 预处理，选取前 z 个主成分向量对数据做投影变换得到新数据 $\boldsymbol{X}^*\in R^{n\times z}$，然

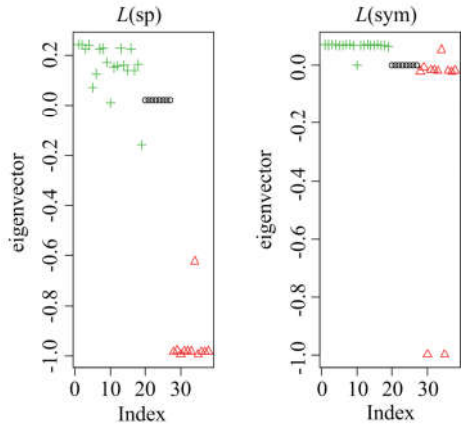

图 2 L_{sp} 及 L_{sym} 的最小特征值对应特征向量与样本索引值

Fig.2 The smallest eigenvectors of L_{sp} and L_{sym} v.s. sample index

后计算新数据 $\boldsymbol{X}^*\in R^{n\times z}$ 的相似度矩阵并进行矩阵稀疏化处理，再寻求拉氏矩阵 L_{sp} 的前 m 个最小特征值对应的特征向量，并构造特征向量空间 $\boldsymbol{G}=(g_1,g_2,\cdots,g_m)$ 及标准化特征向量得到对应的 $\boldsymbol{Y}=(y_{ij})$ 空间，其中 $y_{ij}=g_{ij}/(\sum_j g_{ij}^2)^{1/2}$，最终通过 PAM 算法[12]将空间 $\boldsymbol{Y}$ 聚类，得到聚类集群及相应的指标值.

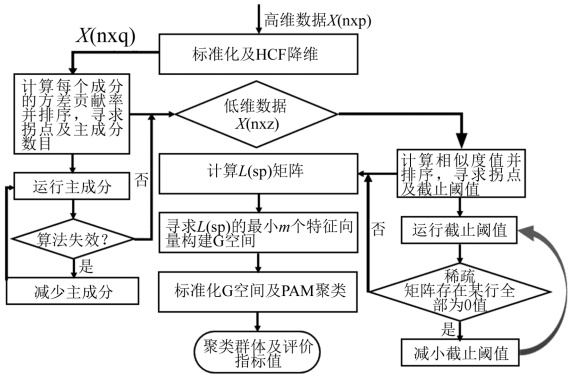

图 3 HSSPAM 算法流程图

Fig3 Scheme of the HSSPAM method

3 算法评价指标

一个算法的有效性或聚类结果如何，需要选择一个合适的评价标准来衡量，并且单纯的依赖一种评价指标往往不具有可靠性，选用多个指标来衡量算法的效果是古往今来只增不减的趋势，评价指标可大致分为内部评价指标和外部评价指标.本文介绍 3 种外部评价指标，Rand 指数(Rand index，RI)[12]、Jaccard 系数(Jaccard index，J)[13]及标准互信息(normalized mutual information，NMI)[14].

Rand 指数.给定拥有 n 个目标点的数据集 S，

设它的一个聚类集群 $P=\{p_1,p_2,\cdots,p_c\}$ 和一个真实划分 $T=\{t_1,t_2,\cdots,t_s\}$，n_{ij} 代表同时属于 p_i 类和 t_j 类的观测点的数目，$n_{i.}$ 代表属于 p_i 类的全部观测点的数目，同理，$n_{.j}$ 代表属于 t_j 类的全部观测点的数目，可通过比较 P 和 T 以及邻近矩阵与 T 来评价聚类的质量.对于基因表达数据集 D 中任意一对基因 (d_i,d_j)，一定满足下列 4 种类型之一：

Type(I)-d_i 和 d_j 在 P 中被聚类成同一簇，且在 T 中属于同一组；

Type(II)-d_i 和 d_j 在 P 中被聚类成不同簇，且在 T 中属于不同组；

Type(III)-d_i 和 d_j 在 P 中被聚类成不同簇，且在 T 中属于同一组；

Type(IV)-d_i 和 d_j 在 P 中被聚类成同一簇，且在 T 中属于不同组.

