

控制与决策

Control and Decision

共享近邻加权和隶属点分配的时空密度峰值聚类算法

赵嘉, 朱伟涛, 肖人彬, 李思维, 白晖峰, 谢海华, 夏志红

引用本文:

赵嘉, 朱伟涛, 肖人彬, 等. 共享近邻加权和隶属点分配的时空密度峰值聚类算法[J]. *控制与决策*, 2026, 41(5): 1415-1426.

在线阅读 View online: <https://doi.org/10.13195/j.kzyjc.2025.0656>

您可能感兴趣的其他文章

Articles you may be interested in

基于混合邻域约束项的改进FCM算法

Mixed neighborhood constraints based fuzzy C-means algorithm

控制与决策. 2021, 36(6): 1457-1464 <https://doi.org/10.13195/j.kzyjc.2019.1321>

区分交通流模式的混合服务路口信号控制策略

Signal control strategies of mixed service intersections to discriminate traffic flow patterns

控制与决策. 2021, 36(6): 1509-1515 <https://doi.org/10.13195/j.kzyjc.2019.1520>

基于相互邻近度的密度峰值聚类算法

Density peaks clustering based on mutual neighbor degree

控制与决策. 2021, 36(3): 543-552 <https://doi.org/10.13195/j.kzyjc.2019.0795>

基于双权重多邻域保持嵌入的间歇过程故障检测

Fault detection of batch process based on double weight and multiple neighborhoods preserving embedding

控制与决策. 2021, 36(12): 3023-3030 <https://doi.org/10.13195/j.kzyjc.2020.0659>

基于协同聚类和权重注意力稀疏自编码网络的变化检测方法

Change detection approach based on cooperative clustering and weighted attention sparse autoencoder

控制与决策. 2021, 36(10): 2442-2450 <https://doi.org/10.13195/j.kzyjc.2019.1633>

共享近邻加权和隶属点分配的时空密度峰值聚类算法

赵嘉^{1,2†}, 朱伟涛^{1,2}, 肖人彬³, 李思维⁴, 白晖峰⁵, 谢海华^{1,2}, 夏志红¹

1. 江西水利电力大学 信息工程学院, 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, 南昌 330099;
3. 华中科技大学 人工智能与自动化学院, 武汉 430074;
4. 北京中电飞华通信有限公司, 北京 100016; 5. 北京智芯微电子科技有限公司, 北京 102200)

摘要: 快速搜索密度峰值的时空聚类算法计算局部密度时, 难以区分所在区域的密度差异, 易引发类簇中心的选择错误; 分配策略缺乏足够的时空约束, 易将时间特征差异明显但空间位置相近的非密度峰值错误分配; 缺乏独立的噪声识别机制, 其检测效能完全依赖样本分配的准确性, 样本分配偏差致噪声识别精度显著降低. 针对这些挑战, 提出一种共享近邻加权和隶属点分配的时空密度峰值聚类 (SNMP-STDPC) 算法. 引入共享近邻加权策略, 构建时空距离相似度矩阵, 精确反映样本间的密度差异, 有效提升密度峰值选择的可靠性; 结合共享近邻增强时空约束, 将非密度峰值分为必然隶属点和可能隶属点, 确保样本分配的准确性; 提出一种新的噪声识别机制, 计算样本的异常分数并使用动态阈值检测噪声, 提高噪声识别的有效性. 将 SNMP-STDPC 算法与当前主流时空聚类方法在模拟数据集和实际地震观测数据上进行比较, 实验结果表明, SNMP-STDPC 算法显著提升了模拟数据集的聚类效果, 并在真实数据集上表现良好.

关键词: 时空聚类; 密度峰值聚类; 共享近邻; 隶属点分配; 噪声识别; 异常分数

中图分类号: TP181 文献标志码: A

DOI: 10.13195/j.kzyjc.2025.0656

引用格式: 赵嘉, 朱伟涛, 肖人彬, 等. 共享近邻加权和隶属点分配的时空密度峰值聚类算法 [J]. 控制与决策, 2026, 41(5): 1415-1426.

Spatial-temporal density peaks clustering with shared neighbors weighting and membership point assignment

ZHAO Jia^{1,2†}, ZHU Wei-tao^{1,2}, XIAO Ren-bin³, LI Si-wei⁴, BAI Hui-feng⁵, XIE Hai-hua^{1,2}, XIA Zhi-hong¹

(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. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China; 4. Beijing Zhongdian Feihua Communication Company Limited, Beijing 100016, China; 5. Beijing Smart Chip Microelectronics Technology Company Limited, Beijing 102200, China)

Abstract: The spatial-temporal clustering by fast search and find of density peaks (ST-CFSFDP) algorithm faces several critical limitations. Its local density calculation struggles to distinguish density variations within regions, leading to erroneous cluster centers selection. The assignment strategy lacks sufficient spatial-temporal constraints, causing misassignment of non-peak points that exhibit significant temporal differences despite spatial proximity. Furthermore, the absence of an independent noise identification mechanism renders noise detection entirely dependent on sample assignment accuracy, significantly degrading precision when assignment errors occur. In order to address these challenges, this paper proposes a spatial-temporal density peak clustering with shared neighbors weighting and membership point allocation (SNMP-STDPC) algorithm. A shared nearest neighbors weighting strategy is introduced to construct a spatial-temporal distance similarity matrix, which precisely reflects density differences among samples

收稿日期: 2025-06-21; 录用日期: 2025-10-20.

基金项目: 国家自然科学基金项目 (62466037); 南昌市重大科技攻关项目 (2024zdxm010, 2024zdxm002); 江西省教育厅科技项目 (GJJ2401408); 江西省职业早期青年科技人才培养专项项目 (20244BCE52231).

责任编辑: 刘宝碁.

†通信作者. E-mail: zhaojia@nit.edu.cn.

and enhances the reliability of density peak identification. The algorithm incorporates enhanced spatial-temporal constraints leveraging shared nearest neighbors, categorizing non-peak points into must-link points and may-link points to ensure accurate sample assignment. Additionally, a novel noise identification mechanism is proposed, calculating sample anomaly scores and employing a dynamic threshold to significantly improve noise identification effectiveness. The SNMP-STDPC algorithm is compared with current mainstream spatial-temporal clustering methods on simulated datasets, and it is further applied to identify foreshock-aftershock sequences in real-world seismic observation data. Experimental results demonstrate that the SNMP-STDPC algorithm significantly enhances clustering performance on synthetic datasets and delivers robust results on real-world data.

Keywords: spatial-temporal clustering; density peaks clustering; shared neighbors; membership point assignment; noise identification; anomaly scores

0 引言

随着信息技术的迅猛发展, 社会各类系统产生的数据持续激增, 传统的结构化数据正逐步迈入多源异构, 社会已进入快速增长的大数据时代^[1]. 在诸如地震监测^[2]、公共卫生^[3]、智慧交通^[4]、环境监测^[5]等领域, 数据深度融合空间位置与时间演化信息, 构成了典型的时空数据^[6]. 如何从时空数据中提取潜在模式^[7], 识别异常事件^[8], 预测未来趋势^[9], 已成为支撑公共安全与资源优化配置等重大决策的核心命题.

在此背景下, 聚类分析方法凭借其强大的模式发现能力, 成为支撑相关领域科学决策的关键技术. 然而, 时空数据固有的空间依赖性、时间动态性及异质性, 使得其聚类分析相较于传统静态数据面临更为严峻的挑战^[10]. 核心在于算法需同步刻画数据的空间分布格局及其时间演化规律, 这对模型的灵活性和适应性提出了更高要求^[11].

