第 34 卷 第 3 期
2019 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 34 No. 3
Jun. 2019

DOI:10.13364/j.issn.1672-6510.20180284

拉普拉斯矩阵在聚类中的应用

刘 颖，张艳邦
（咸阳师范学院数学与信息科学学院，咸阳 712000）

摘 要：高维数据受冗余数据和噪声数据的影响，聚类效率和准确率低，基于拉普拉斯矩阵的特征值和特征向量的特点，介绍了一种适用于高维数据的新的聚类中心选择算法，算法将拉普拉斯矩阵用于候选聚类中心选择前的数据降维处理，经过对数据进行降维处理，提高了候选聚类中心的准确性，增大了聚类准确率，扩大了聚类数据的种类范围. 在 10 个包含不同数量样本、维度、类别数的数据集上进行了聚类分析，实验结果表明了基于拉普拉斯降维的新聚类中心选择方法的有效性.

关键词：拉普拉斯矩阵；聚类；特征值；特征向量
中图分类号：TP391.4 **文献标志码**：A **文章编号**：1672-6510(2019)03-0076-05

Application of Laplacian Matrix in Clustering

LIU Ying，ZHANG Yanbang
（College of Mathematics & Information，Xianyang Normal University，Xianyang 712000，China）

Abstract：High-dimensional data is affected by redundant data and noise data，and the clustering efficiency and accuracy are low. Based on the characteristics of eigenvalues and eigenvectors of Laplacian matrix，a new algorithm for cluster center selection is introduced. The algorithm is suitable for high-dimensional data set. Laplacian matrix is used for data set dimension reduction before the selection of the candidate cluster center. After the dimensionality reduction of the data set，the accuracy of the candidate cluster center is improved，and the clustering accuracy is increased. The types of clustering data has been enriched. Cluster analysis was carried out on ten data sets containing different numbers of samples，dimensions and categories. The experimental results have justified the effectiveness of the new cluster center selection algorithm based on Laplacian matrix dimension reduction.

Key words：Laplacian matrix；clustering；eigenvalue；eigenvector

随着信息时代的发展，各行各业都产生了大量的数据，人们不再满足数据仅仅被电子化，而是希望对数据进行分析挖掘，透过数据的表象，找到隐藏在数据背后的规律和结构[1]. 聚类分析是数据挖掘的一个重要工具，聚类分析的目的是从一个未知数据集中发现隐含在其间的数据内在结构信息，将数据划分为若干个不相交的子集，每个子集成为一个簇，同一个簇内数据相似性大，簇间数据相异性大[2]. 数据聚类分析主要面临两个问题：一是如何确定聚类的结构；二是现在的数据大都是高维数据，如何能在聚类前对数据进行降维，从而提高聚类的效率[3]. 这两个问题也是目前研究的热点. 拉普拉斯矩阵(Laplacian matrix)也称为导纳矩阵，主要应用在图论中，作为一个图的矩阵表示，它广泛地应用在工程中[4–5]. 聚类问题从图的角度看就是对图的分割问题[6]，因此拉普拉斯矩阵被应用到聚类分析中，出现了一种谱聚类算法(spectral clustering)，该算法的核心思想就是把样本空间的聚类问题转化为无向图 G 的图划分问题[7]. 谱聚类算法在寻找聚类方面比传统算法(如 k-means)更有效[8]. 然而，当数据集很大时，谱聚类的时空复杂度都比较大. 为了对大数据集进行聚类，基于拉普拉斯矩阵，结合样本点的密度和距离，介绍了一种新的

收稿日期：2018–08–23；**修回日期**：2018–12–24
基金项目：国家自然科学基金资助项目(61501388)
作者简介：刘 颖(1974—)，女，天津人，讲师，liuying-1974@163.com

候选聚类中心选择方法. 该方法先利用拉普拉斯矩阵对数据集进行降维处理，对经过降维处理的数据求出其密度和距离两个参数，从而形成密度距离决策图；然后利用决策图选出候选聚类中心，对其进行合并，得到最终的聚类中心，最后将剩余点分配给聚类中心，完成聚类. 实验结果表明了算法的有效性.