则 RI 可定义为

$$\mathrm{RI}(P,T)=\frac{\mathrm{Type(I)}+\mathrm{Type(II)}}{\mathrm{Type(I)}+\mathrm{Type(II)}+\mathrm{Type(III)}+\mathrm{Type(IV)}}=\frac{\binom{n}{2}+\sum_{i=1}^{c}\sum_{j=1}^{s}n_{ij}^{2}-\frac{1}{2}\left[\sum_{i=1}^{c}n_{i.}^{2}+\sum_{j=1}^{s}n_{.j}^{2}\right]}{\binom{n}{2}}.$$

RI 的取值在 $(0,1)$ 之间，P 和 T 完全匹配，RI 取 1 值，Rand 统计的值越小，两个划分 P 和 T 的差异越大.这个指标还有一些变种，如 Jaccard 系数.

Jaccard 系数是一个基于 presence-absence 属性的著名相似度衡量指标，常用于聚类分析，重点关注个体间共同具有的属性特征，其计算公式可表示为

$$J(P,T)=\frac{\mathrm{Type(I)}}{\mathrm{Type(I)}+\mathrm{Type(III)}+\mathrm{Type(IV)}}=\frac{n-\sum_{i=1}^{c}\sum_{j=1}^{s}n_{ij}^{2}}{n+\sum_{i=1}^{c}\sum_{j=1}^{s}n_{ij}^{2}-\sum_{i=1}^{c}n_{i.}^{2}-\sum_{j=1}^{s}n_{.j}^{2}}.$$

标准互信息(NMI).把 P 和 T 看作是两个随机变量，P 对应的熵为

$$H(P)=-\sum_{i=1}^{c}\frac{n_{i.}}{n}\log\left(\frac{n_{i.}}{n}\right).$$

T 对应的熵为

$$H(T)=-\sum_{j=1}^{s}\frac{n_{.j}}{n}\log\left(\frac{n_{.j}}{n}\right).$$

P 和 T 的联合熵为

$$H(P,T)=-\sum_{i=1}^{c}\sum_{j=1}^{s}\frac{n_{ij}}{n}\log\left(\frac{n_{ij}}{n}\right);$$

则 P 和 T 的 NMI 定义为

$$\mathrm{NMI}(P,T)=\frac{H(P)+H(T)-H(P,T)}{\sqrt{H(P)H(T)}}=\frac{I(P,T)}{\sqrt{H(P)H(T)}}.$$

式中，$H(P)$ 和 $H(T)$ 表示这两个划分的信息熵，$I(P,T)$ 则代表 P 和 T 的互信息，分母则限定 NMI 的取值在 0 到 1 之间.当 P 和 T 完全一致时，NMI 取 1 值，此时簇的聚类质量最好.

在聚类分配及真实划分已知的情况下，评价指标可忽略划分的期望特征，只注重所得簇的分配有效性，上述 3 类评价指标均如此，并且指标值均是取值越大，簇的分配与真实划分的相似度越高，聚类效果越好.

4　高维数据实例分析

为了测试本文提出算法的有效性，引入高维基因数据及聚类质量评价指标设计多组实验，并将本文提出算法 HSSPAM 与 PAM 算法、K-means 算法、FastSpectral NJW 算法[15]、complete-link 算法[16]、single-link 算法[16]及 average-link 算法[16]进行了比较. CNS tumors 数据集及 St. Jude Leukemia 数据集经过人类基因组 U95 系列基因芯片处理，而 Leukemia(HU6800)数据集经人类基因组 HU6800 系列基因芯片处理，并且 3 组数据集均可下载获得①，Colon 数据集、Leukemia 数据集、SRBCT 数据集、Prostate 数据集及 Lymphoma 数据集均来源于 Dettling (2004)的癌症研究[17]，基因数据集的详细信息如表 1 所示.

表 1 中，HCF 及 PCA 下的数据代表新数据维数，分别指经过相关系数和主成分分析方法处理后的数据维数.