学术界提出了多种面向时空数据的聚类算法, 主要分为基于时空扫描、基于时空距离和基于时空密度 3 类方法. 基于时空扫描的方法通过移动一个预设的、通常为空间圆形区域和时间区间构成的圆柱体的时空扫描窗口遍历研究区域, 识别具有最高统计显著性的潜在时空聚集簇^[12], 该方法在公共卫生领域的传染病时空传播建模^[13]及犯罪地理学中的热点区域画像^[14]等实证研究中展现出显著优势; 基于时空距离的方法通过构建混合时空距离度量^[15], 综合空间与时间维度的相似性以度量样本关联性, 其核心是利用距离函数量化时空数据的复杂特征^[16], 该方法主要聚焦地质灾害时空演化规律研究^[17], 在其他领域的适用性较为有限; 基于时空密度的方法通过度量样本的局部密度划分类簇, 不仅能够识别形状不规则的类簇, 而且无需预先指定类簇数量. Birant 等^[18]提出的 ST-DBSCAN 和 Agrawal 等^[19]提出的 ST-OPTICS 算法分别在 DBSCAN 和 OPTICS 算法的基础上引入时间维度约束条件, 成功扩展了传统的密度聚类方法, 使其能够处理具有时空特征

的数据集. ST-DBSCAN 算法在密度计算时, 低密度类簇的核心对象因所处环境稀疏, 难以满足最小近邻数要求, 易被误判为非核心样本; 样本分配时, 边界样本与已有类簇的密度关联性不强, 导致匹配不足; 噪声判定时, 无法动态调整标准, 低密度类簇样本容易误判为噪声. ST-OPTICS 算法在密度计算与噪声判定上与 ST-DBSCAN 相似, 其在样本分配时, 密度突变点的识别易受数据分布影响, 导致微簇划分偏差. 此外, Liu 等^[20]引入窗口距离和共享近邻思想定义时空密度, 提出了时空共享近邻 (spatial-temporal shared neighbor, STSNN) 算法^[21]. STSNN 算法未区分稀疏与密集区域的密度贡献, 导致稀疏区域样本密度被低估; 样本分配仅依赖时空可达性, 忽视非空间属性; 噪声识别依赖经验参数, 无法适应不同密度类簇判别需求.

密度峰值聚类 (density peaks clustering, DPC) 算法由 Rodrigues 等^[22]提出并在 Science 发表, 其结合局部密度峰值识别与相对距离度量策略, 计算样本的局部密度和相对距离以自动确定类簇中心, 依据更高密度的最近邻所属类簇对剩余样本进行分配. DPC 算法对空间类簇展现出优良性能^[23-25], 但在时空数据挖掘领域的应用潜力尚未充分挖掘.

为使 DPC 算法适配时空聚类任务, 王培晓等^[26]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. ST-CFSFDP 算法利用时间邻域信息与空间距离阈值衡量样本的局部密度, 确定类簇中心后, 依据与邻近类簇中心的空间距离实施剩余样本的定向分配. 该算法弥补了 DPC 算法时空属性处理上的不足, 使其具备复杂时空关联模式的捕获能力. 然而, 其仍然存在以下局限: 1) 局部密度计算难以辨别不同区域的密度差异, 容易导致类簇中心选择失误; 2) 分配策略因时空约束不足, 易将时间特征差异显著且空间位置邻近的非密度峰值样本错误分配; 3) 噪声识别依靠分配策略, 非密度峰

值分配的不准确会导致噪声的识别不精确, 进而影响聚类结果的准确性。

针对上述问题, 本文提出一种共享近邻加权和隶属点分配的时空密度峰值聚类 (spatial-temporal density peaks clustering with shared neighbors weighting and membership point assignment, SNMP-STDPC) 算法。该算法计算局部密度时, 通过定义共享近邻加权的局部密度提升精度, 更准确地反映样本真实分布, 提高类簇中心选择的可靠性; 在非密度峰值样本分配阶段, 提出基于必然与可能隶属点判定的分配策略, 利用时空共享近邻阈值筛选强关联样本以扩展类簇, 并借助动态密度统计与增量优化机制处理不确定性样本, 从而降低误分配率; 设计异常分数噪声检测策略, 计算样本异常分数并动态调整阈值, 实现更精确的噪声识别, 提升算法准确性。

1 ST-CFSFDP 算法

ST-CFSFDP 算法对 DPC 算法局部密度和相对距离的计算策略进行调整, 从而具备处理时空数据的能力。经典的 DPC 算法凭借其独特的密度峰值识别机制, 能够有效识别数据类簇并检测噪声, 在保证较快收敛速度的同时展现出优异的聚类性能。

DPC 算法的特点是: 1) 类簇中心的局部密度较大, 围绕类簇中心的样本密度较小; 2) 类簇中心相互距离较远, 能避免相互影响。通过计算样本的局部密度及相对距离, 算法能够准确识别类簇中心。DPC 的局部密度定义如下:

$$\rho_i = \sum_{i \neq j} \chi(d_{ij} - d_c), \quad \chi(x) = \begin{cases} 1, & x < 0; \\ 0, & x \geq 0. \end{cases} \quad (1)$$

$$\rho_i = \sum_{i \neq j} \exp \left[- \left(\frac{d_{ij}}{d_c} \right)^2 \right]. \quad (2)$$

其中: d_{ij} 为样本 i 与样本 j 的欧氏距离, d_c 为样本的邻域截断距离。

式 (1) 使用样本邻域内样本个数作为该样本的局部密度, 称为截断核的局部密度, 在规模较大的数据集中效果更佳。式 (2) 使用高斯公式定义样本局部密度, 称为高斯核的局部密度, 在规模较小的数据集中效果更佳。

DPC 算法将非局部密度最高样本的相对距离定义为与密度比其高且距离最近样本的距离, 而局部密度最高样本的相对距离定义为所有样本中相对距离的最大值。相对距离公式如下:

$$\delta_i = \begin{cases} \min_{j: \rho_j > \rho_i} (d_{ij}), & \exists j, \text{ s.t. } \rho_j > \rho_i; \\ \max_j (d_{ij}), & \text{otherwise.} \end{cases} \quad (3)$$

为准确选择类簇中心, DPC 算法通过决策值选择样本, 定义如下:

$$\gamma_i = \rho_i \cdot \delta_i. \quad (4)$$

选取决策值较大的样本为类簇中心, 将剩余样本分配给密度较高且距离最近的样本所属类簇。

为适用时空数据, ST-CFSFDP 对局部密度 ρ_i 和相对距离 δ_i 进行修改, 使其不仅考虑空间距离, 还引入时间约束。含有时间约束的截断核的局部密度定义如下:

$$\rho_i = \sum_{i \neq j} \chi(d_{ij} - d_c, t_{ij} - t_c). \quad (5)$$

其中: t_{ij} 为样本 i 与样本 j 之间的时间距离; t_c 为时间截断距离, 表示布尔函数, 仅当两个输入参数同时小于 0 时输出 1, 否则输出 0, 即时间约束下的局部密度表示同时小于空间邻域和时间邻域的样本个数。仅空间和时间上都满足条件的样本, 才会参与局部密度的计算。

ST-CFSFDP 算法综合空间与时间因素, 对相对距离的定义进行优化, 修改后的公式为

$$\begin{cases} \delta_i = \min_{j: \rho_j > \rho_i} (t_{ij}), & i \geq 2; \\ \delta_i = \max_{j \geq 2} (\delta_j), & i = 1. \end{cases} \quad (6)$$

单个类簇内样本数量较多且分布相对连续的情况下, 该类簇在时间上的区间跨度可能较长。若类簇内样本具有较高的局部密度、较大的时间跨度, 则会导致这些样本的相对距离增大。根据 DPC 算法原理, 上述情况可能被识别为多个独立的密度峰值, 导致单个连续簇在聚类结果中被错误地分割。ST-CFSFDP 算法为此引入类簇中心合并策略, 若时间轴上相邻类簇中心的空间距离小于预设空间截断距离, 则将这两类簇中心合并。

ST-CFSFDP 算法的样本分配遵循时序导向策略, 类簇中心沿时间维度双向遍历未分配样本。具体而言, 每个类簇中心沿时间轴正向和逆向扫描未分配样本, 基于最小距离准则将样本分配至最近的类簇中心。遍历完成后, 未分配样本标识为噪声。

2 SNMP-STDPC 算法

2.1 共享近邻加权的局部密度

ST-CFSFDP 算法采用截断核方式计算样本局部密度。该方法仅依赖时空邻域内的有限样本信息, 且对邻域内样本关系的刻画相对简单, 致使不同样本的局部密度易趋同, 降低了类簇中心判定的准确性。为此, 本节提出通过识别时空样本间共享的最近邻样本, 重新定义其局部密度。该方法在有限时空邻

域内,通过窗口距离判定时空近邻,并据此识别时空共享近邻,融合直接邻近性与共享相似性,以精深刻画局部关系及密度差异.

