智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于混合邻域图的复杂结构数据集层次聚类算法

陈仲尚, 冯骥, 杨德刚, 蔡发鹏

引用本文:
陈仲尚, 冯骥, 杨德刚, 等. 基于混合邻域图的复杂结构数据集层次聚类算法[J]. 智能系统学报, 2025, 20(3): 584-593.
CHEN Zhongshang, FENG Ji, YANG Degang, et al. Hybrid neighborhood graph-based hierarchical clustering algorithm for datasets with complex structures[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(3): 584-593.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202407001

DOI: 10.11992/tis.202407001
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250418.1434.005

基于混合邻域图的复杂结构数据集层次聚类算法

陈仲尚, 冯骥, 杨德刚, 蔡发鹏
(重庆师范大学 计算机与信息科学学院, 重庆 401331)

摘 要: 复杂结构数据集通常指包含不同形状 (如球形、非球形、流形)、大小和密度的簇的数据集。自然邻居算法在处理边界模糊、密度变化的数据集时存在局限性, 特别是在数据集中含有大量噪声时, 其性能会显著下降。针对这些问题, 本文提出一种基于混合邻域图的复杂结构数据集层次聚类算法 (hybrid neighborhood graph-based hierarchical clustering algorithm for datasets with complex structures, HCHNG)。该方法提出一种共享自然邻域图方法, 通过邻居关系稀疏数据集以减少噪声样本对聚类结果的影响。随后, HCHNG 将数据集划分为子图并加以合并, 这一策略增强了算法处理变密度数据集的能力, 同时, 定义一种新的子图相似性度量方法, 提高同类子图间的相似性。此外, 对自然邻域图进行改进, 以提升其在识别边界模糊数据集时的性能。在具有复杂结构的人工数据集和真实数据集上的对比实验表明, 本文算法不仅能有效识别变密度球形数据集, 而且在含有大量噪声的复杂数据集中也拥有优越的性能, 在处理具有复杂结构的数据集时比现有方法高效。

关键词: 聚类分析; 混合邻域图; 共享自然邻居; 改进的自然邻域图; 共享自然邻域图; 子图相似性; 复杂数据集; 数据挖掘
中图分类号: TP301 **文献标志码**: A **文章编号**: 1673−4785(2025)03−0584−10

中文引用格式: 陈仲尚, 冯骥, 杨德刚, 等. 基于混合邻域图的复杂结构数据集层次聚类算法 [J]. 智能系统学报, 2025, 20(3): 584–593.
英文引用格式: CHEN Zhongshang, FENG Ji, YANG Degang, et al. Hybrid neighborhood graph-based hierarchical clustering algorithm for datasets with complex structures[J]. CAAI transactions on intelligent systems, 2025, 20(3): 584–593.

Hybrid neighborhood graph-based hierarchical clustering algorithm for datasets with complex structures

CHEN Zhongshang, FENG Ji, YANG Degang, CAI Fapeng
(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: Complex structured datasets typically refer to datasets containing clusters of different shapes (including spherical, non-spherical, and manifold shapes), sizes, and densities. The natural neighbor algorithm exhibits limitations in handling datasets with unclear boundaries and varying densities. Particularly, its performance decreases significantly when the dataset contains a significant amount of noise. To address this drawback, we propose a hybrid neighborhood graph-based hierarchical clustering algorithm for datasets with complex structures (HCHNG). We proposed a method of shared natural neighborhood graph, which uses the neighbor relationships to sparse the dataset and reduce the impact of abnormal samples on clustering results. Subsequently, the algorithm divides the dataset into several subgraphs and enhances the processability of variable density data by merging operations. Concurrently, we propose a new method for defining subgraph similarity, which ensures higher similarity between subgraphs of the same class. Additionally, we improve the performance of the natural neighborhood graph in identifying datasets with blurred boundaries. The experimental results reveal that the HCHNG algorithms can recognize variable density spherical datasets and complex datasets containing a large amount of noise. Therefore, our algorithm is more effective than the existing methods in processing datasets with complex structures.

