



相对密度和马氏距离的时空密度峰值聚类算法

赵嘉, 李辉, 肖人彬, 支妍力, 韩龙哲, 欧清海, 熊小舟

引用本文:

赵嘉, 李辉, 肖人彬, 等. 相对密度和马氏距离的时空密度峰值聚类算法[J]. 智能系统学报, 2026, 21(5): 1335-1347.

ZHAO Jia, LI Hui, XIAO Renbin, et al. Spatial-temporal density peaks clustering algorithm with relative density and mahalanobis distance[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1335-1347.

在线阅读 View online: <https://dx.doi.org/10.11992/tis.202511002>

您可能感兴趣的其他文章

结合度量融合和地标表示的自编码谱聚类算法

An autoencoder-based spectral clustering algorithm combined with metric fusion and landmark representation
智能系统学报. 2020, 15(4): 687-696 <https://dx.doi.org/10.11992/tis.201911039>

加权PageRank改进地标表示的自编码谱聚类算法

An autoencoder spectral clustering algorithm for improving landmark representation by weighted PageRank
智能系统学报. 2020, 15(2): 302-309 <https://dx.doi.org/10.11992/tis.201904021>

基于可拓距的改进k-means聚类算法

Improved k-means algorithm based on extension distance
智能系统学报. 2020, 15(2): 344-351 <https://dx.doi.org/10.11992/tis.201811020>

基于时空约束密度聚类的停留点识别方法

Stay point recognition method based on spatio-temporal constraint density clustering
智能系统学报. 2020, 15(1): 59-66 <https://dx.doi.org/10.11992/tis.201910026>

多层递阶融合模糊特征映射的模糊C均值聚类算法

Fuzzy C-means clustering algorithm for multilayered hierarchical fusion fuzzy feature mapping
智能系统学报. 2018, 13(4): 594-601 <https://dx.doi.org/10.11992/tis.201703047>

多视角模糊双加权可能性聚类算法

Multi-view fuzzy double-weighting possibility clustering algorithm
智能系统学报. 2017, 12(6): 806-815 <https://dx.doi.org/10.11992/tis.201703031>

DOI: 10.11992/tis.202511002

网络出版地址: <https://link.cnki.net/urlid/23.1538.TP.20260324.0843.002>

相对密度和马氏距离的时空密度峰值聚类算法

赵嘉^{1,2}, 李辉^{1,2}, 肖人彬³, 支妍力⁴, 韩龙哲¹, 欧清海⁵, 熊小舟⁶

(1. 江西水利电力大学 信息工程学院, 江西 南昌 330099; 2. 江西省水利大数据智能处理与预警技术工程研究中心, 江西 南昌 330099; 3. 华中科技大学 人工智能与自动化学院, 湖北 武汉 430074; 4. 国网江西省电力有限公司, 江西 南昌 330001; 5. 北京中电普华通信有限公司, 北京 100070; 6. 国网江西省电力有限公司 信息通信分公司, 江西 南昌 330095)

摘要: 快速搜索密度峰值的时空聚类 (spatial-temporal clustering by fast search and find of density peaks, ST-CFSF-DP) 算法在处理时空数据时未充分考虑样本间密度差异, 导致样本的局部密度不能正确反映样本的密度分布; 通过欧氏距离计算样本的相对距离, 无法将时间邻近的簇进行区分。针对以上问题, 提出相对密度和马氏距离的时空密度峰值聚类 (spatial-temporal density peaks clustering algorithm with relative density and mahalanobis distance, ST-DPC-RDMD) 算法。ST-DPC-RDMD 算法在局部密度中引入时空近邻概念, 并结合相对密度思想重新定义样本局部密度, 提高近邻样本的影响程度; 使用空间与时间属性的马氏距离替代欧氏距离, 综合分析各维度的协方差关系实现空间与时间距离权重的合理分配, 以准确识别具有特征差异的簇。在模拟数据集和真实地震数据集上的实验结果表明, ST-DPC-RDMD 算法的聚类效果优势显著。

关键词: 时空聚类; 密度峰值; 聚类算法; 时空近邻; 相对密度; 马氏距离; 局部密度; 相对距离

中图分类号: TP181 **文献标志码:** A **文章编号:** 1673-4785(2026)05-1335-13

中文引用格式: 赵嘉, 李辉, 肖人彬, 等. 相对密度和马氏距离的时空密度峰值聚类算法 [J]. 智能系统学报, 2026, 21(5): 1335-1347.

英文引用格式: ZHAO Jia, LI Hui, XIAO Renbin, et al. Spatial-temporal density peaks clustering algorithm with relative density and mahalanobis distance[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1335-1347.

Spatial-temporal density peaks clustering algorithm with relative density and mahalanobis distance

ZHAO Jia^{1,2}, LI Hui^{1,2}, XIAO Renbin³, ZHI Yanli⁴, HAN Longzhe¹,OU Qinghai⁵, XIONG Xiaozhou⁶

(1. School of Information Engineering, Jiangxi University of Water Resources and Electric Power, Nanchang 330099, China; 2. Jiangxi Province Engineering Research Center for Intelligent Processing and Early Warning Technology of Water Conservancy Big Data, Nanchang 330099, China; 3. Institute of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China; 4. State Grid Jiangxi Electric Power Co., Ltd., Nanchang 330001, China; 5. Beijing Fibrlink Communications Co., Ltd., Beijing 100070, China; 6. Information and Telecommunication Branch, State Grid Jiangxi Electric Power Co. Ltd., Nanchang 330095, China)

Abstract: The spatial-temporal clustering by fast search and find of density peaks (ST-CFSFDP) fails to fully consider density differences among samples when processing spatial-temporal data, resulting in the inability of sample local density to accurately reflect density distribution. Additionally, using Euclidean distance to calculate relative distances between samples prevents the differentiation of clusters that are temporally adjacent but spatially distinct. To address these issues, a spatial-temporal density peaks clustering algorithm with relative density and Mahalanobis distance (ST-DPC-RDMD) is proposed. The ST-DPC-RDMD algorithm introduces the concept of spatial-temporal nearest neighbors into local density and redefines sample local density by incorporating the idea of relative density, thereby enhancing the influence of neighboring samples, amplifying the local density of samples in sparse clusters, and reducing the impact of inter-cluster density heterogeneity. It replaces Euclidean distance with Mahalanobis distance for spatial and temporal attributes, comprehensively analyzing covariance relationships across various dimensions to achieve a reasonable allocation of weights between spatial and temporal distances, thus accurately identifying clusters with characteristic differences. Experimental results on simulated datasets and real earthquake datasets demonstrate that the ST-DPC-RDMD algorithm exhibits significant advantages in clustering performance.

Keywords: spatial-temporal clustering; density peak; clustering algorithm; spatial-temporal nearest neighbors; relative density; Mahalanobis distance; local density; relative distance

收稿日期: 2025-11-03. 网络出版日期: 2026-03-24.

基金项目: 国家自然科学基金项目 (62066037).

通信作者: 赵嘉, E-mail: zhaojia925@163.com.

春秋时期著作中提出的“物以类聚, 人以群分”是聚类思想在认知层面应用的早期朴素体现^[1]。

随着时代的进步和科技的发展,人们开始应用聚类思想分析和解决问题,并随之演化出各种聚类算法。聚类算法已成为数据挖掘中重要的无监督学习手段^[2],它能够对无标签数据按特征进行分组,依据数据间的相似程度,将其划分成具有特定分布模式的簇^[3],使同一簇内数据相似度高,不同簇间数据相似度低^[4]。自 K-means 聚类算法提出至今仅约 60 年,聚类分析已发展为数据分析的核心分支^[5],其在模式识别^[6]、入侵检测^[7]、图像处理^[8]、市场分析^[9]等领域具有广泛应用价值与适用性。

2014 年, Rodrigues 等^[10]于《Science》提出密度峰值聚类(density peaks clustering, DPC)算法。该算法依据样本局部密度和相对距离的计算结果,选取具有较高局部密度且相对距离较大的样本作为簇心^[11];随后,将非簇心样本分配至与之距离最近的高密度样本所属簇^[12]。

随着物联网、遥感监测及定位技术的革新,时空大数据正以 ZB(zettabyte) 级规模呈海量集聚态势^[13]。在空间维度,其显著相关性与传统统计学中独立同分布假设相悖^[14];在时间维度,模式动态演化性导致基于静态模型的分析方法失效^[15]。时空聚类算法聚焦挖掘时空关联模式,成为突破传统聚类算法应用局限的重要路径,为应对时空数据特性提供了新的解决方案。其聚焦于从时空数据库中萃取具备时空相似性的对象集合,本质是传统聚类分析从空间维度向时空维度的延伸拓展^[16],在气候科学^[17]、社会科学^[18]及地球学^[19]等领域发挥关键作用。