定义1 (窗口距离 (DW)) 给定两个时空事件 $ST_i(L_i, t_i)$ 、 $ST_j(L_j, t_j)$, $ST_j(L_j, t_j)$ 在以 $ST_i(L_i, t_i)$ 为中心的圆柱体 (时空邻域) 中, 则 $ST_i(L_i, t_i)$ 与 $ST_j(L_j, t_j)$ 之间的窗口距离定义为 L_i 与 L_j 之间的空间距离, 记为 $D_w(ST_i, ST_j)$.

定义2 (时空 K 近邻 (STKNN)) 同时满足如下两式的时空事件:

$$|t_{i+k} - t_i| \leq DT, \quad (7)$$

$$D_w(ST_i, ST_{i+1}) \leq D_w(ST_i, ST_{i+2}) \leq \dots \leq D_w(ST_i, ST_{i+k}) \quad (8)$$

称为 $ST_i(L_i, t_i)$ 的时空 K 近邻 $STKNN(ST_i)$. 其中: DT 为预设时间窗, $ST_i(L_i, t_i)$ 的时空 K 近邻为 $ST_{i+1}(L_{i+1}, t_{i+1})$, $ST_{i+2}(L_{i+2}, t_{i+2})$, \dots , $ST_{i+k}(L_{i+k}, t_{i+k})$. 若存在时空事件在预定义时间窗中的时空近邻数小于给定的 k 值, 则直接判定为噪声.

将时空事件 $ST_i(L_i, t_i)$ 与其第 k 个时空近邻之间的窗口距离定义为 $ST_i(L_i, t_i)$ 的时空邻域半径. 该时空邻域并非固定大小的圆柱体, 而是根据时空事件与其第 k 个近邻之间的窗口距离进行动态调整.

定义3 (互为时空 K 近邻) 给定时空事件 $ST_i(L_i, t_i)$ 、 $ST_j(L_j, t_j)$, 若 $ST_i(L_i, t_i)$ 属于 $ST_j(L_j, t_j)$ 的时空 K 近邻集合 $STKNN(ST_j)$, 且同时 $ST_j(L_j, t_j)$ 属于 $ST_i(L_i, t_i)$ 的时空 K 近邻集合 $STKNN(ST_i)$, 则称二者互为时空 K 近邻.

定义4 (时空共享近邻 (STSN)) 对于互为时空 K 近邻的两个时空事件 $ST_i(L_i, t_i)$ 和 $ST_j(L_j, t_j)$, 其时空共享近邻是指同时属于 $ST_i(L_i, t_i)$ 和 $ST_j(L_j, t_j)$ 时空邻域的其他时空事件. 有

$$\begin{aligned} STSN(ST_i, ST_j) = \\ STKNN(ST_i) \cap STKNN(ST_j). \end{aligned} \quad (9)$$

空间域中, 共享近邻方法已被证实能有效处理复杂数据分布下的聚类问题^[27]. 该方法通过度量两空间样本共享的近邻, 衡量它们之间的相似性. 若两个空间样本距离相近且共享大量近邻, 则这两个实体极有可能属于同一类簇.

本文进一步将共享最近邻方法拓展到时空域. 基于定义1~定义3, 对依据时空共享近邻定义的样本时空密度做进一步细化.

定义5 (权重系数) 为使样本的局部密度定义更准确地反映样本的时空局部分布, 分别定义时间

权重系数 α_i 和空间权重系数 β_i 如下:

$$\alpha_i = 1 - \frac{\sum_{j \in STKNN(i)} t_{ij}}{\sum_{i \neq j} t_{ij}}, \quad (10)$$

$$\beta_i = 1 - \frac{\sum_{j \in STKNN(i)} d_{ij}}{\sum_{i \neq j} d_{ij}}. \quad (11)$$

定义的时间权重系数 α_i 与空间权重系数 β_i , 可反映样本间的时空邻近性; 若某样本的时空近邻域内样本密度越高, 空间距离越紧凑 (即近邻样本数量越多且分布越紧密), 则该样本被识别为类簇中心的概率越大.

定义6 (样本间的密集程度) 有

$$\begin{aligned} A(i, j) = \\ \left(\sum_{v \in [i, STKNN(i)]} (\alpha_j \cdot e^{-t_{ij}} + \beta_j \cdot e^{-d_{ij}}) + \sum_{v \in [j, STKNN(j)]} (\alpha_i \cdot e^{-t_{vi}} + \beta_i \cdot e^{-d_{vi}}) \right) / (2 \cdot (k + 1)). \end{aligned} \quad (12)$$

其中: $A(i, j)$ 反映了样本 i 与样本 j 在整个数据集中的密集程度; 分子左侧为样本 i 及其时空近邻到样本 j 之间的相似性之和; 分子右侧为样本 j 及其时空近邻到样本 i 之间的相似性之和; 分母起到归一化的作用.

定义7 (时空共享近邻的样本相似度) 有

$$\text{Sim}(i, j) = \frac{|STSN(i, j)|}{k} \cdot A(i, j). \quad (13)$$

其中: $\text{Sim}(i, j)$ 为样本 i 与 j 考虑当前所处环境下的时空相似程度, $|STSN(i, j)|$ 为样本 i 和 j 的时空共享近邻个数. 时空共享近邻可更好地反映样本之间最真实的相似情况.

定义8 (共享近邻加权的局部密度) 有

$$\rho_i = \sum_{j \neq i} \text{Sim}(i, j). \quad (14)$$

样本 i 的局部密度为与其时空近邻样本间的相似度之和, 该局部密度定义综合考虑了样本间距离与局部信息. 当样本间距离相等且稠密程度相似时, 时空共享近邻数成为判别样本相似性的关键指标, 时空共享近邻越多, 样本相似性越高. 若样本间距离相同但区域密度差异显著, 则稀疏区域样本更易符合时空共享近邻条件, 这符合空间分布规律. 算法通过增强样本的密度贡献, 实现对不同分布特征的差异化评估, 能更准确地识别密度峰值.

为验证共享近邻加权的局部密度对类簇中心识别的有效性,通过实验对包含4个类簇的时空数据集STD3进行分析.图1对比了截断核与共享近邻加权两种局部密度定义方式对类簇中心选择的影响.在数据集上对参数调优后计算决策值 γ ,数据集包含4个类簇.如图1所示,在对比分析不同局部密度决策图时发现,共享近邻加权的局部密度决策图呈现出4个五角星形点彼此邻近且与其余数据点间距较大的特征;相较之下,截断核函数所生成的局部密度决策图中,顶部的3个五角星形点出现重合,底部的五角星形点则偏离上方点群并朝下部点群偏移.由此可知,相较于截断核函数,共享近邻加权局部密度在类簇中心的判定过程中展现出更显著的优势.

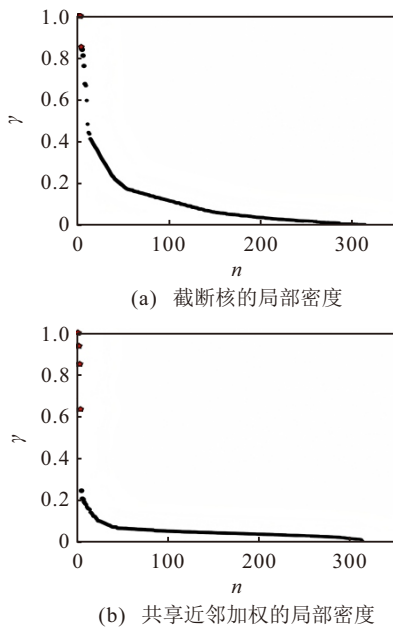


图1 不同局部密度度量下数据集STD3的决策图

2.2 隶属点分配策略

ST-CFSFDP 算法的样本分配策略存在时空约束不足的问题,主要表现为过度依赖样本与类簇中心的空间距离,未能有效融合时间属性以构建准确的时空关联.为此,本节设计了隶属点分配策略,通过将非密度峰值样本划分为必然隶属点与可能隶属点,并利用时空共享近邻数筛选强关联样本以扩展类簇,同时采用动态密度统计与增量优化机制处理不确定性样本.

