

基于 p -Laplacian 的多类均衡聚类算法

张 凯¹ 陶青川^{1*}
ZHANG Kai TAO Qingchuan

摘 要

传统谱聚类算法在面对多维度、非线性复杂数据时聚类效果表现不佳, p -Laplacian 算子的引入一定程度上缓解了此类问题, 然而基于 p -Laplacian 的谱聚类方法只依据第二特征向量, 主要是以二分类的递归方式不断更新聚类结果, 导致在面对多类聚类问题时聚类效果不稳定的情况。因此, 文章基于上述研究提出一种基于 p -Laplacian 的自适应聚类的多类聚类方法, 该方法优化了权重矩阵求解方法, 创新性选择多个特征向量分别进行聚类, 最后根据投票原则进行最终聚类得到更为均衡稳定的聚类结果。

关键词

无监督学习; 谱聚类; p -Laplacian; 多类聚类

doi: 10.3969/j.issn.1672-9528.2026.01.037

0 引言

随着机器学习的日益发展, 聚类方法的研究也在不断优化改进。聚类作为一种无监督学习, 不需要提前知道数据的真实标签类别, 根据数据间的内部关系将数据分为几类^[1], 使其在同一类别中的数据点相似性最大, 在数据分类中有着广泛应用^[2]。传统的聚类方法有: K-means、DBSCAN 等^[3], 这些聚类方法原理简单, 主要根据距离或密度进行分类, 在面对非球形簇或较为复杂的数据类型时并不适用^[4]。而当前流行的谱聚类算法通过对数据的相似性进行谱分解, 一定程度上提高了聚类的准确性和稳定性^[5], 但在面对高维复杂数据和非线性数据时, 同样出现了准确率下降的问题。 p -Laplacian 算子^[6]的引入一定程度缓解了这种问题, 但其在多类聚类问题上聚类表现不稳定。

本文在基于 p -Laplacian 的谱聚类方法的基础上, 对其输入特征向量进行改进, 采用多个特征向量进行聚类后再对结果进行聚类, 在面对多类聚类时, 效果更加均衡, 结果更加稳定。

1 相关工作

基于 p -Laplacian 的谱聚类方法在面对非球形簇或是较为复杂的数据类型时, 特别是二分类表现效果较好, 且不易受到噪声干扰, 对于高维度非线性数据也能较好处理。

但目前聚类数据常常面临着多个聚类类别且类间特征重叠的情况, 上述聚类方法不能很稳定高效地完成聚类目标, 针对该问题, 本文基于多类别且类间特征重叠的聚类数据, 对基于 p -Laplacian 的谱聚类方法进行调整和改进, 相关工作主要有以下方面:

(1) 基于切比雪夫距离方法^[7]求解相似矩阵, 优化权重矩阵求解方式。

(2) 基于谱聚类思想, 采用多初始特征向量进行聚类, 弥补对多类聚类问题不稳定的不足。

(3) 基于投票式聚类原则^[8]对多个聚类结果进行投票聚类得到更为均衡稳定的结果。

2 算法细节

2.1 基于 p -Laplacian 的谱聚类方法

谱聚类算法建立于谱图理论 (spectral graph theory) 基础之上, 是一种广泛应用聚类算法。谱聚类的基本思想是将传统聚类问题转化为图割问题: 该算法将聚类样本看作节点, 在任意两点间形成连接, 度量节点间的相似性并由此定义连接的权重, 通过对所有节点组成的图进行切割, 使得切图后的子图间权重和较低、子图内权重和较高, 从而达到聚类的目的^[9]。对于一个图 G , 用点的集合 V 和边的集合 E 来描述。切图方式 N_{cut} 为:

$$N_{\text{cut}}(A_1, A_2, \dots, A_c) = \sum_{i=1}^c \frac{\text{cut}(A_i, \bar{A}_i)}{\text{vol}(A_i)} \quad (1)$$