近年来,时空聚类挖掘技术不断发展进步,大体上可以划分成时空扫描统计、时空密度聚类及时空混合距离 3 种类型^[20]。其中时空密度聚类方法把空间密度概念延伸到时空维度,将时空对象的局部密度当作相似性的衡量标准,致力于从具有多种密度的异质区域中提取出时空簇^[21]。时空密度聚类在聚类精度上优势突出,所需设置的参数相对较少且能识别任意形状的时空簇^[22]。典型的时空密度聚类算法如 ST-DBSCAN(spatial-temporal density-based spatial clustering of applications with noise)^[23]与 ST-OPTICS(spatio-temporal-ordering points to identify clustering structure)^[24],通过加入时间邻域半径,将 DBSCAN^[25]和 OPTICS^[26]算法的密度可达性扩展为时空密度连通性,允许识别任意形态的时空簇且无需预设簇数量,然而引入的三维参数空间显著增加了调参复杂度,其随机性易使噪声分配至时空簇中或忽略部分低密

度簇^[27]。

近期,在 DPC 算法基础上,王培晓等^[28]提出了快速搜索密度峰值的时空聚类(spatial-temporal clustering by fast search and find of density peaks, ST-CFSFDP)算法。其通过修改样本属性值的计算策略,得以高效且精准地识别出处于相同位置、不同时间的簇。ST-CFSFDP 算法继承了传统 DPC 算法运行效率高、参数设置简易^[29]的特性,相较于 ST-DBSCAN 和 ST-OPTICS 算法,其所需参数数量更少;而与 ST-GRID(spatial-temporal grid method)^[30]算法相比,该算法在聚类精度上表现更为优异。尽管如此,ST-CFSFDP 算法在处理密度差异较大的数据集时,基于截断核的局部密度仅计算域内的样本贡献,无法正确反映样本的真实密度分布,导致选取的簇心可靠性不足;仅依据时间属性计算相对距离,会造成空间范围较大簇的边缘样本出现分配偏差,可能会被错误划分至其他簇,亦或是被误判为异常样本,从而影响聚类结果的准确性。

基于此,本文提出相对密度和马氏距离的时空密度峰值聚类(spatiotemporal density peak clustering with relative sensity and Mahalanobis distance, ST-DPC-RDMD)算法,从局部密度与相对距离计算上对 ST-CFSFDP 算法进行优化。ST-DPC-RDMD 算法在计算样本局部密度时,首先利用样本时空近邻信息定义样本的初始密度,再结合相对密度思想将样本的初始密度与时空近邻内样本的初始密度之和相除计算出样本的局部密度,该方法计算出的局部密度增强了样本与其近邻样本之间的关系,同时放大了稀疏簇样本的局部密度,降低了簇间密度差异的影响;计算样本相对距离时,在马氏距离的基础上提出加权马氏距离,替代传统基于时间属性的计算模式,以此精准考量样本各维度特征相关性与数据尺度差异,提升聚类分析准确性。

1 ST-CFSFDP 算法

ST-CFSFDP 算法是在 DPC 算法的基础上进一步优化和扩展的产物。为了使 DPC 算法更好地适应时间序列数据集的特性,ST-CFSFDP 算法引入第二个超参数 t_c ,修改了局部密度 ρ_i 和相对距离 δ_i 的定义。样本的局部密度计算公式为

$$\rho_i = \sum_{i \neq j} \chi(d_{ij} - d_c, t_{ij} - t_c)$$

$$\chi(x, y) = \begin{cases} 1, & x < 0 \text{ and } y < 0 \\ 0, & \text{其他} \end{cases}$$

式中: d_{ij} 代表样本 x_i 与样本 x_j 的空间欧氏距离; d_c 为人为设定的空间截断距离; t_{ij} 代表样本 x_i 到样本 x_j 的时间距离; t_c 为人为设定的时间截断距离。此种定义方式称为截断核函数计算的样本局部密度。时间约束下的局部密度表示与样本 i 的空间距离小于 d_c 且时间距离小于 t_c 的样本个数。

ST-CFSFDP 算法的相对距离 δ_i 计算公式为

$$\begin{cases} \delta_i = \min_{j: \rho_j > \rho_i} (t_{ij}), & \rho_i \neq \max(\rho_1, \rho_2, \dots, \rho_n) \\ \delta_i = \max_{i \neq j} (\delta_j), & \rho_i = \max(\rho_1, \rho_2, \dots, \rho_n) \end{cases}$$

ST-CFSFDP 算法沿用了 DPC 算法确定簇心的方法, 通过各样本的局部密度 ρ_i 和相对距离 δ_i 的乘积确定决策属性值 γ_i , 属性值大的为初始簇心, 其计算公式为

$$\gamma_i = \rho_i \cdot \delta_i \quad (1)$$

为解决同一簇识别出多个簇心问题, ST-CFSFDP 算法提出在时间轴上两个相邻簇心的空间距离小于截断距离 d_c , 便将这两个簇心合并, 避免同一簇识别出多个簇心。

确定簇心后, 算法按时间轴分配剩余样本, 分配过程如图 1 所示, 其中红色圆点代表簇心, 绿色圆点代表已分配样本, 黑色圆点代表剩余样本, 黄色圆点代表不满足条件样本。以簇 C_2 为例, 对簇心 t_8 进行正向和反向搜索, 将满足空间距离条件的 t_7 和 t_9 样本划入簇 C_2 , 遇到不满足条件样本或其他簇心停止。时间轴首尾簇心分别接收其前、后样本, 搜索完成后, 未分配样本视为异常点。

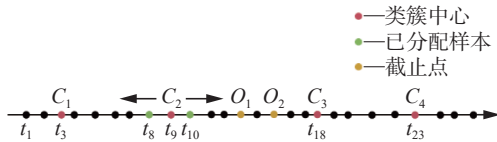


图 1 剩余样本分配过程

Fig. 1 Process of remaining sample allocation

2 ST-DPC-RDMD 算法

ST-CFSFDP 算法的局部密度计算忽略了样本的全局信息, 且未针对样本间的密度差异采取相应措施, 导致簇心的选取易发生偏差; 相对距离在处理时间上相互接近的簇时, 无法捕获簇间的空间差异。针对上述不足, 本文提出 ST-DPC-RDMD 算法, 重新定义局部密度和相对距离的计算方法, 以提高簇心选取及样本分配的准确率。

2.1 相对密度估计的局部密度

ST-CFSFDP 算法采用截断核函数计算样本局部密度, 因仅关注邻近域样本而忽视全局分布, 导致样本局部密度值趋同, 影响簇心识别。本文基于高斯核函数特性, 综合考虑邻近域与域外样

本对样本局部密度的影响, 重新定义局部密度的计算方式, 有效提升簇心识别精度。

首先定义样本 x_i 的时空邻域 $\text{STN}(x_i)$ 为

$$\text{STN}(x_i) = \{x_j \in D \mid d_{ij} \leq \max_{x_j \in N_K(x_i)} \{d_{ij}\} \cap t_{ij} \leq t_c\} \quad (2)$$

式中: D 为时空数据集; $N_K(x_i)$ 表示数据集中距离样本 x_i 最近的 K 个样本, K 为自定义参数。

定义 1 相对密度估计的局部密度。考虑到时间因素对样本局部密度的影响, 本文借助高斯核函数, 在充分考虑时间约束的前提下对样本局部密度进行精准计算, 相对密度估计的局部密度 ρ_i 的计算方式为

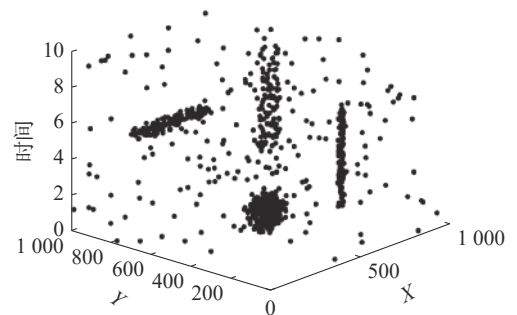
$$\bar{\rho}_i = \sum_{x_j \in \text{STN}(x_i)} \exp(-d_{ij}^2) + \exp(-t_{ij}^2) \quad (3)$$

$$\rho_i = \frac{\bar{\rho}_i}{\sum_{x_j \in \text{STN}(x_i)} \bar{\rho}_j} \quad (4)$$

式中: $\text{STN}(x_i)$ 表示样本 x_i 时空邻域内的样本集合; d_{ij} 表示样本 x_i 到样本 x_j 的空间距离; t_{ij} 表示样本 x_i 到样本 x_j 的时间距离; d_c 为空间截断距离; t_c 为时间截断距离; $\bar{\rho}_i$ 表示样本初始密度, 样本初始密度由其时空邻域内的局部样本信息得到; ρ_i 为样本的局部密度。

通过将样本的初始密度与其时空近邻样本初始密度之和的比值作为样本局部密度的计算方法, 可有效压缩密集区域样本的局部密度值, 同时增强稀疏区域样本的局部密度响应。该方法通过调整样本局部密度的计算范围, 显著提升稀疏区域样本在簇心选取过程中的优先级, 并弱化簇间密度异质性对中心判别结果的干扰。