定义 9 (必然隶属点) 假设样本*i*已分配至某类簇,样本*j*尚未分配,则样本*j*是样本*i*所属类簇的必然隶属点,当且仅当样本*j*满足

$$|\text{STSN}(i, j)| \geq k/2. \quad (15)$$

当时空共享近邻数足够时,即视为足够接近并归为同一类簇,未满足必然隶属条件的未分配样本

则定义为可能隶属点.

当未分配样本与已分配样本的时空共享近邻数 $\geq k/2$ 时,二者时空近邻集合半数以上重合,既体现了空间维度局部结构一致,又能反映时间维度演化特征相近.同时,该阈值能平衡类簇扩展的包容性与准确性,避免阈值过高导致类簇收缩漏纳同类样本、过低导致类簇扩张误纳异类样本.

定义 10 (可能隶属点) 若样本*i*已分配至某类簇但样本*j*未分配,则样本*j*为该类簇的可能隶属点,当且仅当满足

$$0 < |\text{STSN}(i, j)| < k/2. \quad (16)$$

依据式(4)选取决策值较大的样本作为类簇中心后进入样本分配阶段,该过程分为两步:首先,以类簇中心为起点,通过广度优先搜索遍历其时空邻域,若邻域内样本的时空共享近邻数达到预设阈值,则将其标记为必然隶属点,并加入队列以扩展搜索范围,确保每个样本仅被处理一次;随后,处理剩余未分配样本,动态统计各样本邻域内已分配类簇的密度分布,构建分配矩阵并利用最大堆优先队列快速定位当前密度最高的类簇-样本对,每次分配后仅更新受影响邻域的计数,通过堆结构的增量调整持续定位下一个最优分配目标,直至所有样本完成分配.

2.3 异常分数噪声检测策略

由于 ST-CFSFDP 算法缺乏独立噪声检测策略,本节提出异常分数噪声检测策略.该策略融合样本的局部密度与相对距离计算其异常分数,并利用类簇中心异常分数的中位数量化整体异常程度,进而基于异常分数的四分位距动态调整检测动态阈值,将异常分数超过动态阈值的样本标记为噪声.

定义 11 (异常分数 (NAS)) 异常分数计算为

$$\text{NAS}_i = \frac{\delta_i}{\text{median}(\delta_{\text{cluster}}) + \epsilon} \cdot \frac{\text{median}(\rho_{\text{cluster}})}{\rho_i + \epsilon}. \quad (17)$$

其中: NAS_i 为样本*i*的异常分数; $\text{median}(\delta_{\text{cluster}})$ 为所有簇心局部密度的中位数; $\text{median}(\rho_{\text{cluster}})$ 为所有类簇中心相对距离的中位数; ϵ 为一个小常数,用于避免除以零的情况.异常分数由局部密度和相对距离共同决定,用以量化样本的异常程度.

通过动态阈值确定哪些样本被视为噪声,动态阈值的计算基于异常分数的四分位距 (IQR),即

$$\text{NAS}_{\text{threshold}} = \text{median}(\text{NAS}) + 1.5 \cdot \text{IQR}(\text{NAS}). \quad (18)$$

其中: $\text{median}(\text{NAS})$ 为总样本异常分数的中位数; IQR为上四分位数 (Q_3 , 即数据经升序排序后位于前 75% 位置的数值) 与下四分位数 (Q_1 , 即数据经

升序排序后位于前 25% 位置的数值) 之间的差值, 即 $IQR = Q3 - Q1$, 该范围包含了数据中间 50% 的部分, 可以较好地反映数据的集中趋势. 选择 1.5 倍 IQR 作为噪声检测系数, 源于统计学理论将其作为异常值识别常用标准, 该系数基于正态分布假设可涵盖 99.3% 的正常数据, 能有效平衡正常样本保留与噪声检测.

通过与动态阈值进行比较, 将异常分数超过动态阈值的样本标记为噪声, 这种检测策略能够自适应数据分布特征, 显著提升噪声识别的精确度.

为验证 SNMP-STDPC 算法样本分配与噪声检测策略的有效性, 开展相应实验. ST-CFSFDP 算法中样本分配与噪声检测功能未拆分, 样本分配过程同步完成噪声判断, 无法单独评估各策略效果; SNMP-STDPC 算法将样本分配与噪声检测分为相互独立的模块, 可分别优化与验证. 实验验证基于 STD3 数据集, 且确保各实验所用类簇中心一致, 同步检验两种算法在样本分配与噪声识别方面的有效性, 直观展现两种算法策略设计上的差异. 图 2(a) 为使用 ST-CFSFDP 算法的样本分配和噪声检测策略得到的聚类结果. 紫色类簇下半部分与黄色类簇上半部分之间的样本按时间先后顺序进行分配, 并依据与两类簇中心的空间距离决定归属; 黄色类簇上方存在一样本, 其与紫色类簇的时间关系更贴近而

在空间上更靠近黄色类簇, 算法在此处停止搜索; 紫色类簇下方亦存在一样本, 其在时间上更接近黄色类簇而空间上更靠近紫色类簇, 算法同样在此处停止搜索. ST-CFSFDP 算法将时间轴上位于二者之间的所有样本识别为噪声, 紫色类簇下缘与黄色类簇上缘的样本划分错误. 该算法缺乏独立的噪声识别机制, 因此黄色与橙色类簇、橙色与蓝色类簇连接处的噪声未被剔除. 此外, 算法将时间轴上第 1 个类簇中心左侧的样本直接归入该簇中心所属类簇, 导致蓝色类簇左下方的噪声被并入蓝色类簇. 图 2(b) 给出了 SNMP-STDPC 算法隶属点分配策略的聚类结果. 该策略将非密度峰值样本划分为必然隶属点与可能隶属点, 利用时空共享近邻数筛选强关联样本以扩展类簇, 同时引入动态密度统计与增量优化处置不确定样本, 从而在暂不考虑噪声的前提下实现了样本的有效分配. 在图 2(b) 的基础上, 图 2(c) 为进一步使用异常分数噪声检测策略得到的聚类结果. 该方法综合样本的局部密度与相对距离计算异常分数, 以类簇中心异常分数的中位数量化全局异常程度, 进而依据四分位距自适应调整检测阈值, 将超过阈值的样本标注为噪声. 该过程准确识别了各类簇之间及左下方区域的噪声, 改善了边界与过渡区域样本的纯净度.

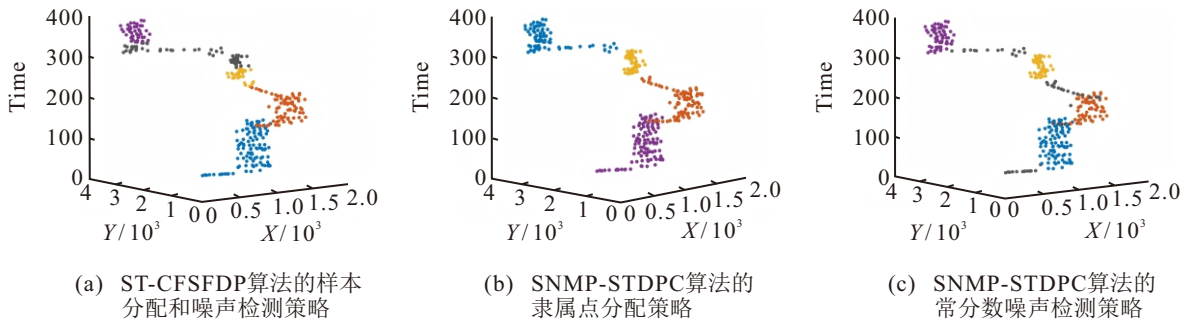


图2 不同分配策略与噪声检测策略下数据集STD3的聚类结果

2.4 算法步骤

输入: 时空数据序列 $seq = \{t_i, x_i, y_i\}$, 时间阈值 t_c , 最近邻个数 k ;

输出: 样本的聚类结果 $labels = \{c_i\}$.

step 1: 归一化数据, 构建样本间空间欧氏距离矩阵和时间距离矩阵.

step 2: 使用式 (7) ~ (14) 计算共享近邻加权的局部密度 ρ_i .

step 3: 使用式 (6) 计算样本的相对距离 δ_i .

step 4: 使用式 (4) 计算决策值 γ_i , 并绘制决策图, 选取类簇中心集合.

step 5: 使用式 (15) 和 (16) 计算样本隶属关系, 首先将满足必然隶属关系的样本进行分配, 然后遍历对所有未分配的可能隶属点进行分配, 直至所有样本分配完成.

step 6: 使用式 (17) 和 (18) 计算样本的异常分数, 与动态阈值进行比较, 检测类簇中的噪声.

step 7: 返回最终结果.