1　拉普拉斯矩阵

1.1　拉普拉斯矩阵的概念

拉普拉斯矩阵[9]主要应用在图论中，是表示图的一种矩阵. 给定一个有 n 个顶点的无向图 $G=(V,E)$，其中 V 表示所有顶点 $v_1,v_2,\cdots,v_n$ 的集合，E 表示顶点之间连接的边的集合，拉普拉斯矩阵的定义如式(1)所示

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W} \tag{1}$$

式中：$\boldsymbol{D}$ 为图的度矩阵，$\boldsymbol{W}$ 为图的邻接矩阵. 用图 1 对拉普拉斯矩阵进行说明.

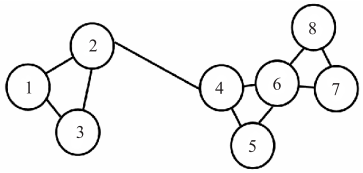


图 1　简单无向图

Fig. 1　Simple undirected graph

图 1 是由 8 个顶点组成的简单无向图，顶点的度简单的说是一个顶点连接的边的个数，度矩阵是一个对角矩阵，图 1 的度矩阵 $\boldsymbol{D}$ 为

$$\boldsymbol{D}=\begin{bmatrix}2&0&0&0&0&0&0&0\\0&3&0&0&0&0&0&0\\0&0&2&0&0&0&0&0\\0&0&0&3&0&0&0&0\\0&0&0&0&2&0&0&0\\0&0&0&0&0&4&0&0\\0&0&0&0&0&0&2&0\\0&0&0&0&0&0&0&2\end{bmatrix} \tag{1}$$

图 1 的邻接矩阵定义为 $\boldsymbol{W}=(w_{ij})_{n\times n}$，$w_{ij}$ 表示顶点 i 和顶点 j 之间是否有连接，w_{ij} 的计算公式如式(2)所示

$$w_{ij}=\begin{cases}1 & \text{顶点}i\text{和顶点}j\text{之间有边连接}\\0 & \text{顶点}i\text{和顶点}j\text{之间无边连接}\end{cases} \tag{2}$$

根据式(2)可知，图 1 的邻接矩阵为

$$\boldsymbol{W}=\begin{bmatrix}0&1&1&0&0&0&0&0\\1&0&1&1&0&0&0&0\\1&1&0&0&0&0&0&0\\0&1&0&0&1&1&0&0\\0&0&0&1&0&1&0&0\\0&0&0&1&1&0&1&1\\0&0&0&0&0&1&0&1\\0&0&0&0&0&1&1&0\end{bmatrix}$$

有了 $\boldsymbol{D}$ 和 $\boldsymbol{W}$，根据式(1)可知，图 1 的拉普拉斯矩阵为

$$\boldsymbol{L}=\begin{bmatrix}2&-1&-1&0&0&0&0&0\\-1&3&-1&-1&0&0&0&0\\-1&-1&2&0&0&0&0&0\\0&-1&0&3&-1&-1&0&0\\0&0&0&-1&2&-1&0&0\\0&0&0&-1&-1&4&-1&-1\\0&0&0&0&0&-1&2&-1\\0&0&0&0&0&-1&-1&2\end{bmatrix}$$

1.2　拉普拉斯矩阵的性质

拉普拉斯矩阵具有以下 4 个性质[10]：

(1) 拉普拉斯矩阵的最小特征值为 0，其所对应的特征向量为 1.

(2) 拉普拉斯矩阵是对称的半正定矩阵.

(3) 拉普拉斯矩阵有 n 个非负的实数特征值 $\lambda_0\leqslant\lambda_1\leqslant\cdots\leqslant\lambda_{n-1}$.