Keywords: cluster analysis; hybrid neighborhood graph; shared natural neighbors; improved natural neighborhood graph; shared natural neighborhood graph; subgraph similarity; complex dataset; data mining

在大数据时代, 如何从海量数据中提取有价值的信息一直是研究的焦点。数据挖掘作为一种辅助决策过程[1], 能够高度自动化地分析数据, 通

收稿日期: 2024−07−01. **网络出版日期**: 2025−04−18.
基金项目: 重庆市教委科学技术研究项目 (KJZD-M202300502, KJQN201800539).
通信作者: 冯骥. E-mail: jifeng@cqnu.edu.cn.

过归纳推理来挖掘潜在模式, 从而帮助用户调整策略[2]、降低风险[3], 并做出正确的决策。

聚类方法通常根据其特性进行分类, 包括基于分区的聚类[4]、基于密度的聚类[5-6]和层次聚类等[7-8]。其中, 层次聚类算法是聚类算法中性能较好的一种算法。它通过分层分解数据集并进行合并, 直至达到预期的聚类效果, 以此构建聚类的层次结构。BIRCH(balanced iterative reducing and clustering using hierarchies)、CURE(clustering using representatives) 以及 Chameleon 是经典的层次聚类算法[9-11]。BIRCH 算法通过分层划分来构建聚类特征树, 只需对数据集进行一次扫描就能获得相对较好的聚类结果。然而, 当处理复杂结构数据集时, BIRCH 的性能并不理想。CURE 算法首先随机创建分区, 并在分区中随机选择样本点作为聚类中心, 然后按照一定的比例将样本点向聚类中心收缩, 使得算法能够识别非球形聚类。尽管如此, CURE 在处理包含噪声的数据集时仍存在困难。Chameleon 算法首先构造 K-近邻图, 然后利用 hMctis 算法[12]分解数据集, 最后根据子簇相似性从高到低进行合并 (子簇相似性由相互连通性和紧密性构成)。由于 Chameleon 算法通过对聚类的互联性和紧密性进行选择来衡量聚类的相似性, 因此它对形状、大小和密度差异较大的聚类具有敏锐的洞察力。然而, Chameleon 算法需要首先构造 K-近邻图, 不同的 K 值对聚类结果的影响较大, 且 hMetis 算法环境相当具有挑战性。吕端端[13]提出利用自然邻域加权图代替 Chameleon 中的 K-近邻图。Cheng 等[14]提出一种基于模块化对自然邻域图进行划分的层次聚类算法。在层次合并的阶段, 该算法参考 Chameleon 算法的子簇的互连性和紧密性, 使其在识别任何形状的数据集中具有更好的效果。

基于分区的聚类方法因其高效性而广受欢迎, 其中 K-means 算法应用尤为广泛。该算法通过随机选取初始聚类中心, 并迭代更新直至达到稳定状态。然而, 由于其对初始聚类中心的随机选择, K-means 对噪声和异常值敏感, 且在处理非凸簇时效果欠佳。为改善初始中心点的选择, K-means++算法[15]在选定一个中心点后, 会优先选择距离已选中心点较远的样本作为新的中心点, 从而提升了初始中心点的质量, 但依旧存在一定的对噪声敏感问题。而 Cheng 等[16]提出了利用自然密度峰 (natural density peaks, NDPs) 改进初始聚类中心的选择。首先, NDP-Kmeans 计算样本点的密度, 并暂时移除低密度点, 以降低噪声对聚类结果的影响; 然后, 将 NDPs 作为 K-means 的初始聚类中心, 并结合图距离, 优化了样本点到聚类中心的分配策略, 使其能有效发现各种形状的聚类。另一方面, Tzortzis 等[17]改进了 K-means 的分配策略, 通过计算所有子聚类的方差并为其分配权重, 并通过同时学习权重和迭代过程进行聚类分配来优化 K-means 目标函数, 但在处理含有大量噪声的数据集时, 该算法仍不理想。