2.5 算法时间复杂度分析

设数据集的样本规模为 n , ST-CFSFDP 算法的时间复杂度为 $O(n^2)$. SNMP-STDPC 算法的时间复杂度主要由 5 个部分组成: 1) 构建样本时间距离矩

阵和空间距离矩阵的时间复杂度为 $O(n^2)$; 2) 计算样本间相似性的时间复杂度为 $O(n^2)$; 3) 计算样本时空近邻矩阵的时间复杂度为 $O(n \lg n)$; 4) 计算样本共享近邻矩阵的时间复杂度为 $O(n^2)$; 5) 计算样本分配矩阵的时间复杂度为 $O(n \lg n)$; 6) 噪声的检测, 计算每个样本的异常分数和计算动态阈值, 时间复杂度为 $O(n \lg n)$. 综上, 本文算法时间复杂度为 $O(n^2)$, 与 ST-CFSFDP 算法量级相同.

3 实验结果与分析

3.1 实验设置

为验证 SNMP-STDPC 算法时空数据集上的聚类表现, 使用模拟数据集和真实数据集进行实验, 并与 ST-CFSFDP^[26]、ST-DBSCAN^[18]、ST-OPTIC^[19] 和 ST-AGNES^[28] 聚类算法进行比较.

本节模拟了 4 组数据量和类簇个数不同的数据集 STD1 ~ STD4. 各数据集的构建流程包括 3 个步骤: 1) 确定核心类簇的数量及其几何特征; 2) 设定这些类簇在空间中的具体分布位置; 3) 按照实验要求添加适当比例的噪声, 以增强数据集的真实性和复杂性. 表 1 为数据集的基本信息, 在数据集 STD1 和 STD3 中, 各类簇于时间维度上呈现出显著差异, 借此对 SNMP-STDPC 算法在时间属性突出的数据集上的表现予以检验. 而数据集 STD2 和 STD4 具备的特性为在特定时段内涵盖多个处于不同空间位置的类簇, 特别是数据集 STD2, 其内类簇的地理边界相对模糊, 可用于评估 SNMP-STDPC 算法密集数据集上的聚类性能.

表1 模拟数据集

数据集	实例数	维度	类簇个数	数据形式
STD1	350	3	2	$\langle \text{time}, x, y, \text{label} \rangle$
STD2	1153	3	3	$\langle \text{time}, x, y, \text{label} \rangle$
STD3	314	3	4	$\langle \text{time}, x, y, \text{label} \rangle$
STD4	1550	3	5	$\langle \text{time}, x, y, \text{label} \rangle$

采用中国地震台网记录的实测地震数据 (73°E ~ 125°E, 18°N ~ 48°N) 作为真实地震数据集, 该地震数据集涵盖了时间、经度、纬度以及强度 4 个关键维度的信息. 时间维度能够展现地震活动不同历史阶段的演变规律; 经度与纬度信息则精准定位地震的发生位置, 勾勒出地震地理空间上的分布态势; 强度数据直观反映了地震的破坏力大小. 这 4 个维度的有机结合, 为深入剖析地震特征, 验证本文所提出算法在地震数据分析中的有效性提供了精准且全面的数据支持, 地震数据信息见表 2.

表2 真实数据集

发震日期	经度	纬度	震级类型	震级	震中位置
2023-12-31 22:27:37	105.02	36.74	Ms	4.9	甘肃白银
2023-12-24 21:22:11	122.65	24.94	Ms	4.7	台湾东部海域
⋮	⋮	⋮	⋮	⋮	⋮
2022-1-1 15:27:57	123.37	24.91	Ms	4.5	台湾宜兰海域
2022-1-1 10:18:01	74.12	39.03	mB	4.9	新疆阿克陶

所有实验环境均为 Win11 64bit 操作系统, Matlab R2023a 软件, 16 GB 内存, Intel(R) Core(TM) i5-12500H 2.50 GHz.

为评估聚类结果的有效性, 选用调整互信息 (AMI)^[29]、调整兰德系数 (ARI)^[30] 以及 FM 指数 (FMI)^[31]. 这些指标的取值范围均为 0 ~ 1, AMI 指标用于衡量聚类结果与真实标签之间的相似度; ARI 指标用于评估聚类结果与真实标签的一致性; FMI 指标综合了精确度和召回率, 以评估聚类结果的准确性. 各聚类评估指标与算法性能呈显著正相关, 指标数值增加直接体现聚类准确性的提升. 评估过程中, 噪声作为独立类别参与最终的指标计算. 采用参数搜索策略, 在预设参数范围内以固定步长进行调参, 通过多次实验验证确定最优参数组合, 以确保各算法在给定数据集上达到最佳的聚类效果. 对数据集类簇结构进行几何特征分析, 计算各类簇的长宽尺寸并采用半径 r 的内切圆完全覆盖类簇内样本, 同时设定最大覆盖半径 s_{\max} 并记录各类簇最大垂直高度 t_{\max} . ST-DBSCAN 与 ST-OPTICS 算法的参数相同且设置范围一致, 空间邻域 $\text{esp1} \in [0.1, s_{\max}]$, 时间邻域 $\text{esp2} \in [0.1, t_{\max}/2]$, 搜索增量为 0.1, 最小邻域数量 min Pts 由数据集规模 N 的对数得到. ST-CFSFDP 及 SNMP-STDPC 算法中, 时间截距 t_c 的取值区间设定为 $[1, t_{\max}/2]$, 搜索增量为 0.1. 另外, SNMP-STDPC 算法中时空 K 近邻参数 k 的取值区间为总样本数 N 的 [1%, 2%], 以步长为 1 进行迭代搜索, 找到最高聚类精度下的近邻参数 k ; ST-CFSFDP 在 $d_c \in [0.1, s_{\max}]$ 范围内设置空间截断距离. ST-AGNES 算法 $\text{dis}_{\text{thred}} \in [1, N/2]$, 以 1 为固定增量进行参数搜索.

3.2 模拟数据集的实验结果分析

表 3 展示了 5 种算法在模拟数据集上的聚类性能对比结果, 其中加粗字体显示的数据表示在处理相同数据集时, 算法展现的最优聚类效果. 分析表 3 可知, SNMP-STDPC 算法在 4 个数据集上的聚类效果最佳. ST-OPTICS 与 ST-DBSCAN 算法在不同数据集上表现次优, 其中 ST-OPTICS 算法在数据集

STD1、STD3和STD4上表现次优, ST-DBSCAN算法则在数据集STD2上表现次优. 相对而言, ST-CFSFDP和ST-AGNES两种算法在5个测试集中均表现较差.

为更直观地展示各算法在4个模拟数据集的真实聚类结果, 绘制了数据集的聚类结果对比图, 如图3~图6所示. 聚类结果的可视化展示中, 采用不同颜色区分不同类簇, 其中黑色标记对应噪声.

数据集STD1包含2个密度不同的时空类簇. 如图3所示, SNMP-STDPC、ST-DBSCAN和ST-OPTICS算法成功识别出预定义的两个类簇. SNMP-STDPC算法凭借其有效的噪声检测策略, 实现了最佳表现. 相比之下, ST-CFSFDP和ST-AGNES算法因缺少独立的噪声识别机制, 将噪声识别为正常类簇样本, 导致聚类结果不理想.

图4(a)展示了数据集STD2的真实散点图, 该数

据集包含3个大小和密度各异且彼此相邻的类簇. ST-CFSFDP算法在确定聚类中心及合并操作时存在机制性缺陷, 导致3个本应水平连续的类簇被错误分割为2个垂直分布的独立类簇; 同时, 由于样本分配策略存在局限性, 中间区域的样本被误标记为噪声. 与ST-CFSFDP算法类似, ST-AGNES算法在处理时空数据时也存在固有局限, 致使整体性能指标偏低. ST-OPTICS和ST-DBSCAN算法因左侧两个类簇过于紧密, 未能完全正确分类边界样本, 在3种评价指标上均略低于SNMP-STDPC算法.