(4) 对于任意向量 $f\in R^n$ 有式(3)成立

$$f^{\mathrm{T}}Lf=\frac{1}{2}\sum_{i,j=1}^{n}w_{ij}(f_i-f_j)^2 \tag{3}$$

1.3　拉普拉斯矩阵特征映射

拉普拉斯特征映射(Laplacian Eigenmaps)[11]是从局部的角度出发来处理图，尽量在保留原图基本结构的情况下，将其映射到低维下表示. 根据拉普拉斯矩阵的性质，得出拉普拉斯特征映射的基本思想是希望相互有关系的点如图 1 中的顶点 1 和顶点 2，在降维后的空间中尽可能的靠近，相互之间没有关系的顶点如图 1 中的顶点 1 和顶点 6，在降维后的空间中尽可能的远离. 从拉普拉斯矩阵特征映射，发现其非常符合从高维数据中提取出能代表原始数据的低维表达这一情况.

2　拉普拉斯矩阵在降维中的应用

对于高维数据集来说，其包含有冗余信息以及噪声信息，使得这些数据在聚类的过程中准确率低，时

空复杂度高. 因此，为了高效发现数据的内在结构，通常在数据分析之前对数据进行降维处理，使用拉普拉斯的特征进行降维是目前流行的一种降维方法.

拉普拉斯特征映射的基本思想是对给定的有 n 个数据对象的高维数据集 $\boldsymbol{A}=\{x_1,x_2,\cdots,x_n\}$，在保持局部临近关系特征不变的情况下，找到数据集 $\boldsymbol{A}$ 对应的低维数据集 $\boldsymbol{B}=\{y_1,y_2,\cdots,y_n\}$.

降维的过程如下：

(1) 根据 K 近邻求出数据集的邻接矩阵 $\boldsymbol{W}$，对任意一对数据对象 x_i 和 x_j，若 x_i 在 x_j 的最近 K 个点内，则 $w_{ij}=1$，否则 $w_{ij}=0$；由于 x_i 与周围最近的 K 个点之间的距离大小不一，为了精确刻画点之间的距离，根据数据的需要，还可以为每条边赋权，利用热核函数为每条边赋权 $\boldsymbol{W}$ 如式(4)所示.

$$w_{ij}=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{t}\right) \tag{4}$$

(2) 利用拉格朗日乘数法计算拉普拉斯矩阵 $\boldsymbol{L}$ 的特征值：

$$Ly=\lambda Dy \tag{5}$$

(3) 计算式(5)中最小的前 $m+1$ 个特征值 $0=\lambda_0\leqslant\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_m$ 及其对应的特征向量，去除 $\lambda_0=0$ 所对应的特征向量，剩余的 m 个特征向量构成降维后的矩阵.

3　拉普拉斯矩阵在聚类分析中的应用

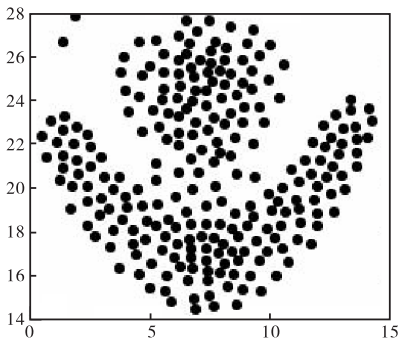
大多聚类算法都需要人为的选择聚类中心，经典的 k-means 算法需在开始随机给出 k 个点作为初始聚类中心，经过不断迭代，最后选择稳定下来的聚类分布作为最后的结果，由于初始聚类中心的任意性，在很大程度上影响了聚类效果. 2014 年发表在 Science 上的一个聚类算法 FSDP[12]，摒弃了 k-means 随机选择聚类中心的不确定性. FSDP 算法提出聚类中心一定是那些在某一个邻域内密度最高，且较其他高于自己密度的数据点的距离较远的点. 算法构造一个横坐标为密度 ρ、纵坐标为距离 δ 的决策图，供用户选择合适的聚类中心. 其中：第 i 个点的密度 ρ_i 表示在 i 点的指定邻域 dc 内包含的数据点的个数，即式(6)

$$\rho_i=\sum_{j\in I_x,j\neq i}\exp-\left(\frac{D\left(x_i,x_j\right)}{dc}\right)^2 \tag{6}$$