如图 2 所示, 数据集 X 包含多组不同密度的时空簇, 不同颜色的六角星代表二维空间中所选簇心。ST-CFSFDP 算法采用截断核函数计算样本局部密度, 对密度稀疏的簇敏感性不足, 易导致识别不充分, 其结果如图 2(b) 所示, 在空间范围狭小但时间连续性较高的细长高密度柱状簇中识别出 2 个中心。基于相对密度估计的局部密度在 4 个簇中得到图 2(c) 中的结果, 成功识别出正确的簇心。



(a) 数据集 X 的时空分布

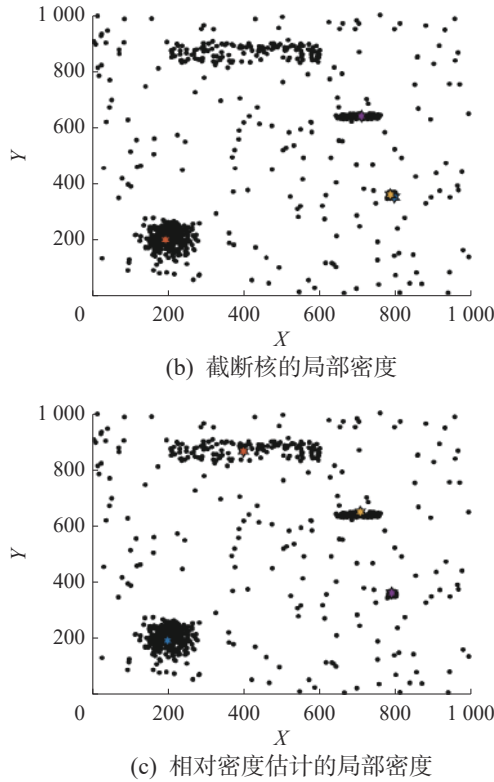


图 2 不同局部密度定义方式识别出的簇心

Fig. 2 Cluster centers identified by different local density definition methods

2.2 加权马氏距离定义的相对距离

针对 ST-CFSFDP 算法的相对距离仅考虑单一距离的影响而造成算法在某些情况下无法准确识别簇心的弊端,综合考虑时间和空间的共同影响,使用马氏距离代替原算法中的相对距离。

马氏距离是一种用于衡量样本间距离的指标,其优点在于考虑了各个特征之间的相关性。相比于欧氏距离,马氏距离不仅考虑各维度之间的联系,还受量纲的影响,能够更好地处理不同特征之间的相关性和数据的不同尺度。对于经度、纬度、时间三元组来说,其前两个空间属性属于地理属性,相关性较大,但时间属性与地理属性之间差异较大。使用马氏距离,可以忽略这些属性在数值和性质上的差异,将 3 种属性值更好地融合在一起,从而提高簇心识别的准确性。马氏距离公式为

$$D_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{y})} \quad (5)$$

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \quad (6)$$

式中: n 为样本总数, $\mathbf{x}$ 和 $\mathbf{y}$ 为不同样本的向量形式, $\boldsymbol{\mu}$ 是总体均值向量, $\mathbf{S}$ 是协方差矩阵, $\mathbf{S}^{-1}$ 是协方差矩阵的逆矩阵, $D_M(\mathbf{x}, \mathbf{y})$ 表示两样本间的马氏距离。

定义 2 维度方差。考虑不同维度的样本离散程度不统一,分别计算样本在二维空间和时间维度的方差,用于刻画样本的离散程度;

$$\begin{cases} \sigma_x^2 = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j \neq i} d_x(\mathbf{x}_i, \mathbf{x}_j) - \mu_x \right)^2 \\ \sigma_y^2 = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j \neq i} d_y(\mathbf{x}_i, \mathbf{x}_j) - \mu_y \right)^2 \\ \sigma_t^2 = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j \neq i} d_t(\mathbf{t}_i, \mathbf{t}_j) - \mu_t \right)^2 \end{cases} \quad (7)$$

式中: σ^2 表示各维度的方差, d 为样本在各维度上的欧氏距离, μ 代表各维度的均值。

定义 3 维度权重。在评估样本间关系时需分别考虑时间和空间维度的重要性。为此,定义维度权重 w_i , 以确保时空关系的合理度量:

$$w_i = 1 - \frac{2\sigma_c^2}{\sigma_x^2 + \sigma_y^2 + \sigma_t^2}, (i \in \{x, y, t\}) \quad (8)$$

由于方差反映维度数据的离散程度,方差大意味着数据分布更分散,对距离计算的干扰相对更强。因此,对方差较大的维度,赋予其更小权重;方差小的维度,则设计更大权重,以平衡不同维度对距离度量结果的影响。

定义 4 加权马氏距离。仅使用马氏距离仍无法充分反映时间属性与空间属性于不同场景数据集集中的重要性,本文进一步提出加权马氏距离,为时间属性和空间属性分配不同的权重,可更准确地反映出不同场景数据集中时间属性和空间属性的相对贡献度。加权马氏距离公式为

$$D_{WM}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{W} \mathbf{S}^{-1} \mathbf{W} (\mathbf{x} - \mathbf{y})} \quad (9)$$

$$\mathbf{W} = \begin{bmatrix} w_x & & \\ & w_y & \\ & & w_t \end{bmatrix} \quad (10)$$

式中: $\mathbf{W}$ 为归一化权重矩阵,其对角线元素 w_x 、 w_y 、 w_t 分别对应 3 个维度空间的权重系数。基于加权马氏距离的相对距离定义为

$$\begin{cases} \delta_i = \max_{i \neq j} (D_{WM}(\mathbf{x}_i, \mathbf{x}_j)), & \rho_i = \max(\rho) \\ \delta_i = \min_{j: \rho_j > \rho_i} (D_{WM}(\mathbf{x}_i, \mathbf{x}_j)), & \text{其他} \end{cases} \quad (11)$$

根据式 (4) 计算样本相对密度估计的局部密度 ρ_i , 式 (11) 计算样本加权马氏距离的相对距离 δ_i ; 根据式 (1) 计算 γ_i 并排序,选择前若干个样本作为簇心,再进行样本分配。

针对 ST-CFSFDP 算法在样本分配策略上的不足,尤其是在处理时间轴上第一个簇心之前的样本和最后一个簇心之后的样本时只能分别划分到第一个簇和最后一个簇,导致异常点不能被有效识别而被错误划分,影响聚类效果。以图 3 为

例,其相应的分配流程如图4所示。以簇心 C_1 为起点向右进行搜索,由于样本 t_6 与 C_1 之间的空间距离相对较短,符合聚类规则中距离判定标准,因此该样本被划分至该参照样本所在的簇当中。然而,由于样本 t_7 与下一簇心 C_2 的空间距离更为接近,下一轮分配任务将直接从 C_2 开始向左执行,而不再对位于 t_7 后的样本做进一步分配,导致原本应归属于 C_1 的样本 t_8 和 t_9 无法被正确分配。

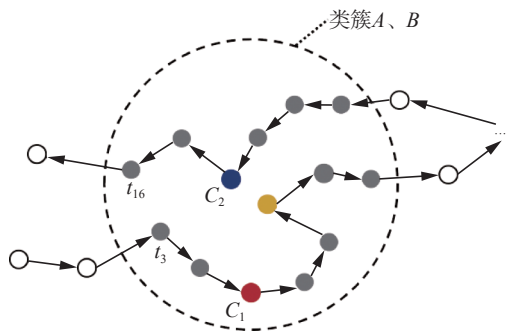


图3 空间分布不均样本的排列情况

Fig. 3 Arrangement of samples with uneven spatial distribution

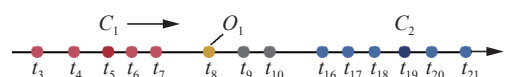


图4 空间分布不均样本时间轴上的排列情况

Fig. 4 Arrangement of samples with uneven spatial distribution on the time axis

为解决这一问题,本研究采用就近分配策略。该策略旨在降低异常点的不利影响,确保所有样本都被合理划分到相应的簇中。

2.2.1 就近分配

对非簇点集合 $(x_1, x_2, \dots, x_n)$,从中选取一个点 x_i ,先寻找所有局部密度比它大的点集,再确定与 x_i 加权马氏距离最近的那个点 x_j ,将 x_i 分配给 x_j 所在的簇。这种处理方式的核心逻辑在于高密度点往往处于簇心位置或者紧邻簇心,它们能够很好地代表所在簇的特征,所以更适合作为簇的代表性点。本文采用的就近分配策略通过寻找距离剩余样本最近的高密度点,并将其分配到该高密度点所在的簇,尽可能保证所有剩余样本都能准确地划分到各自所属的簇中。这种策略有效减少了样本错分的情况,使聚类结果能够涵盖更多有效信息;同时,也极大地提升了聚类结果的准确性,让各个簇能够更精准地反映数据的真实分布情况。