数据集STD3包含4个规模不等的类簇及相互连接的线型噪声, 各类簇在时间属性上差异显著. 如图5所示, 5种算法均能识别出各类簇的分布边界. SNMP-STDPC和ST-CFSFDP算法准确识别出正确的类簇中心, 但ST-CFSFDP算法仅识别出紫色与黄色类簇交界区域的部分噪声, 并将部分黄色类簇样

表3 5种算法在4个模拟数据集上的聚类结果

聚类算法	STD1			STD2		
	AMI	ARI	FMI	AMI	ARI	FMI
SNMP-STDPC	0.7870	0.8426	0.9158	0.9825	0.9928	0.9961
ST-CFSFDP	0.4729	0.5548	0.7414	0.1783	0.1446	0.4914
ST-DBSCAN	0.6493	0.7229	0.8353	0.9813	0.9920	0.9945
ST-OPTICS	0.7399	0.8200	0.8952	0.9689	0.9875	0.9913
聚类算法	STD3			STD4		
	AMI	ARI	FMI	AMI	ARI	FMI
ST-AGNES	0.3118	0.3742	0.5938	0.1490	0.1496	0.4228
SNMP-STDPC	0.9199	0.9337	0.9512	0.8795	0.9045	0.9213
ST-CFSFDP	0.6806	0.6686	0.7440	0.2103	0.0630	0.4388
ST-DBSCAN	0.8940	0.8836	0.9082	0.8098	0.7906	0.8274
ST-OPTICS	0.9105	0.9269	0.9424	0.8777	0.9035	0.9205
ST-AGNES	0.7283	0.7645	0.8249	0.1579	0.0996	0.3206