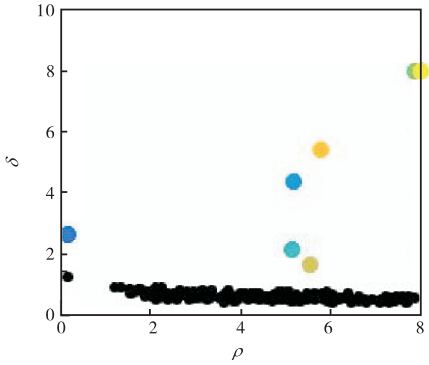
式中：$D\left(x_i,x_j\right)$ 表示 x_i 与 x_j 之间的距离；dc 表示指定的邻域. 第 i 个点的距离 δ_i 指的是到比其密度高的点的距离集合中最短的距离，如式(7)所示.

$$\delta_i=\begin{cases}\max\limits_{j:\rho_i>\rho_j}(d_{ij}) & \forall j,\ \rho_j\leqslant\rho_i\\ \min\limits_{j:\rho_j>\rho_i}(d_{ij}) & \text{其他}\end{cases} \tag{7}$$

该算法对低维数据聚类分析有众多优点，以数据集 flame 为例，如图 2 所示. 图 2(a)为二维数据集的图形表示，图 2(b)为由密度 ρ 和距离 δ 组成的决策图，算法通过选择找到聚类中心(ρ 和 δ 数值都大的点). 首先找到彩色的点即候选聚类中心，然后采用合并原则，对候选聚类中心进行合并，得到真正的聚类中心，最后完成聚类.



(a) 二维数据集的图形表示



(b) 决策图

图 2　Flame 数据集及其决策图

Fig. 2　Flame data set and its decision graph

由于此数据集维度低，且无噪声点，因此聚类效果好. 对于维度高、噪声多的数据集，由于其特征冗余、噪声点影响，所以从决策图很难准确找到聚类中心，如图 3 所示.

对此类数据集的聚类分析过程如下：

(1) 计算数据集 $\boldsymbol{S}=(s_{ij})_{a\times b}$ 的邻接矩阵 $\boldsymbol{M}$.

(2) 计算数据集 $\boldsymbol{S}=(s_{ij})_{a\times b}$ 的度矩阵 $\boldsymbol{D}$.

(3) 计算数据集 $\boldsymbol{S}=(s_{ij})_{a\times b}$ 的拉普拉斯矩阵 $\boldsymbol{L}$.

(4) 利用拉格朗日乘数法计算拉普拉斯矩阵的各

特征值和特征向量.

(5) 对特征值按升序进行排序，自次小特征值起取 m 个特征值，并求出其对应的 m 个特征向量，这些特征向量组成新的矩阵 $\boldsymbol{S}'=(s'_{ij})_{a\times m}(m<b)$.

(6) 对新的矩阵按式(6)和式(7)分别计算ρ和δ，最后画出新的决策图，找到聚类中心，完成聚类.

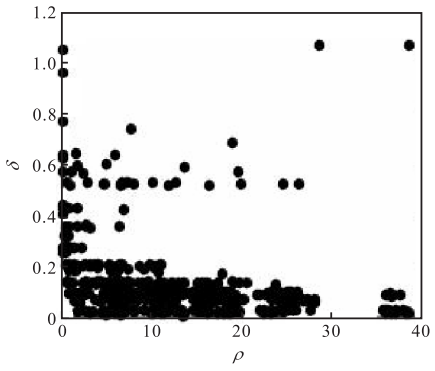


图 3　高维含噪声数据集决策图

Fig. 3　Decision diagram of high-dimensional noise-containing data set

4　实验结果与分析

为了测试算法对不同样本数目、维度大小、类别数目的高维数据集的有效程度，特别从 UCI(http://archive.ics.uci.edu/ml/datasets.html) 中选出了 Dermatology、Credit approval、German credit、Wine、Ionosphere 5 个数据集(维度大于 10 维、样本数目大小不一、类别数目不同)进行聚类分析实验，结果见表 1.