4.1　选取主成分

主成分的选取直接影响数据的降维效果，下面介绍本文算法采用的选取准则，该选取标准满足截断原理：将方差贡献率较小的成分压缩为 0，即截断，保留贡献率较大的成分作为主成分.其选取方法描述如下：

① ttp://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi

表 1 基因表达数据集信息描述

Tab.1 The summary of gene-expression datasets' characteristics

原始数据($\boldsymbol{X}$)				新数据($\boldsymbol{X}^*$)	
数据集群	样本大小	维数	类组	HCF	PCA
Lymphoma	62	4 026	3	3 564	6
CNS tumors	48	989	5	930	5
St. Jude Leukemia	248	985	6	946	7
Leukemia	72	3 571	2	3 534	2
Colon	62	2 000	2	1 485	4
SRBCT	83	2 308	4	2 297	10
Prostate	102	6 033	2	5 339	1
Leukemia(HU6800)	999	38	3	930	6

(a)寻求初始主成分数目.将个体成分方差贡献率降序排列,寻求拐点(方法雷同截止阈值中拐点),拐点对应的索引值即为初始主成分数目 pc_0;

(b)确定有效主成分个数.pc_0 需满足:确保 PAM 算法的有效性,若出现失效情况,则减少初始主成分数目,直到失效情况消失,此时对应的主成分个数 pc 为有效主成分个数.

4.2 主成分选取实例分析

以 Colon 数据为例,结肠癌微阵列数据集(Colon cancer microarray data,Colon 数据)包括 2 000个不同的基因在 62 个样本条件下的基因表达水平,第一步经过相关性预处理,保留了 1 485 个基因(此时,相关性 $\rho<0.9$);第二步采取 PCA 降维方法,数据集投影变换到 4 维主成分空间中.

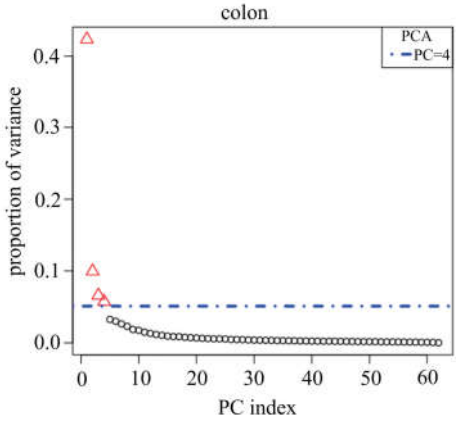

图 4 方差贡献比率与成分索引值

Fig.4 Variance contribution rate v.s. component index

图 4 展示了结肠癌微阵列数据集的前 50 个成分及每个特征值所解释的方差比率的散点.从图 4 可看到,方差贡献率呈现降序排列,它的取值开始出现了大幅度跳跃式减小的情形,随后又出现连续式缓慢减小并向 0 值靠近的情况,且以第 4 个点为临界点,因此可选取前 4 个成分为主要成分,其余成分截断,将数据集投影变换到 4 维主成分空间即可.

4.3 实验设计

为了对聚类结果的有效性进行评价,本文采取 3 个有效性评价指标对聚类结果进行评价.评价指标包括 Rand 指数(RI)、Jaccard 系数(J)及标准互信息(NMI).

实验中,由于某些比较算法受初始聚类中心选择的影响而导致聚类结果不稳定,因此以下报告的实验结果均为算法独立重复运行 20 次,聚类结果评价指标的平均值及方差,方差被标记在平均值后面的括号内.

4.4 不同指标下的实验结果

表 2~4 报告的实验数据为 7 种聚类算法在高维基因表达数据集上 20 次重复运行后的不同聚类指标下的指标均值及方差,每一列中,加黑数据为最优的实验结果,值得注意的是:斜体数值表示 HSSP0 与 PSSPAM 算法的不同指标值,其中 HSSP0 算法采用了 $\boldsymbol{L}_{\mathrm{sym}}$ 拉氏矩阵,其余运行与 HSSPAM 完全一致.