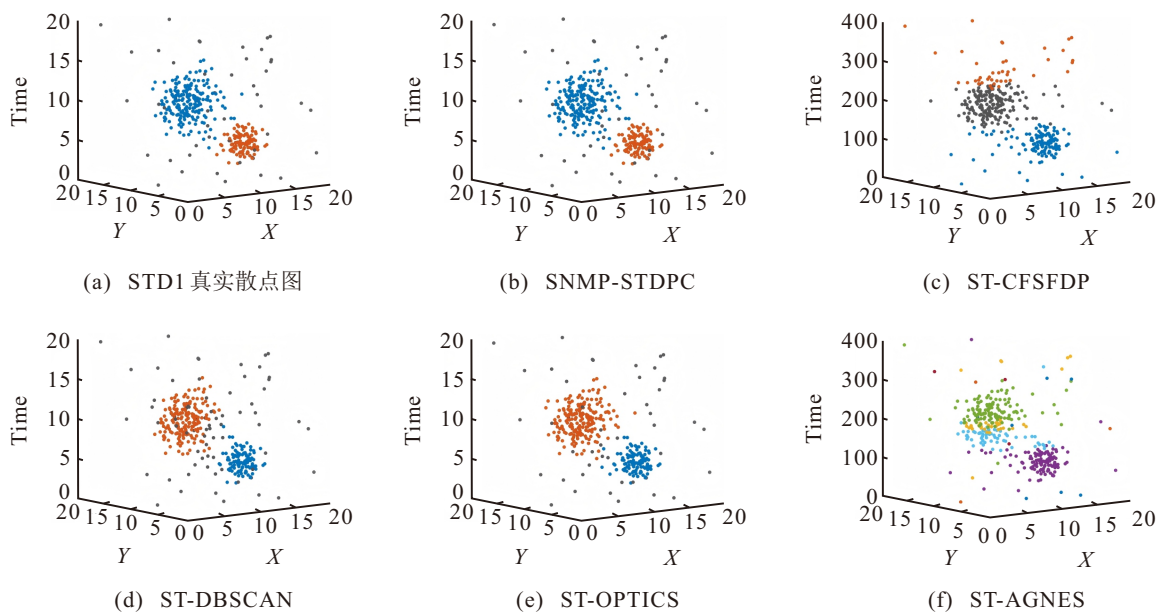


图3 5种算法在数据集STD1的时空聚类结果

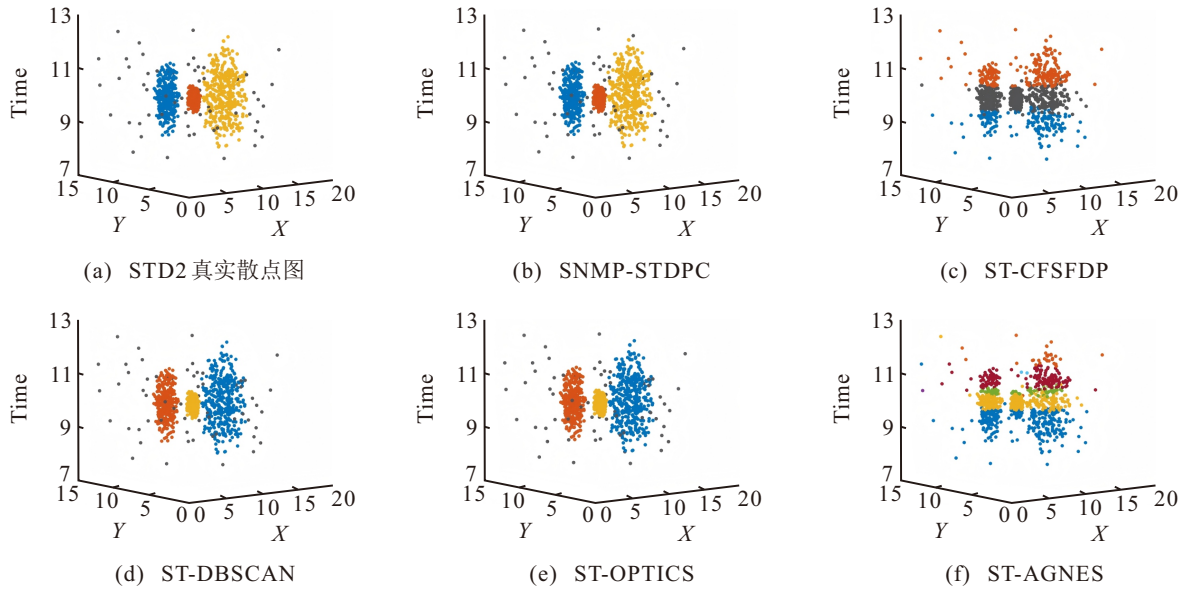


图4 5种算法在数据集STD2的时空聚类结果

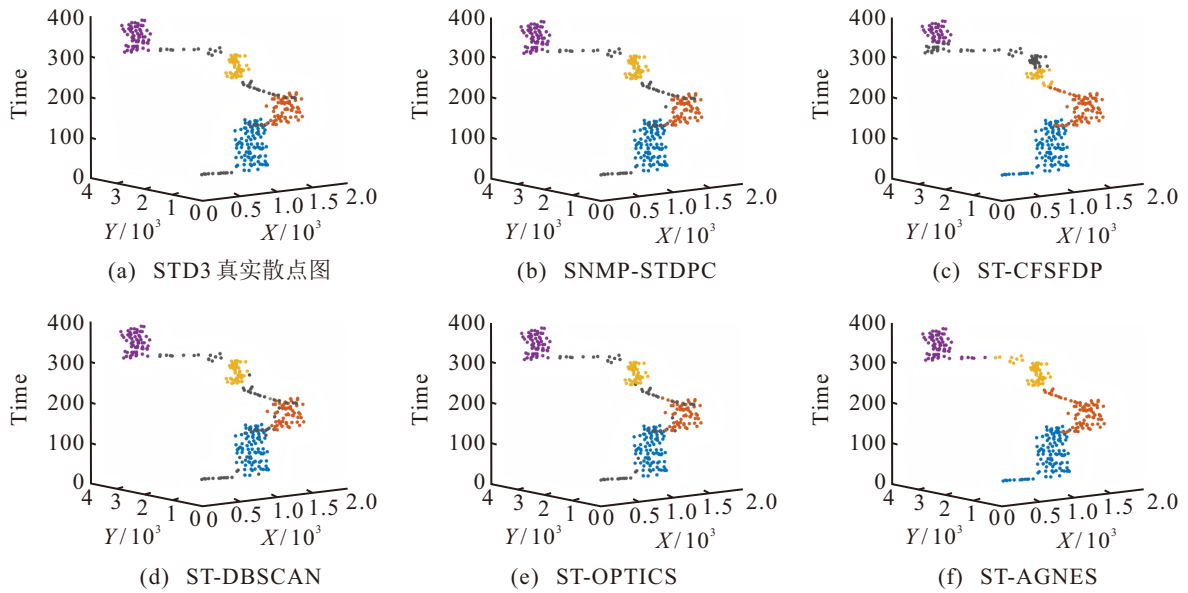


图5 5种算法在数据集STD3的时空聚类结果

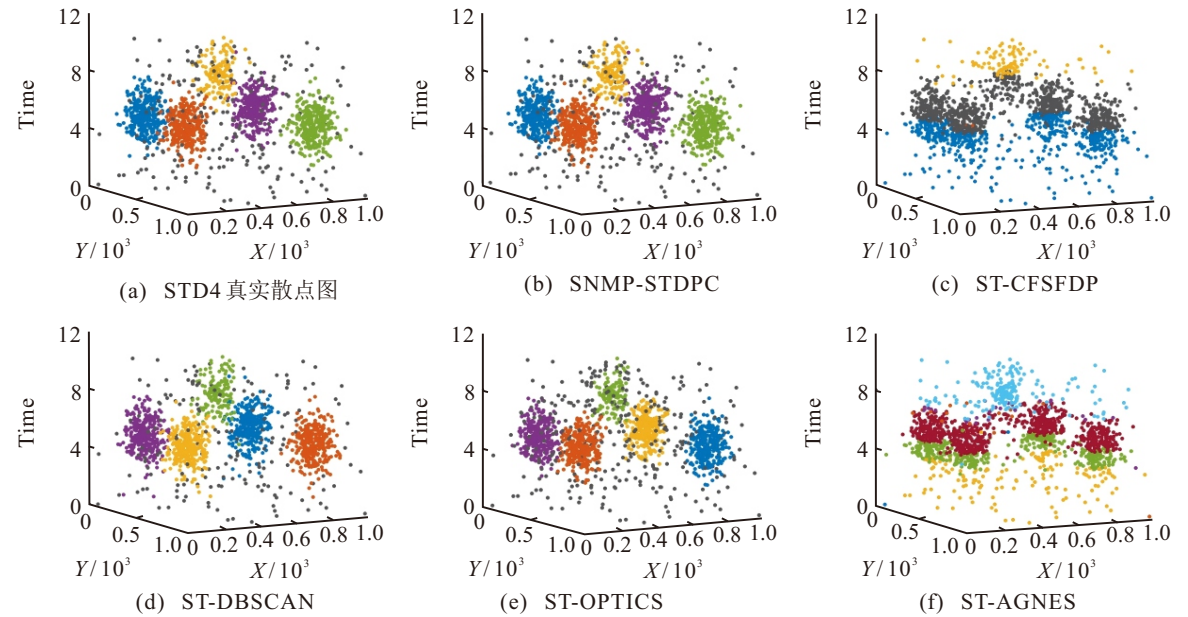


图6 5种算法在数据集STD4的时空聚类结果

本误判为噪声. ST-AGNES 算法因噪声识别能力不足, 使噪声被错误划分至邻近类簇中. ST-OPTICS 和 ST-DBSCAN 算法虽评价指标表现良好, 但在处理类簇边界样本时存在偏差, 其聚类性能略逊于 SNMP-STDPC 算法.

数据集 STD4 包含 5 个显著类簇, 其中底部的 4 个类簇规模相似, 呈现时间相邻但空间独立的分布模式, 顶部的较小类簇在时间维度上与其他类簇显著分离, 但空间层面与其余 4 个类簇保持邻近. 由图 6 分析可知, ST-CFSFDP 算法基于决策图识别出 5 个类簇中心, 但簇心合并后仅保留 3 个. 由于分配策略对空间分布突变敏感, 一个预设类簇对应区域的样本在分配初期即被误分配至相邻类簇, 导致最终形成的该类簇样本规模显著偏小; ST-AGNES 算法仅依据时间维度将数据集划分为 4 个聚类区域, 导致聚类结果在空间特征匹配度上表现显著不足. ST-OPTICS 和 ST-DBSCAN 算法因缺乏独立噪声判别机制, 其性能表现略逊于 SNMP-STDPC 算法. SNMP-STDPC 算法通过异常分数噪声检测标记噪声样本, 相比其他算法在数据集上展现出更优的聚类效果.

3.3 真实数据集的实验结果分析

本节运用 SNMP-STDPC 算法对地震数据集进行时空聚类分析, 能够识别出强震的前震和余震^[32]. 前震通常指主震发生前出现的相对较小规模的地震活动, 而余震则指主震发生后的系列地震现象. 通过对地震序列的聚类研究, 揭示了强震的形成机制, 同时为大地震的预测提供了重要参考依据^[33]. 选用的地震数据集涵盖了 2022 年 1 月 ~ 2023 年 12 月期间我国境内震级超过 4 M 的 602 次地震事件.

SNMP-STDPC 算法在地震数据集上的时空聚类结果如图 7 所示, 识别出了 3 个显著类簇, 分别为橙色类簇 STC1、黄色类簇 STC2 和蓝色类簇 STC3, 其统计特征详见表 4. 类簇 STC1 以 2022 年 3 月 23 日台东海域地震为主震, 包含 71 次余震; 类簇 STC2 共记录 126 次地震, 主震为 2022 年 9 月 18 日

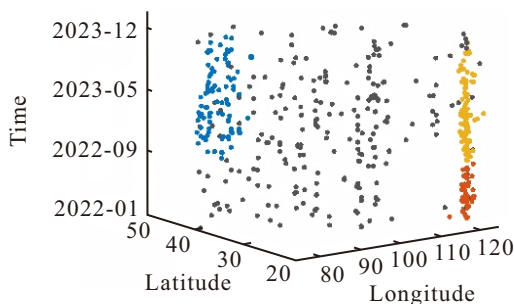


图7 SNMP-STDPC 算法在地震数据集上时空聚类结果

发生的台湾花莲地震, 在其 24 小时前发生了一起 7.2 M 强震, 还包括 6 次前震, 其余 118 次地震均为余震; 类簇 STC3 揭示了 2023 年 1 月 30 日新疆沙雅地震序列的时空聚集特征, 检测到含 120 次前震-余震序列.

表4 地震数据集聚类结果统计信息

类簇	地震数	地震名称	强震位置	震级
STC1	72	台东海域地震	23.45°N, 121.55°E	7.2 M
STC2	126	花莲地震	23.15°N, 121.3°E	7.5 M
STC3	121	沙雅地震	40.1°N, 82.29°E	6.1 M

4 结论

ST-CFSFDP 算法的局部密度计算采用截断核方法, 难以准确甄别密度峰值; 非密度峰值分配因依赖局部邻域关系且缺乏时空约束, 易产生分配偏差; 噪声识别策略受限分配过程的准确性, 噪声检测能力较弱. 针对上述问题, 本文提出一种共享近邻加权和隶属点分配的时空密度峰值聚类算法. 该算法定义时空共享近邻加权的局部密度, 提高了类簇中心选择的可靠性; 设计隶属点分配策略, 融合时空双重约束机制, 增强了非密度峰值分配的准确性; 构建异常分数噪声检测模型, 通过量化噪声可能性, 提升了算法检测噪声的能力. 实验结果表明, SNMP-STDPC 算法在局部密度计算、非密度峰值分配及噪声识别精度等方面均有显著提升. 未来仍需进一步优化算法的时间复杂度, 并探索如何自适应选择参数, 以提高其在实际问题上的应用效果.