2.2.2 异常识别

当所有样本完成分配,使用局部离群因子识别异常样本,若样本所处环境的聚集程度明显低于周围邻域内的其他样本,便可判定其为异常。

定义5 局部可达距离。对于样本 x_i ,其局部

可达距离基于时空邻域 $STN(x_i)$ 得到。先计算 x_i 与时空邻域内所有样本 x_j 的加权马氏距离的平均值,再取该平均值的倒数作为当前样本的局部可达距离,即

$$LRD(x_i) = \frac{1}{\text{mean} \left(\sum_{x_j \in STN(x_i)} D_{WM}(x_i, x_j) \right)} \quad (12)$$

式(12)取时空邻域内所有样本到 x_i 的加权马氏距离的平均值倒数,以量化 x_i 周围区域的聚集程度,数值越大说明该区域内样本分布越密集。

定义6 局部离群因子。样本 x_i 的局部离群因子为其时空邻域 $STN(x_i)$ 内所有样本局部可达距离的均值与自身局部可达距离的比值:

$$LOF(x_i) = \frac{\text{mean} \left(\sum_{x_j \in STN(x_i)} LRD(x_j) \right)}{LRD(x_i)} \quad (13)$$

局部离群因子反映的是样本 x_i 与邻域内其他样本所处环境的相对差异。若样本的局部离群因子 $LOF(x_i)$ 接近1,表明其与时空邻域内样本周围的聚集程度处于相近水平,不太可能属于异常样本;当 $LOF(x_i)$ 远大于1时,证明 x_i 周围的聚集程度明显低于其时空邻域内的其他样本,由此可判定该样本表现为异常^[31]。

2.3 算法步骤

结合前文,给出ST-DPC-RDMD算法的流程如下。

1) 数据预处理与距离矩阵构建。采用min-max归一化对数据 X 进行归一化,得到 X_{norm} ;基于欧氏距离计算样本空间距离,构建空间距离矩阵 D_s ;基于绝对差计算样本时间距离,构建时间距离矩阵 D_t 。

输入: 时空数据序列 $X = \{t_i, x_i, y_i\}$ 。

输出: 归一化后的时空数据 X_{norm} 、空间距离矩阵 D_s 、时间距离矩阵 D_t 。

2) 局部密度计算。结合时空距离特征,通过式(2)~(4)计算每个样本的局部密度 ρ_i ,其中式(2)定义时空融合距离范式,式(3)算样本间融合距离,式(4)统计样本内数量并转化为局部密度。

输入: 归一化后的时空数据、距离阈值 d_c 、时间阈值 t_c 、时空近邻样本数 k 。

输出: 样本局部密度集合 $\{\rho_i\}$ 。

3) 加权马氏距离计算。首先筛选出局部密度大于当前样本的样本,然后基于加权马氏距离计算当前样本与这些密度样本的距离,通过式(5)~(11)完成相对距离 δ_i 的计算,其中式(5)定义加

权马氏距离公式, 式 (6)~(10) 求解协方差矩阵及逆矩阵, 式 (11) 确定加权马氏距离作为相对距离。

输入: 样本局部密度集合 $\{\rho_i\}$ 、空间距离矩阵 D_s 、时间距离矩阵 D_t 。

输出: 样本相对距离集合 $\{\delta_i\}$ 。

4) 决策图绘制与簇心选择。以局部密度 ρ_i 为横轴、相对距离 δ_i 为纵轴绘制决策图; 根据式 (1) 样本的决策属性值 $\gamma_i = \rho_i \times \delta_i$, 设定 γ 阈值, 将决策属性值大于阈值的样本选为核心。

输入: 样本局部密度集合 $\{\rho_i\}$ 、样本相对距离集合 $\{\delta_i\}$ 。

输出: 类簇中心集合 $C_{\text{center}} = \{C_1, C_2, \dots, C_m\}$ 。

5) 样本归属确定与异常检测。剩余非核心样本依据距离最近的核心所在类簇确定归属; 之后基于式 (12)~(13) 计算各样本的局部离群因子, 将局部离群因子远大于 1 的样本从聚类结果中剔除, 标记为异常样本, 形成异常样本集合 O , 其中式 (12) 计算局部可达距离, 式 (13) 推导局部离群因子。

输入: 类簇中心集合 $C_{\text{center}} = \{C_1, C_2, \dots, C_m\}$ 、样本相对距离集合 $\{\delta_i\}$ 。

输出: 初步结果 C 、异常样本集合 O 。

6) 结果输出。整理聚类的样本簇、核心坐标等信息, 明确异常样本的类别及特征值, 以结构化形式输出最终结果。

输入: 初步结果 C 、异常样本集合 O 。

输出: 样本聚类结果 $c = \{c_i\}$ 。

2.4 算法时间复杂度分析

设数据集规模为 n , 簇个数为 c , 现对 ST-DPC-RDMD 算法时间复杂度进行分析: 1) 对数据按时间升序排序, 时间复杂度为 $O(n \log n)$; 2) 对数据集的每一对样本进行距离计算并构建空间距离矩阵和时间距离矩阵, 时间复杂度为 $O(n^2)$; 3) 获取各样本的时空近邻集合并计算其局部密度, 需要遍历数据集中的每个样本以及其近邻样本, 该过程的时间复杂度为 $O(n^2)$; 4) 计算各样本的加权马氏距离, 时间复杂度也为 $O(n^2)$; 5) 绘制决策图, 选取密度峰值作为簇心, 时间复杂度为 $O(n \log n)$; 6) 将剩余的非核心样本分配到相应的簇中, 时间复杂度为 $O(cn)$ 。由于 $c \ll n$, 即 ST-DPC-RDMD 算法的时间复杂度为 $O(n^2)$ 。

2.5 算法空间复杂度分析

设数据集规模为 n , 簇个数为 c , ST-CFSFDP 算法的空间复杂度为 $O(n^2)$ 。现对 ST-DPC-RDMD 算法空间复杂度进行分析: 1) 构建并存储样本间的空间距离矩阵与时间距离矩阵, 空间复杂度为

$O(n^2)$; 2) 存储归一化样本数据, 空间复杂度为 $O(n)$; 3) 存储时空邻域集合, 空间复杂度为 $O(n)$; 4) 存储簇心标记以及样本归属关系数据, 空间复杂度为 $O(n)$ 。即 ST-DPC-RDMD 算法空间复杂度同为 $O(n^2)$, 与 ST-CFSFDP 算法的空间复杂度为同一量级。

3 实验结果与分析

3.1 实验设置

为探究 ST-DPC-RDMD 算法的聚类性能, 选用 5 个不同规模和簇数量的时序数据集展开实验。这 5 个数据集涵盖了多种数据特征, 能有效模拟出不同应用场景的真实情况。

实验选取了 ST-DBSCAN^[23]、ST-OPTICS^[24]、ST-AGNES(spatial-temporal agglomerative nesting)^[32]、STSN-DPC(spatio-temporal shared nearest neighbors and density peaks based clustering)^[33] 以及 ST-CFSFDP^[28] 算法, 与 ST-DPC-RDMD 算法进行对比分析, 通过对比这些主流算法, 从多个维度对 ST-DPC-RDMD 算法的性能进行评估, 清晰展现出该算法在聚类准确性、稳定性等方面的优势与不足, 为算法的进一步优化和实际应用提供有力依据。实验所选的 5 种数据集具有不同的特征信息, 其生成过程如下: 先确定簇心的数量和形状, 并在空间中指定它们的位置, 然后在空间中加入噪声点使数据更贴近现实世界的情况, 最终聚合此类数据后模拟生成数据集。表 1 给出了这些数据集的基本信息。其中 D_1 数据集为 ST-CFSFDP 算法中所用数据集; 数据集 $D_2 \sim D_5$ 具有相同的时间段内包含多个不同簇的特点, 用于对比评估 ST-DPC-RDMD 算法在时间密集型数据集上的聚类性能。真实数据集 E 研究区域界定于东经 $73^\circ \sim 125^\circ$ 、北纬 $18^\circ \sim 48^\circ$ 范围内, 检验 ST-DPC-RDMD 算法在复杂实际场景下的性能表现, 数据集详细信息见表 2。所有聚类算法的实验环境均为: Windows 10 的 64 位操作系统、11th Gen Intel(R) Core(TM) i5-1035G1 @ 1.00 GHz CPU、8 GB 内存。

表 1 模拟数据集信息
Table 1 Information of simulated datasets

数据集	实例数	维度	簇个数
D_1	314	3	4
D_2	1 550	3	5
D_3	350	3	2
D_4	1 153	3	3
D_5	300	3	2

表 2 真实数据集信息
Table 2 Information of real dataset

发震日期	经度	纬度	震级	震中位置
2008-01-05 13:45:05	102.118	33.889	3.9	甘肃玛曲
2008-01-17 23:07:05	100.222	38.12	4.9	青海祁连
...	...	...	...	...
2009-11-23 07:25:22	103.386	30.959	4.6	四川汶川
2009-11-27 16:04:05	103.787	31.186	4.9	四川彭州