表 1　高维数据集的聚类信息表

Tab. 1　Clustering information table for high-dimensional data sets

数据集名称	样本数	维度	类数	正确率
Dermatology	366	34	6	0.81
Credit approval	653	15	2	0.85
German credit	699	20	2	0.72
Wine	178	13	3	0.94
Ionosphere	351	34	2	0.71

为了测试算法在对高维数据有效的同时，不失去对低维数据的效力，特别从 Clustering datasets(http://cs.uef.fi/sipu/datasets/) 中选出了 Flame、Jain、Smiles、Aggregation、Spiral 5 个低维数据集(维度 2 维、样本数目大小不一、类别数目不同)在同样的环境下进行了聚类实验分析测试，实验结果见表 2.

表 2　低维数据集的聚类信息表

Tab. 2　Clustering information table for low-dimensional data sets

数据集名称	样本数	维度	类数	正确率
Flame	240	2	2	0.95
Jain	373	2	2	0.98
Smiles	266	2	3	0.98
Aggregation	788	2	7	0.93
Spiral	312	2	3	0.98

实验在个人电脑上运行，采用 Matlab R2014b 编程工具进行聚类分析.

以 Dermatology 数据集为例对实验过程进行描述：

(1) 设数据集 Dermatology 为 $\boldsymbol{S}=(s_{ij})_{a\times b}$，$a$ 为样本数 366，b 为数据集的维数 34.

(2) 构造邻接矩阵 $\boldsymbol{M}=(m_{ij})_{a\times a}$，$m_{ij}$ 为样本 i 与样本 j 的相似度.

(3) 计算数据集 $\boldsymbol{S}=(s_{ij})_{a\times b}$ 的度矩阵 $\boldsymbol{D}=(d_{ij})_{a\times a}$.

(4) 计算数据集 $\boldsymbol{S}=(s_{ij})_{a\times b}$ 的拉普拉斯矩阵 $\boldsymbol{L}=(l_{ij})_{a\times b}$.

(5) 利用拉格朗日乘数法计算拉普拉斯矩阵的各特征值和特征向量.

(6) 对特征值按升序进行排序，自次小特征值起取 m 个特征值，并求出其对应的 m 个特征向量，这些特征向量组成新的矩阵 $\boldsymbol{S}'=(s'_{ij})_{a\times m}(m\leqslant b)$.

(7) 对矩阵 $\boldsymbol{S}'$ 按公式(6)、(7)分别计算ρ和δ，最后画出新的决策图，找到聚类中心.

(8) 将剩余的样本点按照最近邻原则分配到各个聚类中心，完成聚类.

表 1 的实验结果显示聚类数目选择正确，聚类效果良好，这表明通过使用拉普拉斯矩阵对数据集进行降维处理，能有效处理冗余数据和噪声数据，从而纯化数据集，提高候选聚类中心选择的正确率，进而达到了提高聚类效率和正确率的目的. 表 2 的实验结果表明，算法不仅对高维数据集有效，对低维数据集效力并没有消失，聚类数目选择正确，聚类结果正确率高，因此证明算法对聚类数据集一定的宽泛性.

5　结　语

本文讨论了拉普拉斯矩阵的原理，针对拉普拉斯矩阵的特征，首先，使用拉普拉斯矩阵对数据集进行降维处理，为数据的高效聚类打好基础；其次，结合 FSDP 算法，对降维后的数据集求出密度和距离，得

到候选聚类中心，对候选聚类中心进行合并，获得最终的聚类中心；最后，将剩余样本点分配到各个聚类中心，求出最终的聚类结果. 拉普拉斯矩阵的应用降低了冗余数据和噪声数据对数据结构的影响，在 10 个不同样本数、不同维度、不同类别数的数据集上进行聚类分析实验，实验结果表明算法的有效性. 拉普拉斯矩阵在聚类中的使用，凸显了拉普拉斯矩阵特征的实用性，为在其他领域使用提供了启示.

参考文献：

[1] Yang Y，Ma Z，Yang Y，et al. Multitask spectral clustering by exploring inter task correlation[J]. IEEE Transactions on Cybernetics，2015，45(5)：1083–1094.