式中各部分定义为:

$$\text{cut}(A_i, \bar{A}_i) = \sum_{i \in A_i, j \in \bar{A}_i} \omega_{ij} \quad (2)$$

$$\text{vol}(A_i) = \sum_{p \in A_i, q \in V_{pq}} \omega_{pq} \quad (3)$$

由于子图样本的个数多并不一定权重就大, 切图时基于权重也更符合本文目标, 因此一般来说 N_{cut} 切图优于 Ratio-Cut 切图。Luo 等^[10]基于 N_{cut} 提出了一种新的切图方式 C_{cut} , 该切图方式基于 N_{cut} 进行了改进。

$$C_{\text{cut}}(A_1, \dots, A_c) = \sum_{i=1}^c \frac{\text{cut}(A_i, \bar{A}_i)}{\min(A_i)} \quad (4)$$

然而根据瑞利商原理计算式 (4) 的最优解是一个 NP 难

1. 四川大学电子信息学院 四川成都 610065

的问题, 所以为求得 C_{cut} 的最优解, 引入 p -Laplacian 算子解决这一问题, p -Laplacian 可定义为:

$$\langle f, \Delta_p f \rangle_i = \frac{1}{2} \sum_{i,j=1}^n w_{ij} (f_i - f_j)^p \quad (5)$$

根据公式 (5) 可以推导出以下表达式:

$$\langle \Delta_p^{(u)} f \rangle_i = \sum_{j \in V} w_{ij} \phi_p(f_i - f_j) \quad (6)$$

$$\langle \Delta_p^{(n)} f \rangle_i = \frac{1}{d_i} \sum_{j \in V} w_{ij} \phi_p(f_i - f_j) \quad (7)$$

其中:

$$\phi_p(x) = |x|^{p-1} \text{sign}(x) \quad (8)$$

令 $\mathbf{1} \in \mathbf{R}^n$ 表示全 1 向量, 求取特征向量的最小化目标函数定义为:

$$F_p(f) = \frac{\langle f, \Delta_p^{(u)} f \rangle}{\min_{c \in \mathbf{R}} \|\mathbf{f} - c\mathbf{1}\|_p^p} \quad (9)$$

$$\mathbf{v} = \underset{\mathbf{f}}{\text{argmin}} \frac{\langle f, \Delta_p^{(u)} f \rangle}{\min_{c \in \mathbf{R}} \|\mathbf{f} - c\mathbf{1}\|_p^p} \quad (10)$$

通过上式推导求得 p -Laplacian 矩阵的特征向量:

$$\mathbf{v}_p^{(n)} = \mathbf{u}^* - c^* \mathbf{1} \quad (11)$$

$$c^* = \underset{c}{\text{argmin}} \sum_{i=1}^n |u_i^* - c|^p \quad (12)$$

本文将上述方法简称为 PSC。

2.2 邻接矩阵构建

对于 p -拉普拉斯聚类算法中邻接矩阵的构建, 选择一个能够真实描述数据样本间结构关系的相似性计算方法非常关键。各方面研究中最普遍使用的相似性计算方法是依赖于欧氏距离的高斯核函数, 但当数据集中存在离群点或异常值时, 该种计算方式无法准确反映数据的空间分布结构, 因此本文提出一种基于切比雪夫距离的相似度计算方式, 数据样本之间的权重值可表示为:

$$w_{ij} = W(K_{\text{dist}}) = e^{-\frac{\alpha \cdot K_{\text{dist}}}{\text{gamma}}} \quad (13)$$

式中: K_{dist} 表示每个数据点到前 k 个最近邻的切比雪夫距离; α 表示控制高斯核函数衰减的参数; $\text{gamma} = \max\{K_{\text{dist}1}, K_{\text{dist}2}, \dots, K_{\text{dist}k}\}$ 表示局部自适应的缩放参数。

该方式通过局部连接叠加高斯核的方式构建出邻接矩阵 $\mathbf{W}$, 着重关注数据点之间的维度差异, 在面对异常值或噪声数据时具有更好的表现和稳健性。

2.3 投票式聚类原则

针对多个聚类结果发现, 部分元素一直同属一个类别。而对于剩余的数据, 由于聚类算法存在不稳定性, 给了一定的聚类允许误差^[11]。由此, 定义了一个投票机制。

首先, 先给了一个隶属度的概念。函数定义为:

$$g(x_i) = \frac{m_j}{N} \quad (14)$$

式中: x_i 表示第 i 个聚类结果; N 表示聚类次数; m_j 表示 N 中属于 j 类簇的次数。

然后规定函数:

$$h(x_i) = \begin{cases} x_i \in j, & g(x_i) \geq 0.5 \\ x_i \in k, & g(x_i) < 0.5 \end{cases} \quad (15)$$