在评估算法聚类性能时, 本文采用了 3 种在学术领域广泛认可的外部评估指标: 调整互信息 (adjusted mutual information, AMI)^[34]、调整兰德系数 (adjusted Rand index, ARI)^[34] 以及 FM 指数 (Fowlkes-Mallows index, FMI)^[35]。这 3 个指标的满分均为 1, 其数值大小与聚类结果的精确度呈正相关, 指标值越接近 1, 意味着聚类结果越精确^[36]。实验过程中, 为确保每种算法在不同数据集上都能达到最佳聚类效果, 本文针对各算法设定了合理的参数取值范围。通过逐步改变参数值, 并进行一系列的测试, 对每次测试的聚类结果进行分析比较, 最终筛选出每个算法在相应数据集上的最优参数组合, 以此获得最佳的聚类结果。

为直观分析聚类结果, 通过绘制二维空间散

点图, 对聚类结果进行可视化处理。分析过程中, 以半径 r 内切圆覆盖样本, 定义 r 最大值为 $s_{\max}$, 记录各簇垂直高度最大值 $t_{\max}$ 。ST-DBSCAN 与 ST-OPTICS 算法的空间邻域参数的取值区间设为 $[0.1, s_{\max}]$, 时间邻域参数的取值范围是 $[0.1, t_{\max}/2]$, 步长 0.1, 邻域内最少点数的取值为数据集规模 n 的对数 $\ln n$; ST-CFSFDP 和 ST-DPC-RDMD 算法在空间截断距离 d_c 的取值范围是 $[0.1, s_{\max}]$, 时间截断距离 t_c 的取值范围为 $[0.1, t_{\max}/2]$, 步长同样为 0.1。ST-AGNES 算法距离参数的取值区间是 $[1, n/2]$, 步长为 1。

3.2 模拟数据集的实验结果分析

表 3 给出了 6 种算法在模拟数据集上的聚类性能对比结果, 数据中标注加粗的数值代表同一数据集下各算法中最优的聚类效果评估指标值。从表 4 的统计数据可以发现, ST-DPC-RDMD 算法在数据集 D_1 、 D_2 、 D_3 及 D_4 这 4 个样本集合上的聚类表现均优于对比方法。紧随其后的是 ST-OPTICS 和 ST-DBSCAN 算法, 其中 ST-OPTICS 算法在数据集 D_1 、 D_2 、 D_3 及 D_5 上取得了第二优的成绩; ST-DBSCAN 算法则在数据集 D_5 上表现最为突出, 在数据集 D_4 上获得次优效果。相较之下, ST-CFSFDP 和 ST-AGNES 算法的聚类效果与前三者存在明显差距, 在各数据集上的综合表现欠佳。

表 3 6 种算法在模拟数据集上的聚类效果对比
Table 3 Comparison of clustering effects of six algorithms on simulated datasets

数据集	评价指标	STSN-DPC	ST-CFSFDP	ST-OPTICS	ST-DBSCAN	ST-AGNES	ST-DPC-RDMD
D_1	AMI	0.7517	0.6806	0.9105	0.8940	0.7283	0.9246
	ARI	0.7752	0.6686	0.9269	0.8836	0.7645	0.9311
	FMI	0.8326	0.7440	0.9424	0.9082	0.8249	0.9457
D_2	AMI	0.7828	0.2103	0.8777	0.8098	0.1579	0.8860
	ARI	0.8103	0.0630	0.9035	0.7906	0.0996	0.9102
	FMI	0.8494	0.4388	0.9205	0.8274	0.3206	0.9260
D_3	AMI	0.5005	0.4729	0.7399	0.6493	0.3118	0.8162
	ARI	0.6595	0.5548	0.8200	0.7229	0.3742	0.8955
	FMI	0.8282	0.7414	0.8952	0.8353	0.5938	0.9412
D_4	AMI	0.6764	0.1783	0.9689	0.9813	0.1490	0.9835
	ARI	0.7329	0.1446	0.9875	0.9920	0.1496	0.9937
	FMI	0.8213	0.4914	0.9913	0.9945	0.4228	0.9954
D_5	AMI	0.2902	0.2128	0.8732	0.8972	0.1427	0.8108
	ARI	0.3217	0.1538	0.9124	0.9220	0.1084	0.8490
	FMI	0.6133	0.4401	0.9414	0.9479	0.4395	0.8992

注: 加粗表示最优结果。

表 4 6 种算法在合成数据集上各项聚类评估指标的秩均值
Table 4 Mean ranks of various clustering evaluation metrics for six algorithms on synthetic datasets

算法	秩均值		
	AMI	ARI	FMI
ST-SNN-DPC	3.0	2.4	2.4
ST-CFSFDP	1.8	1.8	2.2
ST-OPTICS	4.8	4.8	4.8
ST-DBSCAN	4.6	4.4	4.4
ST-AGNES	1.2	2.0	1.6
ST-DPC-RDMD	5.6	5.6	5.6

注: 加粗表示最优结果。

为检验 ST-DPC-RDMD 算法在常规时空数据集方面的适用性, 分析 6 种算法在模拟数据集的聚类表现情况, 从而评估 ST-DPC-RDMD 算法在该数据集上相对于其他算法的优势与不足。

图 5~7 为数据集 D_1 、 D_2 和 D_3 绘制的聚类结果对比图, 其中, 黑色轮廓的六角星符号代表算法所得的各簇的簇心, 而不同颜色的数据点则组成了不同的簇。