[2] 杜辉，王宇平，董晓盼. 采用万有引力定律自动确定类数的 K 均值算法[J]. 西安交通大学学报，2014，48(10)：115–119.

[3] 王伟文. 拉普拉斯特征映射新增样本点问题及正则化降维研究[D]. 广州：暨南大学，2017.

[4] 谢德喜. 拉普拉斯变换在工程方程中的应用[J]. 天津轻工业学院学报，1990(1)：103–110.

[5] 汪玉美，陈代梅，赵根保. 基于目标提取与拉普拉斯变换的红外和可见光图像融合算法[J]. 激光与光电子学进展，2017，54(1)：98–106.

[6] Hagen L，Kahng A B. New spectral methods for ratio cut partitioning and clustering[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，1992，11(9)：1074–1085.

[7] 胡乾坤，丁世飞. 局部相似性优化的 p–谱聚类算法[J]. 计算机科学与探索，2018，12(3)：462–471.

[8] 郭磊，杨静，宋乃庆. 谱聚类算法在不同属性层级结构诊断评估中的应用[J]. 心理科学，2018，41(3)：735–742.

[9] 朱晓欣. 拉普拉斯矩阵特征值的图论意义[J]. 江苏教育学院：自然科学版，2006，23(1)：19–20.

[10] 郭继明. 图的拉普拉斯特征值[D]. 上海：同济大学，2006.

[11] 侯臣平，吴翊，易东云. 新的流形学习方法统一框架及改进的拉普拉斯特征映射方法[J]. 计算机研究与发展，2009，46(4)：676–682.

[12] Rodriguez A，Laio A. Machine learning. Clustering by fast search and find of density peaks[J]. Science，2014，344(6191)：1492–1496.

责任编辑：常涛，郎婧

（上接第 37 页）

boxymethylcellulose solutions[J]. Cellulose，2007，14(5)：409–417.

[7] Cuissinat C，Navard P. Swelling and dissolution of cellulose part I：Free floating cotton and wood fibres in N-methylmorpholine-N-oxide-water mixtures[J]. Macromolecular Symposia，2010，244(1)：1–18.

[8] Cuissinat C，Navard P. Swelling and dissolution of cellulose part II：Free floating cotton and wood fibres in naoh-water-additives systems[J]. Macromolecular Symposia，2010，244(1)：19–30.

[9] 张志慧，徐立新. 关于水溶性纸的生产方法和溶解性能的研究[J]. 黑龙江造纸，2002，30(3)：6–8.

[10] Zhang H，Zeng X，Xie J，et al. Study on the sorption process of triclosan on cationic microfibrillated cellulose and its antibacterial activity[J]. Carbohydrate Polymers，2016，136(10)：493–498.

[11] Fatehi P，Hamdan F C，Ni Y. Adsorption of lignocelluloses of pre-hydrolysis liquor on calcium carbonate to induce functional filler[J]. Carbohydrate Polymers，2013，94(1)：531–538.

[12] Derecskei B，Derecskei-Kovacs A. Molecular dynamic studies of the compatibility of some cellulose derivatives with selected ionic liquids[J]. Molecular Simulation，2006，32(2)：109–115.

[13] Kozbial A，Li Z，Sun J，et al. Understanding the intrinsic water wettability of graphite，graphene，and 2D materials[C]. APS March Meeting 2014. Denver：American Physical Society，2014.

[14] Barrett P，Glennon B. Characterizing the metastable zone width and solubility curve using Lasentec FBRM and PVM[J]. Chemical Engineering Research and Design，2002，80(7)：799–805.

[15] Heath A R，Fawell P D，Bahri P A，et al. Estimating average particle size by focused beam reflectance measurement(FBRM)[J]. Particle & Particle Systems Characterization，2002，19(2)：84–95.

[16] Etter M C. Encoding and decoding hydrogen-bond patterns of organic compounds[J]. Accounts of Chemical Research，1990，23(4)：120–126.

责任编辑：周建军