表 2~4 是基于不同聚类质量评价指标的聚类结果,其中 RI 是典型的聚类算法评价指标,该指标重点考察聚类后数据的划分与基于先验知识的数据划分的一致性,Type(I)和 Type(II)表示两个划分的一致程度,Type(III)和 Type(IV)表示两个划分的不一致程度;聚类基因表达数据旨在通过同一类中已知功能的基因来推测未知基因的功能,因此对基因表达数据聚类的效果评价也应该重点考察在同一类中的基因,通过聚类算法聚类后是否仍属同一类[18].本文引入 J 指标来修正 RI 指标的不足,重点考察个体间共同具有的属性特征;NMI 指标在簇数目已知的情况下重点衡量每个簇的质量、聚类划分的相对随机性程度及聚类结构与真实划分的互信息,并以聚类熵来度量.

表 2　聚类算法的 RI 均值及方差

Tab.2　RI mean value and deviation for clustering algorithms

Data	HSSP0	HSSPAM	PAM	K-means	FastSpectral NJW	complete-link	single-link	average-link
Lymphoma	1.00	**1.00**	0.922	0.773(0.0140)	0.840(0.0080)	0.776	0.894	0.894
CNS tumors	*0.847*	**0.990**	0.935	0.815(0.0040)	0.843(0.0005)	0.614	0.417	0.696
St. Jude Leukemia	0.990	**0.990**	0.903	0.926(0.0006)	0.925(0.0002)	0.741	0.241	0.508
Leukemia	*0.495*	**0.894**	0.570	0.709(0.0020)	$0.627(5.5\times10^{-5})$	0.583	0.524	0.532
Colon	0.535	**0.535**	0.494	0.506(0.0010)	$0.493(1.8\times10^{-6})$	0.491	0.525	0.500
SRBCT	*0.554*	**0.719**	0.693	0.624(0.0016)	$0.638(5.1\times10^{-5})$	0.391	0.296	0.296
Prostate	**0.523**	**0.523**	0.504	0.507	$0.507(6.7\times10^{-6})$	0.511	0.496	0.507
Leukemia (HU6800)	*0.707*	**0.963**	0.771	0.881(0.012)	$0.957(7.4\times10^{-5})$	0.771	0.411	0.771

表 3　聚类算法的 J 均值及方差

Tab.3　J mean value and deviation for clustering algorithms

Data	HSSP0	HSSPAM	PAM	K-means	FastSpectral NJW	complete-link	single-link	average-link
Lymphoma	1.00	**1.00**	0.849	0.576(0.0370)	0.690(0.0230)	0.612	0.813	0.813
CNS tumors	*0.478*	**0.579**	0.485	0.432(0.0040)	0.455(0.0030)	0.247	0.230	0.338
St. Jude Leukemia	0.956	**0.956**	0.608	0.690(0.0100)	0.709(0.0010)	0.432	0.216	0.298
Leukemia	*0.446*	**0.827**	0.424	0.436	$0.470(5.4\times10^{-5})$	0.442	0.527	0.527
Colon	0.469	0.469	0.381	0.388(0.0008)	$0.367(6.8\times10^{-6})$	0.404	0.520	0.469
SRBCT	*0.292*	0.350	0.280	0.215(0.0012)	0.214(0.0001)	0.255	0.259	0.259
Prostate	0.439	0.439	0.347	0.352	$0.341(1.2\times10^{-6})$	0.439	0.491	0.423
Leukemia (HU6800)	*0.514*	0.901	0.583	0.769(0.031)	0.884(0.0009)	0.584	0.362	0.584

表 4　聚类算法的 NMI 均值及方差

Tab.4　NMI mean value and deviation for clustering algorithms

Data	HSSP0	HSSPAM	PAM	K-means	FastSpectral NJW	complete-link	single-link	average-link
Lymphoma	1.00	**1.00**	0.805	0.658(0.0140)	0.760(0.0120)	0.588	0.723	0.723
CNS tumors	*0.695*	**0.756**	0.705	0.632(0.0060)	0.677(0.0020)	0.436	0.401	0.571
St. Jude Leukemia	0.963	**0.963**	0.765	0.827(0.0003)	0.757(0.0010)	0.646	0.070	0.491
Leukemia	*0.118*	**0.722**	0.236	0.202	0.263(0.0003)	0.130	0.027	0.027
Colon	0.087	**0.087**	0.046	0.044(0.0001)	$0.005(2.5\times10^{-5})$	0.045	0.030	0.087
SRBCT	*0.329*	**0.452**	0.383	0.187(0.011)	0.177(0.0011)	0.076	0.083	0.083
Prostate	**0.067**	**0.067**	0.014	0.019	$0.017(3.6\times10^{-6})$	0.055	0.036	0.034
Leukemia (HU6800)	*0.571*	**0.911**	0.688	0.832(0.019)	0.903(0.0002)	0.688	0.138	0.688