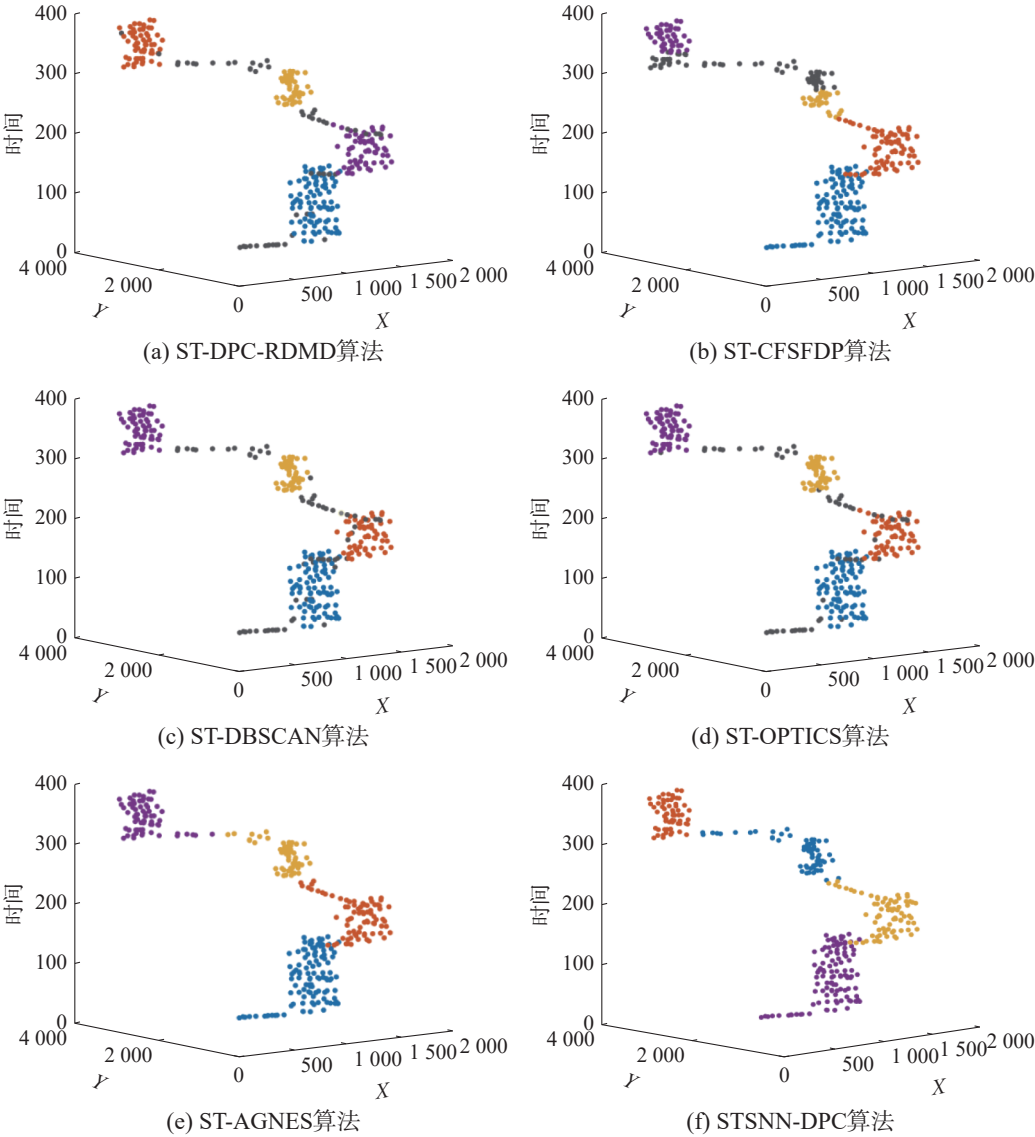
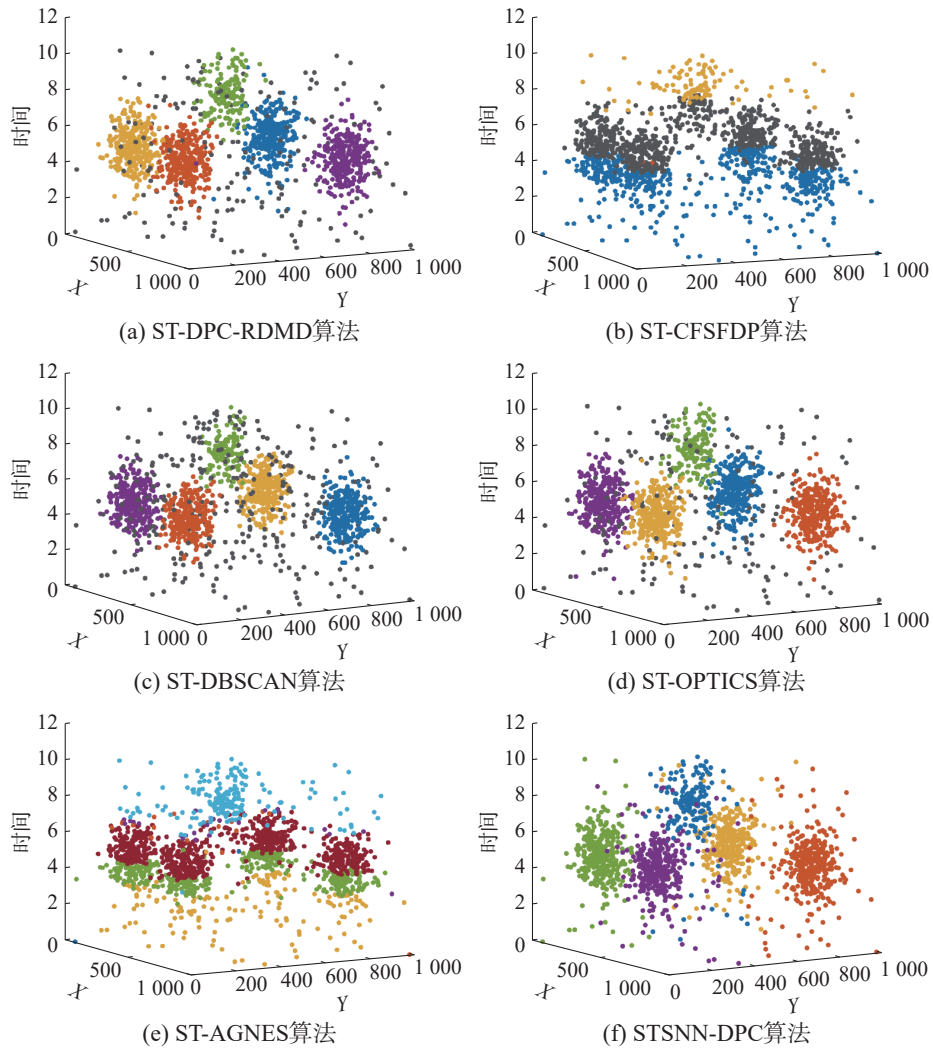
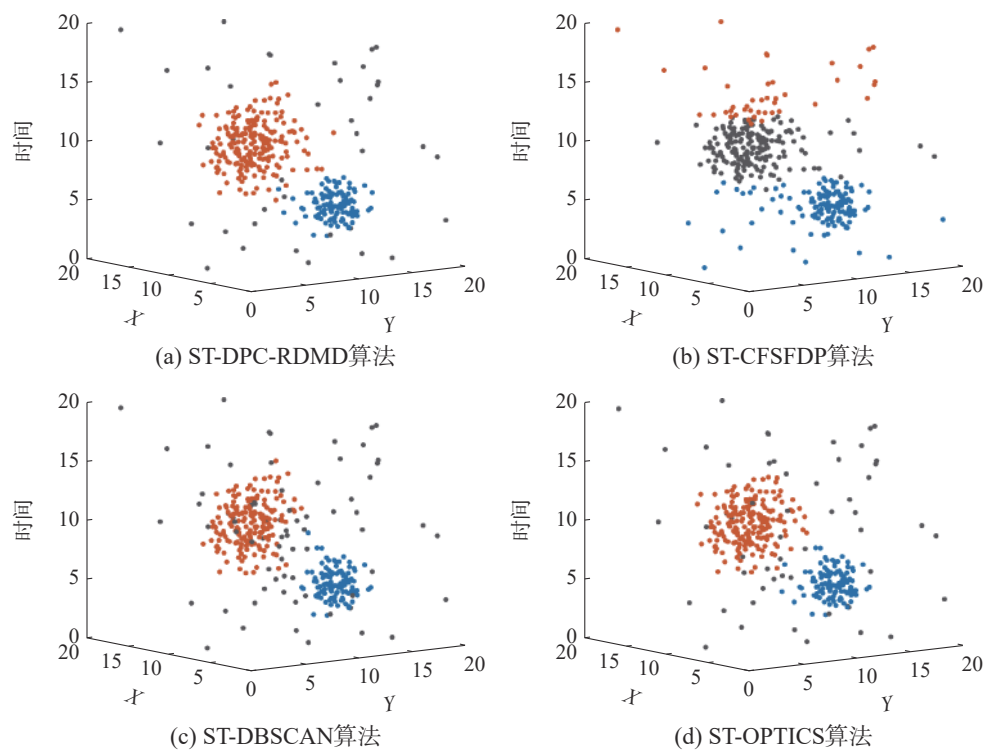
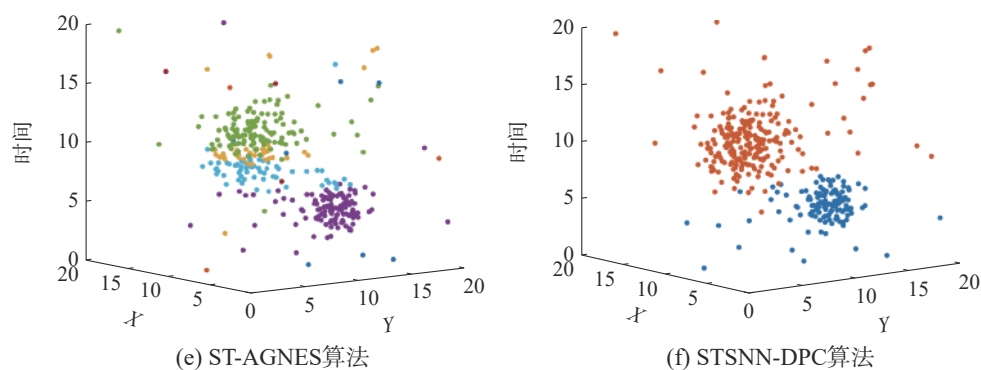


图 5 6 种算法在数据集 D_1 上的时空聚类结果
Fig. 5 Clustering results of six algorithms on dataset D_1

图 6 6 种算法在数据集 D_2 上的聚类结果Fig. 6 Clustering results of six algorithms on dataset D_2 

图 7 6种算法在数据集 D_3 上的聚类结果Fig. 7 Clustering results of six algorithms on dataset D_3

数据集 D_1 由5个不同规模的簇以及两两之间相互连接的线型异常点构成,其中各簇在时间属性上展现出了显著的区分度。根据图5所示的结果,ST-DPC-RDMD和ST-CFSFDP算法均能准确识别出簇心。但ST-CFSFDP算法在处理黄色簇时,不恰当地将部分样本误判为异常。ST-AGNES算法在异常点的识别上存在缺陷,错误地将所有异常点纳入了各簇中。ST-OPTICS、STSNN-DPC和ST-DBSCAN算法在处理簇边界时出现错误分配,使得其性能指标低于ST-DPC-RDMD算法,证明后者相对密度机制考虑了邻域内所有样本的贡献,在聚类精度上存在优势。