参考文献 (References)

- [1] Tosi D, Kokaj R, Rocchetti M. 15 years of big data: A systematic literature review[J]. *Journal of Big Data*, 2024, 11(1): 73.
- [2] Gu Y J, Sun W H, Yu T C, et al. Peace river induced seismic monitoring (PRISM) nodal seismic array[J]. *Seismological Research Letters*, 2025, 96(1): 562-575.
- [3] Yongheng D, Shan X, Fei L, et al. GIS-based assessment of spatial and temporal disparities of urban health index in Shenzhen, China[J]. *Front Public Health*, 2024, 12: 1429143.
- [4] 赵嘉, 段发祥, 潘正祥, 等. 基于时空加权和密度峰值的轨迹聚类算法[J]. *控制与决策*, 2026, 41(2): 470-480. (Zhao J, Duan F Y, Pan J S, et al. Trajectory clustering algorithm based on spatio-temporal weighting and density peak[J]. *Control and Decision*, 2026, 41(2): 470-480.)
- [5] 刘金培, 罗瑞, 陈华友, 等. 基于多尺度 3D-CNN-CBAM 的空气质量指数时空预测[J]. *控制与决策*, 2025, 40(2): 404-412. (Liu J P, Luo R, Chen H Y, et al. Spatio-temporal forecasting of air quality index based on multi-scale 3D-

- CNN-CBAM[J]. *Control and Decision*, 2025, 40(2): 404-412.)
- [6] Hamdi A, Shaban K, Erradi A, et al. Spatiotemporal data mining: A survey on challenges and open problems[J]. *Artificial Intelligence Review*, 2022, 55(2): 1441-1488.
- [7] Jin J L, Liu P, Huang H L, et al. Analyzing urban traffic crash patterns through spatio-temporal data: A city-level study using a sparse non-negative matrix factorization model with spatial constraints approach[J]. *Applied Geography*, 2024, 172: 103402.
- [8] 徐天慧, 郭强, 张彩明. 一种采用串行自编码器的时序数据异常检测方法[J]. *控制与决策*, 2023, 38(12): 3507-3515.
(Xu T H, Guo Q, Zhang C M. A serial autoencoders based method for detecting time series anomalies[J]. *Control and Decision*, 2023, 38(12): 3507-3515.)
- [9] 王迎春, 王志硕, 刘洋, 等. 基于联邦学习的海上分布式光伏超短期功率预测[J]. *控制与决策*, 2025, 40(2): 441-450.
(Wang Y C, Wang Z S, Liu Y, et al. Ultra-short-term power prediction of offshore distributed PV based on federated learning[J]. *Control and Decision*, 2025, 40(2): 441-450.)
- [10] Portugal I, Alencar P, Cowan D. A framework for spatial-temporal cluster evolution representation and analysis based on graphs[J]. *Scientific Reports*, 2024, 14(1): 22137.
- [11] Wu Y, Shi B, Zheng Z, et al. Adaptive spatiotemporal neural networks through complementary hybridization[J]. *Nature Communications*, 2024, 15(1): 7355.
- [12] Ansari M Y, Mainuddin, Ahmad A, et al. Spatiotemporal trajectory clustering: A clustering algorithm for spatiotemporal data[J]. *Expert Systems with Applications*, 2021, 178: 115048.
- [13] Moon J, Kim M, Jung I. Optimizing the maximum reported cluster size for the multinomial-based spatial scan statistic[J]. *International Journal of Health Geographics*, 2023, 22(1): 30.
- [14] Cesario E, Lindia P, Vinci A. Multi-density crime predictor: An approach to forecast criminal activities in multi-density crime hotspots[J]. *Journal of Big Data*, 2024, 11(1): 75.
- [15] Zhao B, Wu G S, Li J J, et al. Spatio-temporal heterogeneous ensemble learning method for predicting land subsidence[J]. *Applied Sciences*, 2024, 14(18): 8330.
- [16] Mann S K, Chawla S. A proposed hybrid clustering algorithm using K-means and BIRCH for cluster based cab recommender system (CBCRS)[J]. *International Journal of Information Technology*, 2023, 15(1): 219-227.
- [17] He S T, Wang D J, Zhao P, et al. Dynamic simulation of debris flow waste-shoal land use based on an integrated system dynamics-geographic information systems model[J]. *Land Degradation & Development*, 2022, 33(12): 2062-2075.
- [18] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data[J]. *Data & Knowledge Engineering*, 2007, 60(1): 208-221.
- [19] 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.
- [20] Liu Q L, Deng M, Bi J T, et al. A novel method for discovering spatio-temporal clusters of different sizes, shapes, and densities in the presence of noise[J]. *International Journal of Digital Earth*, 2014, 7(2): 138-157.
- [21] Leiva L A, Vidal E. Warped K-Means: An algorithm to cluster sequentially-distributed data[J]. *Information Sciences*, 2013, 237: 196-210.
- [22] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. *Science*, 2014, 344(6191): 1492-1496.
- [23] 赵嘉, 马清, 肖人彬, 等. 面向流形数据的共享近邻密度峰值聚类算法[J]. *智能系统学报*, 2023, 18(4): 719-730.
(Zhao J, Ma Q, Xiao R B, et al. Density peaks clustering based on shared nearest neighbor for manifold datasets[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(4): 719-730.)
- [24] 陈蔚昌, 赵嘉, 肖人彬, 等. 面向密度分布不均数据的近邻优化密度峰值聚类算法[J]. *控制与决策*, 2024, 39(3): 919-928.
(Chen W C, Zhao J, Xiao R B, et al. Density peaks clustering algorithm with nearest neighbor optimization for data with uneven density distribution[J]. *Control and Decision*, 2024, 39(3): 919-928.)
- [25] 吕莉, 赵姐, 肖人彬, 等. 混合近邻和多簇合并的密度峰值聚类算法[J]. *控制与决策*, 2025, 40(7): 2194-2202.
(Lv L, Zhao N, Xiao R B, et al. Density peak clustering algorithm with mixed nearest neighbors and multi-cluster merging[J]. *Journal of Frontiers of Computer Science and Technology*, 2025, 40(7): 2194-2202.)
- [26] 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 P X, Zhang H C, Wang H B, et al. Spatial-temporal clustering by fast search and find of density peaks[J]. *Acta Geodaetica et Cartographica Sinica*, 2019, 48(11): 1380-1390.)
- [27] Zhao J, Wang G, 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.
- [28] 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 P X, Wang H B, Fu M Y, et al. Research on semantic location prediction of indoor users[J]. *Journal of Geo-information Science*, 2018, 20(12): 1689-1698.)
- [29] Romano S, Bailey J, Nguyen V, et al. Standardized mutual information for clustering comparisons: one step

- further in adjustment for chance[J]. *International Conference on Machine Learning*, 2014, 32(2): 1143-1151.
- [30] Vinh N X, Epps J, Bailey J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance[J]. *Journal of Machine Learning Research*, 2010, 11(1): 2837-2854.
- [31] 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.
- [32] Peng Z G, Lei X L, Wang Q Y, et al. The evolution process between the earthquake swarm beneath the Noto Peninsula, central Japan and the 2024 M 7.6 Noto Hanto earthquake sequence[J]. *Earthquake Research Advances*, 2025, 5(1): 100332.
- [33] Nicolis O, Delgado L, Peralta B, et al. Space-time clustering of seismic events in Chile using ST-DBSCAN-EV algorithm[J]. *Environmental and Ecological Statistics*, 2024, 31(2): 509-536.

作者简介

赵嘉 (1981-), 男, 教授, 博士, 主要研究方向为机器学习、数据挖掘、智能计算, E-mail: zhaojia@nit.edu.cn;

朱伟涛 (1999-), 男, 硕士生, 主要研究方向为数据挖掘, E-mail: 2023313024@nit.edu.cn;

肖人彬 (1965-), 男, 教授, 博士, 博士生导师, 主要研究方向为复杂系统建模与分析、群集智能, E-mail: rbxiao@hust.edu.cn;

李思维 (1988-), 男, 高级工程师, 博士, 主要研究方向为电力负荷管理、电力物联网、微电网技术, E-mail: lisiwei@sgtg.sgcc.com.cn;

白晖峰 (1984-), 男, 高级工程师, 博士, 主要研究方向为电力物联网, E-mail: baihuifeng1984@163.com;

谢海华 (1994-), 男, 讲师, 博士, 主要研究方向为群智能优化算法, E-mail: xie_haihua@nit.edu.cn;

夏志红 (1977-), 女, 讲师, 硕士, 主要研究方向为智慧水利, E-mail: xzh@nit.edu.cn.