观察表 2～4，可以得出以下结论：

(Ⅰ) HSSP0 算法应用于 CNS tumors 数据、Leukemia 数据、SRBCT 数据及 Leukemia (HU6800)数据，指标值用斜体标注，数据值远小于 HSSPAM 算法的指标，其余指标数值与 HSSPAM 算法一致，表明对这 4 组数据集而言，HSSP0 算法并不适用，尤其是 Leukemia 数据集，在 NMI 指标下的值仅为 0.118，同时体现了 HSSPAM 算法较

HSSP0算法的适用数据集范围更加广泛.

(Ⅱ)经典 K-means 算法及改进的谱聚类算法(FastSpectral NJW)的聚类结果均不稳定,甚至有些波动较大,由标注的聚类质量评价指标的方差可知,在J和NMI两种指标下,k-means算法仅应用在Leukemia数据、Prostate数据集上的聚类结果稳定,这种现象不仅说明了数据集直接影响聚类结果的稳定性,还进一步指明了聚类评价指标的选取对聚类结果质量的好坏有不可忽视的影响,因此选取合适的评价指标至关重要.

(Ⅲ)在3种评价指标下,除个别数据外,改进的谱聚类算法较 k-means 算法的聚类有小的波动性.

(Ⅳ)single-link算法在Colon数据、Prostate数据集上及average-link算法在Colon数据上出现了最优的表现,由表3和表4的加黑数据可说明,PAM算法及complete-link算法没有较突出的聚类结果.

(Ⅴ)综合表2-4,在不同评价指标下,本文提出的高维版本稀疏谱聚类算法下的聚类结果数据都被加黑,仅有J指标下的Colon数据、Prostate数据集除外,但就此指标及此数据而言,新算法的表现仅劣于single-link算法,J值位居第二,较其余算法均有明显的聚类优势.另外,新算法的聚类结果显然有很不错的稳定性,没有被标记任何方差.

4.5 聚类结果统计显著性分析

为判断本文提出算法的聚类结果与其余算法聚类结果差异,引入无参数的Wilcoxon符号秩检验(Wilcoxon signed-rank test)设计了统计显著性检验实验.Wilcoxon符号秩检验是非参数统计中符号检验法的改进,此检验法是在符号检验的基础上发展起来的,较符号检验相比,考虑了差值的大小,其检验效率较传统的正负号检验(符号检验)有所提高,适用于非正态或形态不清的样本分布[19].在不同聚类评价指标下,将本文提出新算法的聚类实验结果指标均值分别与其余算法实验结果指标均值进行统计显著性检验,其中,原假设为在给定聚类评价指标下,提出算法的聚类结果与比较算法的聚类结果没有显著性异,备择假设为提出算法的聚类结果较比较算法的聚类结果有显著提高,置信区间为95%,即显著水平 α 设置为0.05.检验结果 $H(p)$ 由表5展示,H 表示假设检验结果,取1值代表拒绝原假设,此时 p 值小于0.05,即本文提出算法较比较算法的聚类结果有显著提高,取0值则提出算法与比较算法的聚类结果没有显著差异,此时 p 值大于或者等于0.05.

观察表5可知,在5组高维基因数据集上,本文提出算法获得的聚类结果在规定的允许误差范围内,比其余聚类算法获得的结果有显著提高的概率高达100%,这一结果与预期的结果十分吻合.

综合上述,所有的聚类评价指标值及聚类结果统计显著性分析,均能体现本文提出的算法对聚类高维基因数据是有效的,较其余算法能够获得更高的聚类质量,而且比其他聚类算法获得的结果具有一定的显著提高.