最后根据投票函数得出最均衡的聚类结果。

2.4 基于 p -Laplacian 的多类均衡聚类算法

原始 p -谱聚类算法的基本理论, 仅通过第二小特征向量来解决二分问题。然而, 对于多类划分问题, 需要采用层次化策略, 并通过递归的方式执行二分过程, 同时设置适当的阈值。仅依赖第二小特征向量来处理复杂数据集往往会导致图划分效果不佳, 且聚类结果的不稳定性。本文引入了多初始特征向量进行聚类^[12], 并对相应的聚类结果根据投票策略再次聚类, 最终得到效果更为稳定均衡的聚类结果。将该方法简称为 SPSC。具体如表 1 所示。

表 1 基于 p -拉普拉斯的多类均衡聚类算法流程

步骤	操作
输入:	样本分布矩阵 $\mathbf{X}$, 期望簇数 k , p -拉普拉斯算子的值, 特征向量数目 n
Step1	根据切比雪夫距离计算权重矩阵
Step2	对于每个集群 C_i ($i = 1, 2, \dots, s$), 最小化目标函数 $F_p^{(s)}$
Step3	求解第 2,3,..., n 小特征向量
Step4	根据特征向量求解聚类结果
Step5	对 $n-1$ 个特征向量聚类结果再进行投票式聚类
Step6	输出期望簇数 k , 算法结束

3 实验对比

3.1 评价指标

在数值计算中引入 ACC (Accuracy) 和 NMI^[13] 作为衡量标准, 两种衡量指标的计算公式为:

$$\text{ACC} = \frac{\sum_{i=1}^n \delta(s_i, \text{map}(r_i))}{n} \quad (16)$$

$$\text{NMI}(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}} \quad (17)$$

3.2 实验数据集

本文进行实验的数据集为经典的人工生成的数据集: Two-Moon、ThreeCircle、Smile、Multi-scale、Blobs and Circle, 为了验证本方法的优越性, 并在其基础上增加了多维度的噪声, 使其变得更加复杂。此外还在 UCI 真实数据集 WDBC、GLASS, 人脸数据集 AT&T、UMIST 上进行了对比试验。

3.3 实验结果

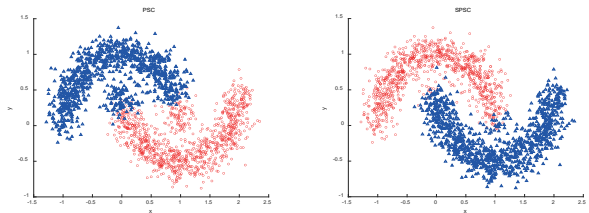
3.3.1 人工数据集验证结果

为了验证 SPSC 的聚类效果，对比 PSC 在五种人工数据集上的 ACC 以及 NMI，实验结果如表 2 所示。

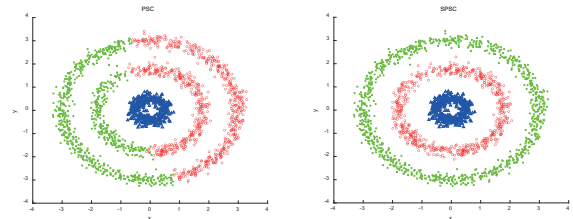
表 2 两种聚类方法在人工数据集上的聚类指标

数据集名称	评价指标	PSC	SPSC
TwoMoon	ACC	0.900 5	0.972 5
	NMI	0.532 6	0.818 6
ThreeCircle	ACC	0.696 5	1.000
	NMI	0.696 2	1.000
Smile	ACC	0.798 5	0.986 0
	NMI	0.692 5	0.982 7
Multi-scale	ACC	0.751 0	0.960 0
	NMI	0.614 7	0.859 2
Blobs and Circle	ACC	0.789 0	0.994 5
	NMI	0.699 1	0.971 1

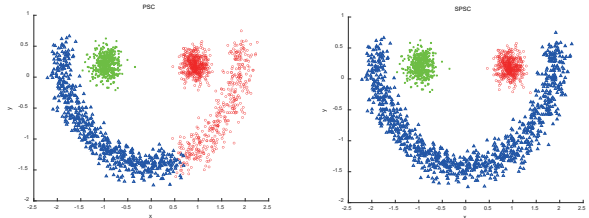
图 1 为 PSC 和 SPSC 在人工数据集上的检测效果，可以看出本文提出的 SPSC 能更好地区分出聚类边界。聚类准确率达有 7%~30% 左右的提高，NMI 有 20%~30% 左右的提高。