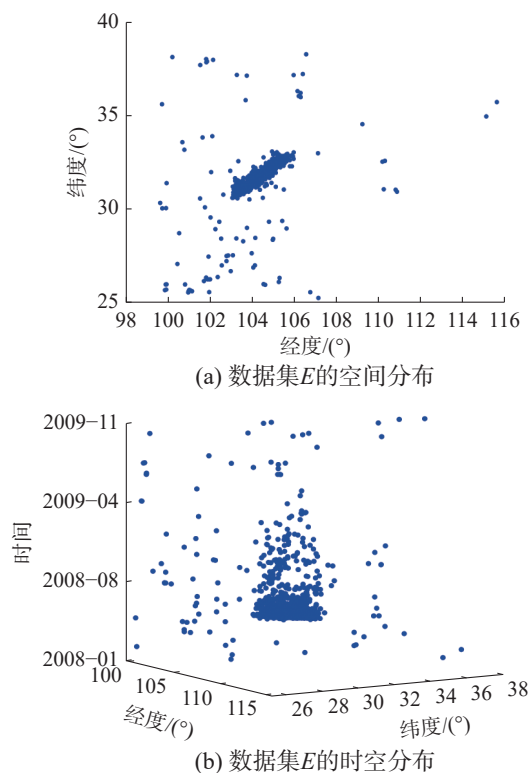
数据集 D_2 由5个簇组合而成,其中下方的4个簇呈现出时间上紧密相邻、空间上彼此独立的分布特性,观察图6的聚类结果,可以发现ST-CFSFDP算法在聚类上存在明显不足。基于决策图完成5个簇心的初步筛选后,经簇心合并处理,最终仅生成3个簇心,这一现象源于分配策略缺陷。ST-AGNES算法在聚类时,仅关注了空间维度的分割,而未充分考虑时间维度的关键作用,同样导致了聚类结果的错误。ST-OPTICS和ST-DBSCAN算法由于未能提供独立的异常点判别机制,其聚类表现相对较弱。STSNN-DPC算法虽然表现不错,但相较之下,ST-DPC-RDMD算法通过相对化处理放大了稀疏簇核心样本的密度响应、压缩了密集簇边缘样本的密度值,使簇心识别更贴合数据真实分布,在全局聚类水平上展现了更为出色的性能。

数据集 D_3 由2个大小与密度均有差异且相互邻近的簇构成,然而,从图7的聚类分析中可以看出,ST-CFSFDP算法在簇心选择和后续的合并过程中未能精确地执行任务,导致2个本应分离的簇未能被正确划分。ST-AGNES算法虽然以时间序列为基础进行聚类,但其输入参数并未考量时

间维度,导致其分配策略与ST-CFSFDP算法出现了类似的错误,从而使得其整体性能指标较低。ST-OPTICS、STSNN-DPC和ST-DBSCAN算法在面对数据集中2个簇空间上的紧密邻近时,也未能有效地对边界区域的样本进行正确归类,而ST-DPC-RDMD算法中提出的加权马氏距离通过协方差矩阵捕捉“经度、纬度、时间”的维度关联,结合动态维度权重平衡各维度贡献进行正确聚类。

3.3 真实数据集的实验结果分析

如图8所示,数据集 E 详细记录了2008年1月5日13时45分05秒至2009年11月27日16时04分05秒这两年时间内,发生在以中国四川为主体区域内、震级大于3.0级的共计1090起地震事件。

图 8 数据集 E 的可视化结果Fig. 8 Visualization results of dataset E

将 ST-DPC-RDMD 算法应用于地震数据集 E , 聚类结果如图 9 所示。通过聚类分析生成了 3 个层次清晰的时空簇, 依据时间递增顺序, 依次标记为橙色簇 C_1 、黄色簇 C_2 和蓝色簇 C_3 。这 3 个地震簇所涵盖的地震事件信息详见表 5。簇 C_1 以 2008 年 5 月 12 日震级为 7.9 级的四川汶川地震, 其后发生的 879 次地震认定为汶川地震的余震, 符合实际情况中主震后余震频发的特征。簇 C_2 中的地震事件大多发生于四川省内的平武、汶川、彭州等地, 2009 年 3 月 12 日于四川青川发生的震级为 5.3 级的地震较为突出, 此前的 170 次地震判定为前震, 后续还引发了 5 次后震。簇 C_3 中的强震位置为四川什邡, 发生时间为 2009 年 6 月 29 日, 其后续识别出的 11 次地震事件均为余震。

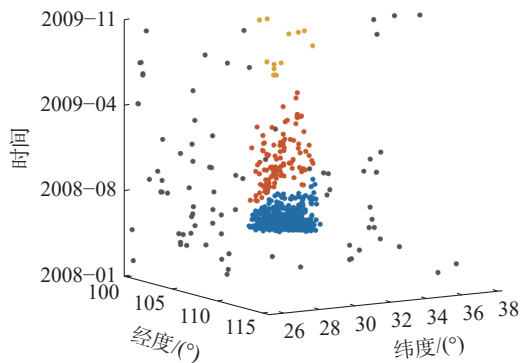


图 9 ST-DPC-RDMD 算法在数据集 E 上的聚类结果

Fig. 9 Clustering results of the ST-DPC-RDMD algorithm on dataset E

表 5 数据集 E 的聚类结果统计信息

Table 5 Statistical information of clustering results for dataset E

簇	地震数量	地震名称	强震位置	震级/级
C_1	880	汶川地震	31.002°N, 103.322°E	7.9
C_2	108	青川地震	32.386°N, 105.096°E	5.3
C_3	14	什邡地震	31.426°N, 104.012°E	5.3

通过对 3 个簇的空间分布以及表中统计的信息进行分析可知, 在 2008 年 1 月至 2009 年 11 月时间跨度内, 中国境内的地震活动呈现出显著的空间聚集性, 主要集中于四川地区。从地质构造角度阐释, 四川地处地中海至喜马拉雅地震带, 该区域受欧亚板块与印度洋板块相互作用的影响极为强烈^[37]。两大板块的持续挤压与碰撞, 致使该地区地质构造异常复杂, 断裂带广泛发育, 龙门山断裂带便是其中典型代表^[38]。这种特殊的地质构造背景, 为地壳能量的不断积累与突然释放创造了条件, 进而使四川地区成为地震频发的区域。

4 结束语

ST-CFSDP 算法在处理时空数据时局部密度无法反映密度不均样本的分布状况; 在空间范围较大的数据集中, 沿时间轴的分配策略致使边缘样本因时间特征不突出而被错误分配。为此, 本文提出相对密度和马氏距离的时空密度峰值聚类算法。该算法引入时空近邻并结合相对密度思想重新定义样本局部密度, 通过对比每个样本与其邻域内其他样本的关系, 全面衡量样本的局部密度, 避免相同局部密度值样本过多问题; 提出加权马氏距离对时间和空间维度进行归一化计算, 有效衡量经度、纬度、时间 3 种属性在样本特征中的贡献程度, 保证算法对样本的合理聚类。实验结果表明, ST-DPC-RDMD 算法在处理时空数据集时展现出优异的性能。后续研究将进一步探索群智能算法在参数优化中的应用, 以聚类评价指标作为适应度函数, 逐步寻找空间截断距离 d_c 和时间截断距离 t_c 的最优组合, 以减少人工干预; 同时尝试引入近邻搜索技术改进样本距离的计算流程, 通过构建 K-D 树或球树等空间索引结构加速距离求解, 降低算法时间复杂度, 提升其在大规模时空数据集上的处理效率。

参考文献:

- [1] SCHWYCK M E, DU Meng, LI Yuchen, et al. Similarity among friends serves as a social prior: the assumption that “birds of a feather flock together” shapes social decisions and relationship beliefs[J]. *Personality & social psychology bulletin*, 2024, 50(6): 823–840.
- [2] 张远鹏, 周洁, 邓赵红, 等. 代表点一致性约束的多视角模糊聚类算法[J]. *软件学报*, 2019, 30(2): 282–301.
ZHANG Yuanpeng, ZHOU Jie, DENG Zhaohong, et al. Multi-view fuzzy clustering approach based on medoid invariant constraint[J]. *Journal of software*, 2019, 30(2): 282–301.
- [3] 吕莉, 朱梅子, 康平, 等. 面向密度分布不均数据的混合近邻密度峰值聚类算法[J]. *控制理论与应用*, 2024, 41(10): 1821–1830.
LÜ Li, ZHU Meizi, KANG Ping, et al. Multiplex neighbor density peaks clustering for uneven density data sets[J]. *Control theory & applications*, 2024, 41(10): 1821–1830.
- [4] 李智冈, 吕莉, 谭德坤, 等. 基于加权核密度估计与微簇合并的密度峰值聚类算法[J]. *信息与控制*, 2024, 53(3): 302–314.
LI Zhigang, LÜ Li, TAN Dekun, et al. Density peaks clustering algorithm based on weighted kernel density es-