表5 提出算法与比较算法 Wilcoxon 符号秩检验结果 H(p)

Tab.5 Results $H(p)$ of Wilcoxon signed-rank test for HSSPAM algorithm versus other algorithms

指标	HSSPAM与PAM	HSSPAM与 K-means	HSSPAM与FastSpectral NJW	HSSPAM与complete-link	HSSPAM与single-link	HSSPAM与average-link
RI	1(0.005859)	1(0.005859)	1(0.005859)	1(0.005859)	1(0.005859)	1(0.005859)
J	1(0.005859)	1(0.005859)	1(0.005859)	1(0.00898)	1(0.01785)	1(0.00898)
NMI	1(0.005859)	1(0.005859)	1(0.005859)	1(0.005859)	1(0.005859)	1(0.00898)

5 结论

本文在对高维基因表达数据聚类时,提出一种新的具有针对性的谱聚类算法-高维版本的稀疏谱聚类(HSSPAM),该算法综合了降维方法、稀疏化相似矩阵、压缩式拉氏矩阵及PAM算法的优点,以解决高维数据的有效聚类问题.在不同的有效性聚类评价指标下,实例分析结果充分表明,该算法确实能够很好地应用在高维基因表达数据集上,不仅提高了聚类精度,还有效降低了数据维度,而且比其他方法的聚类结果有显著的提高,更重要的是,该算法的聚类结果比较稳定,输出的聚类指标值为0方差(表2~4,显示新算法输出值的方差为0),因此本文提出的聚类高维数据算法具有合理性及可行性.

参考文献(References)

[1] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA: AMS Press, 1967: 281-297.

[2] JEFF WU C F. On the convergence properties of the EM algorithm[J]. Annals of Statistics, 1982, 11(1): 95-103.

[3] PARK H S, JUN C H. A simple and fast algorithm for *K*-medoids clustering [J]. Expert Systems with Applications, 2009, 36(2): 3336-3341.

[4] DING S F, ZHANG L W, ZHANG Y. Research on spectral clustering algorithms and prospects [C]// Proceedings of the 2nd International Conference on Computer Engineering and Technology. ChengDu, China: IEEE Press, 2010: 149-153.

[5] LUXBURG U V. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395-416.

[6] HAGEN L, KAHNG A. New spectral methods for ratio cut partitioning and clustering [J]. IEEE Transactions on Computer-Aided Design, 2006, 11 (9): 1074-1085.

[7] SHI J B, MALIK J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22 (8): 888-905.

[8] NG A Y, JORDAN M, WEISS Y. On spectral clustering: analysis and an algorithm[A]. Advances in Neural Information Processing Systems [M]. MIT Press, 2002: 849-856.

[9] 徐天顺. 谱聚类算法研究[J]. 电脑知识与技术, 2012, 8(16): 3948-3950.

[10] Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: An *R* package for affinity propagation clustering [J]. Bioinformatics, 2011, 27 (17): 2463-2464.

[11] 谢国瑞. 线性代数及应用[M].北京:高等教育出版社,1999.

[12] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of Classification, 1985, 2(1):193-218.

[13] SAPORTA, YOUNESS G. Comparing two partitions: Some proposals and experiments [J]. Compstat. Physica-Verlag HD, 2002: 243-248.

[14] MARTIN L C, GLOOR G B, DUNN S D, et al. Using information theory to search for Co-evolving residues in proteins[J]. Bioinformatics, 2005, 21(22): 4116-24.

[15] MIYAHARA S, KOMAZAKI Y, MIYAMOTO S. An algorithm combining spectral clustering and DBSCAN for core points[J]. Advances in Intelligent Systems & Computing, 2014, 245: 21-28.

[16] MURTAGH F, LEGENDRE P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? [J]. Journal of Classification, 2014, 31(3): 274-295.

[17] DETTLING M. BagBoosting for tumor classification with gene expression data[J]. Bioinformatics, 2004, 20 (18): 3583-3593.

[18] 邓小燕, 甘晓玲, 唐宜. 谱聚类算法在基因表达数据分析中的应用[J]. 现代计算机, 2014, (6): 8-12, 24.

[19] 祝国强,杭国明,滕海英,等. 谈谈两总体比较的非参数检验方法[J]. 数理医药杂志, 2011, 24(5): 524-525.