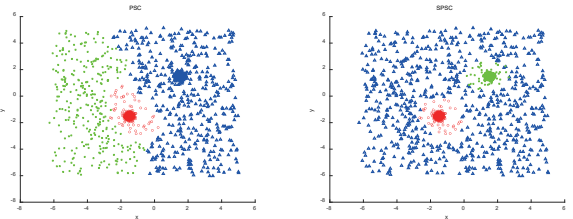
(a) TwoMoon 聚类效果



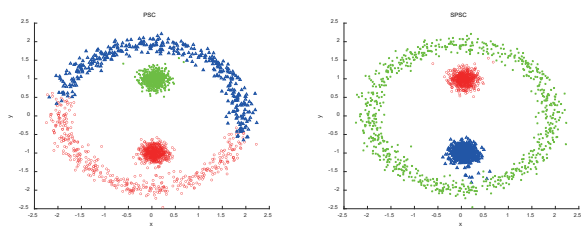
(b) ThreeCircle 聚类效果



(c) Smile 聚类效果



(d) Multi-scale 聚类效果



(e) Blobs and Circle 聚类效果

图 1 PSC 和 SPSC 在人工数据集上的检测效果

3.3.2 真实数据集验证结果

为了验证 SPSC 的聚类效果，对比 PSC 在真实数据集上 WDBC、GLASS，人脸数据集 UMIST、AT&T 上进行了对比试验，ACC 以及 NMI 实验结果如表 3 所示。

表 3 两种聚类方法在真实数据集上的聚类指标

数据集名称	评价指标	PSC	SPSC
WDBC	ACC	0.715 3	0.854 1
	NMI	0.276 5	0.467 2
GLASS	ACC	0.616 8	0.649 5
	NMI	0.348 3	0.426 1
UMIST	ACC	0.742 1	0.804 3
	NMI	0.806 6	0.856 9
AT&T	ACC	0.816 0	0.852 0
	NMI	0.862 7	0.891 3

从聚类效果看，SPSC 在真实数据集上的 ACC 和 NMI 都有一定程度的提升。

上述实验结果表明，本文提出的 SPSC 聚类算法效果较好，在聚类数据复杂且聚类类别较多时，能够使得聚类结果更加稳定均衡，对比 PSC 能够完成更为复杂的聚类任务。

4 结语

本文针对传统谱聚类算法在处理多维度、非线性复杂数据时存在的不足，提出了一种基于 p -Laplacian 的自适应多类聚类方法——SPSC。该方法不仅优化了权重矩阵的求解方式，而且创新性地采用了多个特征向量进行聚类，并结合投票式聚类原则，显著提升了多类聚类问题的稳定性和准确性。通过在多种人工及真实数据集上的实验验证，SPSC 在聚类效果上相较于传统 PSC 方法有了显著提升，在 ACC 和 NMI 等评价指标上表现优异。这表明 SPSC 能够有效应对复杂数据环境下的聚类挑战，为解决实际应用中的多类聚类问题提供了新的思路和解决方案。未来的工作将致力于进一步优化算法性能，探索其在更多领域中的应用潜力。

参考文献：

[1] OYEWOLE G J, THOPIL G A. Data clustering: application and trends[J]. Artificial intelligence review, 2023, 56: 6439-6475.

(下转第 167 页)

对比表 2、图 4 可以得出结论：设计方法调控后电力系统线损率不超过 2%，弃风光率不超过 2.5%，两项指标均明显优于两组对照方法，因此实验证明设计方法可以实现对电力系统负荷有效调控，具有良好的适应性与可靠性。

4 结语

本文在电力负荷智能调控中引入了深度学习算法，提出了一种新的调控方法，有效降低了电力系统线损率和弃风光率，提高了电力系统运行经济效益，为电力系统负荷智能化调控和相关系统的搭建提供了理论支撑，具有良好的现实意义和理论价值。

参考文献：

[1] 戴剑丰, 阎诚, 汤奕. 基于分布式通信架构的温控负荷参与电力系统频率调控模型 [J]. 电网技术, 2023, 47(5):1810-1818.