- timation and micro-cluster merging[J]. *Information and control*, 2024, 53(3): 302–314.
- [5] JAIN A K. Data clustering: 50 years beyond K-means[J]. *Pattern recognition letters*, 2010, 31(8): 651–666.
- [6] SINGH J, SINGH D. A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects[J]. *Advanced engineering informatics*, 2024, 62: 102799.
- [7] AHMED U, NAZIR M, SARWAR A, et al. Signature-based intrusion detection using machine learning and deep learning approaches empowered with fuzzy clustering[J]. *Scientific reports*, 2025, 15: 1726.
- [8] WANG Nian, CUI Zhigao, LI Aihua, et al. Structured doubly stochastic graph-based clustering[J]. *IEEE transactions on neural networks and learning systems*, 2025, 36(6): 11064–11077.
- [9] THAKUR J, KUSHWAHA B P. Artificial intelligence in marketing research and future research directions: Science mapping and research clustering using bibliometric analysis[J]. *Global business and organizational excellence*, 2024, 43(3): 139–155.
- [10] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. *Science*, 2014, 344(6191): 1492–1496.
- [11] ZHAO Jia, WANG Gang, PAN J S, et al. Density peaks clustering algorithm based on fuzzy and weighted shared neighbor for uneven density datasets[J]. *Pattern recognition*, 2023, 139: 109406.
- [12] LI Qingpeng, HU Xinyue, ZHAO Jia, et al. Density peak clustering algorithm based on shared neighbors and natural neighbors and analysis of electricity consumption patterns[J]. *Concurrency and computation: practice and experience*, 2025, 37(3): e8387.
- [13] 陈芳森, 黄慧萍, 贾坤. 时空大数据在城市群建设与管理中的应用研究进展[J]. *地球信息科学学报*, 2020, 22(6): 1307–1319.
- CHEN Fangmiao, HUANG Huiping, JIA Kun. Study on the administration and construction of urban agglomeration with spatiotemporal big data: a progress review[J]. *Journal of geo-information science*, 2020, 22(6): 1307–1319.
- [14] YOO M, KOO H. Exploring the impact of spatial autocorrelation on optimistic bias in cross-validation and assessing the effectiveness of spatial cross-validation[J]. *Cartography and geographic information science*, 2025, 52(5): 596–609.
- [15] HUANG Ruixi, LONG Yin, ZIA T. A physical-enhanced spatio-temporal graph convolutional network for river flow prediction[J]. *Applied sciences*, 2025, 15(16): 9054.
- [16] TANTOUSH M, SENGUPTA U, HAN Liangxiu. Exploring city dynamics through tweets: a framework for capturing urban activities as complex spatiotemporal patterns[J]. *Cities*, 2025, 162: 105894.
- [17] GAETAN C, GIRARDI P, MUSAU V M. Spatial quantile clustering of climate data[J]. *Advances in data analysis and classification*, 2025, 19(1): 147–175.
- [18] BUCHDADI A D. Temporal crime pattern analysis using seasonal decomposition and k-means clustering[J]. *Journal of cyber law*, 2025, 1(1): 65–87.
- [19] SPERNA WEILAND R, VAN DER WIEL K, SELTEN F, et al. Spatiotemporal clustering of preferred trajectories of the euro-Atlantic summer circulation[J]. *Journal of climate*, 2025, 38(10): 2435–2459.
- [20] ANSARI M Y, AHMAD A, KHAN S S, et al. Spatiotemporal clustering: a review[J]. *Artificial intelligence review*, 2020, 53(4): 2381–2423.
- [21] CHOI C, HONG S Y. MDST-DBSCAN: a density-based clustering method for multidimensional spatiotemporal data[J]. *ISPRS international journal of geo-information*, 2021, 10(6): 391.
- [22] YOU G. Spatiotemporal data-adaptive clustering algorithm: an intelligent computational technique for city big data[J]. *Annals of the American association of geographers*, 2022, 112(2): 602–619.
- [23] BIRANT D, KUT A. ST-DBSCAN: an algorithm for clustering spatial-temporal data[J]. *Data & knowledge engineering*, 2007, 60(1): 208–221.
- [24] AGRAWAL K P, GARG S, SHARMA S, et al. Development and validation of OPTICS based spatio-temporal clustering technique[J]. *Information sciences*, 2016, 369: 388–401.
- [25] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996: 226–231.
- [26] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. OPTICS: ordering points to identify the clustering structure[J]. *ACM SIGMOD record*, 1999, 28(2): 49–60.
- [27] 吴文浩, 吴升. 多时间尺度密度聚类算法的案事件分析应用[J]. *地球信息科学学报*, 2015, 17(7): 837–845.
- WU Wenhao, WU Sheng. Application of density-based clustering algorithm in crime cases analysis considering multiple time scale[J]. *Journal of geo-information science*, 2015, 17(7): 837–845.
- [28] 王培晓, 张恒才, 王海波, 等. ST-CFSFDP: 快速搜索密度峰值的时空聚类算法[J]. *测绘学报*, 2019, 48(11): 1380–1390.

- WANG Peixiao, ZHANG Hengcai, WANG Haibo, et al. Spatial-temporal clustering by fast search and find of density peaks[J]. *Acta geodaetica et cartographica sinica*, 2019, 48(11): 1380–1390.
- [29] 赵嘉, 姚占峰, 吕莉, 等. 基于相互邻近度的密度峰值聚类算法[J]. *控制与决策*, 2021, 36(3): 543–552.
- ZHAO Jia, YAO Zhanfeng, LÜ Li, et al. Density peaks clustering based on mutual neighbor degree[J]. *Control and decision*, 2021, 36(3): 543–552.
- [30] WANG Min, WANG Aiping, LI Anbo. Mining spatial-temporal clusters from geo-databases[C]//Advanced Data Mining and Applications. Berlin: Springer, 2006: 263–270.
- [31] 沈江飞, 李怀洲, 黄立军, 等. 基于局部离群因子和神经网络模型的设备状态在线监测方法研究[J]. *核动力工程*, 2021, 42(3): 160–165.
- SHEN Jiangfei, LI Huaizhou, HUANG Lijun, et al. Study on online monitoring of equipment condition based on local outlier factor and artificial neural networks model [J]. *Nuclear power engineering*, 2021, 42(3): 160–165.
- [32] 王培晓, 王海波, 傅梦颖, 等. 室内用户语义位置预测研究[J]. *地球信息科学学报*, 2018, 20(12): 1689–1698.
- WANG Peixiao, WANG Haibo, FU Mengying, et al. Research on semantic location prediction of indoor users[J]. *Journal of geo-information science*, 2018, 20(12): 1689–1698.
- [33] ZHANG Fengling, HUANG Shengqiang, ZHANG Haiyan, et al. STSNN-DPC: spatio-temporal shared nearest neighbors and density peaks based clustering method[EB/OL]. (2024–11–25)[2026–01–07]. <https://doi.org/10.21203/rs.3.rs-5381784/v1>.
- [34] NGUYEN X V, EPPS J, BAILEY J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance[J]. *J Mach learn res*, 2010, 11: 2837–2854.
- [35] FOWLKES E B, MALLOWS C L. A method for comparing two hierarchical clusterings[J]. *Journal of the American statistical association*, 1983, 78(383): 553–569.
- [36] 赵嘉, 王刚, 吕莉, 等. 面向流形数据的测地距离与余弦互邻近邻密度峰值聚类算法[J]. *电子学报*, 2022, 50(11): 2730–2737.
- ZHAO Jia, WANG Gang, LÜ Li, et al. Density peaks clustering algorithm based on geodesic distance and cosine mutual reverse nearest neighbors for manifold datasets[J]. *Acta electronica sinica*, 2022, 50(11): 2730–2737.
- [37] 张建勇, 王新, 陈凌, 等. 2022—2023 年四川泸定 MS6.8、MS5.0 和 MS5.6 地震序列的发震构造及成因[J]. *地球物理学报*, 2024, 67(4): 1471–1486.
- ZHANG Jianyong, WANG Xin, CHEN Ling, et al. Seismotectonics and seismogenesis of the 2022–2023 Sichuan Luding MS 6.8, MS5.0 and MS5.6 earthquake sequences [J]. *Chinese journal of geophysics*, 2024, 67(4): 1471–1486.
- [38] ZHANG Zhe, FANG Lihua, XU Lisheng. Primary source characteristics of the 2022 Sichuan Luding MS 6.8 Earthquake[J]. *Chinese journal of geophysics*, 2023, 66(4): 1397–1408.

作者简介:



赵嘉, 教授, 博士生导师, 博士, 主要研究方向为智能计算与计算智能、模式识别与大数据挖掘。主持国家自然科学基金项目 3 项, 发表学术论文 100 余篇, 出版专著 1 部。E-mail: zhaojia925@163.com。



李辉, 硕士研究生, 主要研究方向为数据挖掘。E-mail: lihui66688@126.com。



肖人彬, 博士生导师, 主要研究方向为群体智能、大规模个性化定制、复杂系统与复杂性科学。主持并承担国家自然科学基金项目 11 项, 作为第一完成人获得教育部自然科学奖 1 项和湖北省自然科学奖及科技进步奖 4 项。发表学术论文 300 余篇, 出版学术专著和教材 10 余部。E-mail: rbxiao@hust.edu.cn。