[2] 陈光宇, 杨锡勇, 江海洋, 等. 高比例新能源接入下计及工业负荷特性的电网需求响应调控策略 [J]. 电力自动化设备, 2023, 43(4):177-184.

(上接第 163 页)

[2] CAI J H, HAO J, YANG H F, et al. A review on semi-supervised clustering[J]. Information sciences, 2023, 632: 164-200.

[3] YANG M S, CHANG-CHIEN S J, HUNG W L. An unsupervised clustering algorithm for data on the unit hypersphere [J]. Applied soft computing, 2016, 42:290-313.

[4] 陈新泉, 周灵晶, 刘耀中. 聚类算法研究综述 [J]. 集成技术, 2017, 6(3): 41-49.

[5] 白璐, 赵鑫, 孔钰婷, 等. 谱聚类算法研究综述 [J]. 计算机工程与应用, 2021, 57(14): 15-26.

[6] BÜHLER T, HEIN M. Spectral clustering based on the graph p -Laplacian [C]//ICML'09: Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009: 81-88.

[7] ARORA J, KHATTER K, TUSHIR M. Fuzzy c-means clustering strategies: a review of distance measures [J]. Software engineering, 2018: 153-162.

[8] 谷留全, 柴瑞林, 王平心. 基于投票理论的三支聚类分析 [J]. 计算机科学与应用, 2019, 9(12): 2349-2356.

[9] STANKOVIĆ L, DAKOVIĆ M, BRAJOVIĆ M, et al. A p -Laplacian inspired method for graph cut [C/OL]//2019 27th Telecommunications Forum (TELFOR). Piscataway: IEEE, 2019 [2025-12-31]. <https://ieeexplore.ieee.org/document/8971207>. DOI:10.1109/TELFOR48224.2019.8971207.

[3] 王顺江, 任守东, 于博, 等. 智能电网下计及用户侧效益最优的负荷调控方法 [J]. 现代电力, 2023(1):117-124.

[4] 陆建宇, 吉斌, 张怀宇, 等. 新型电力系统的负荷调控电价补偿机制设计 [J]. 电力需求侧管理, 2024, 26(3):27-33.

[5] 孔德全, 韩忠修, 王颖, 等. 基于可调负荷的电网运行优化策略研究 [J]. 自动化与仪器仪表, 2024(1):108-111.

[6] 刘昊, 郝福忠, 贺翔, 等. 基于数据驱动的拓扑识别技术在台区负荷调控的应用 [J]. 电气自动化, 2024, 46(4):63-65.

[7] 张斌, 司大军, 李文云, 等. 计及多类型可调度柔性负荷响应的电力系统经济调度策略 [J]. 电工电能新技术, 2023, 42(4):39-47.

【作者简介】

贾蓉蓉 (1986—), 女, 陕西榆林人, 硕士研究生, 副高级工程师, 研究方向: 负荷管理。

雷瑞雪 (1990—), 女, 陕西西安人, 本科, 副高级工程师, 研究方向: 负荷管理。

(收稿日期: 2025-07-07 修回日期: 2025-12-31)

[10] LUO D J, HUANG H, DING C, et al. On the eigenvectors of p -Laplacian [J]. Machine learning, 2010, 81: 37-51.

[11] 周林, 平西建, 徐森, 等. 基于谱聚类的聚类集成算法 [J]. 自动化学报, 2012, 38(8): 1335-1342.

[12] DING S F, CONG L, HU Q K, et al. A multiway p -spectral clustering algorithm [J]. Knowledge-based systems, 2019, 164: 371-377.

[13] BOUTSIDIS C, KAMBADUR P, GITTENS A. Spectral clustering via the power method-provably [EB/OL]. (2015-05-12) [2025-12-31]. <https://doi.org/10.48550/arXiv.1311.2854>.

【作者简介】

张凯 (2000—), 男, 河南汝州人, 硕士研究生, 研究方向: 智能信息系统。

陶青川 (1972—), 通信作者 (email:taoqingchuan@scu.edu.cn), 男, 四川南充人, 博士, 副教授, 研究方向: 模式识别与智能系统。

(收稿日期: 2025-09-02 修回日期: 2026-01-08)