基于密度的聚类方法假设聚类是由稀疏区域分隔的密集区域构成。DBSCAN(density-based spatial clustering of applications with noise)[18]作为典型的基于密度的聚类方法, 依赖两个关键参数: 局部邻域半径和邻域内点数。DBSCAN 在检测任意形状簇方面表现出色, 但对参数设置高度敏感。RNN-DBSCAN(recurrent neural network DBSCAN)[19]利用反向最近邻策略减少参数的依赖。Gholizadeh 等[20]则通过分析样本与其第 K 个最近邻居之间的距离分布来优化参数选择。密度峰值聚类算法 (density peak clustering, DPC)[21]是一种简单高效的基于密度算法, 它选择密度最高且截止距离最小的点作为聚类中心, 然后依据最近距离且密度较大的样本来确定所有余下样本的聚类标签。然而, DPC 算法难以确定合适的截止距离。Tong[22]和 Chen[23]等运用自然邻居的概念[24]避免了设定截止距离的问题。张清华和位雅等[25-26]通过改进 DPC 的密度概念, 提高了 DPC 算法识别聚类中心的能力。Cheng 等[27]将粒球概念引入到聚类算法中, 提出了 GBDPC(granular-ball-based density peaks clustering) 算法。该算法将数据集划分为多个粒球, 并按照 DPC 算法进行合并, 直至达到预定的簇数量, 从而提升了 DPC 算法的处理速度。Liu 等[28]提出的 SNNDPC(shared-nearest-neighbor DPC) 算法基于共享邻居重新定义了密度, 减少了人为选择中心点的随机性, 增强了算法的鲁棒性。此外, SNNDPC 采用基于共享邻居的两阶段方法分配剩余样本, 从而提升了 DPC 算法发现任意形状聚类的能力。

然而, 上述算法在处理包含复杂数据的数据集时均难以取得较好的聚类结果。现有邻域图通常在单个结构的数据集表现优秀, 但在处理复杂结构数据集时, 性能下降显著。通过大量文献调研发现, 基于图的聚类算法相对较少, 并且在识别复杂结构数据集时性能较弱。因此, 本文提出了一种基于混合邻域图的层次聚类算法 (hybrid neighborhood graph-based hierarchical clustering algorithm for datasets with complex structures, HCHNG)。

HCHNG 充分利用邻域图的特性，其主要创新为：1) 提出邻域图方法——共享自然邻域图，该图能有效识别变密度和噪声数据集的分布规律。2) 改进自然邻域图，该图更有效地表达稀疏数据集的分布状况。3) 提出衡量子图相似性的方法，该方法最大化两个相似子图间的相似性。4) 提出基于混合邻域图的层次聚类算法。在合成数据集、噪声数据集和真实数据集的实验表明，HCHNG 算法在处理具有复杂结构的数据集时优于现有方法。

1　相关工作

1.1　自然邻居

自然邻居方法能够自动适应数据集的分布规律，从而无需手动选择 K 参数。自然邻居已在聚类分析、实例近似和异常值检测等领域得到广泛应用[29-31]。其核心思想主要体现在 3 个方面：邻居、搜索算法和邻居数量。

定义 1　(稳定搜索状态) 当数据集进入稳定状态时，对于任意的样本点 x_i 和 x_j，都有：

$$(\forall x_i)\left(\exists x_j\right)\wedge\left(x_i\neq x_j\right)\rightarrow \left(x_i\in \mathrm{NN}_\lambda\left(x_j\right)\wedge\left(x_j\in \mathrm{NN}_\lambda(x_i)\right)\right) \tag{1}$$

式中：参数 λ 表示自然邻居算法获得的自然特征值，$\mathrm{NN}_\lambda(x_i)$ 表示 x_i 的 λ 最近邻，$\mathrm{NN}_\lambda\left(x_j\right)$ 表示 x_j 的 λ 最近邻。

定义 2　(自然邻居, natural neighbors, NaN) 假设搜索状态在 λ^{th} 轮搜索后稳定下来。对于任何样本 x_i 和 x_j，如果它们是彼此的邻居，则它们有关系：

$$x_j\in \mathrm{NaN}(x_i)\Leftrightarrow\left(x_i\in \mathrm{NN}_\lambda\left(x_j\right)\right)\wedge\left(x_j\in \mathrm{NN}_\lambda(x_i)\right) \tag{2}$$

自然邻居具有不变性和稳定性。不变性意味着，如果 x_i 在算法搜索过程中处于 x_j 的自然邻居集 ($\mathrm{NaN}\left(x_j\right)$) 中，当算法达到稳态时，$x_i$ 仍然是 x_j 的自然邻居。稳定性意味着，对于相同的数据集，无论算法重复多少次，自然邻居搜索算法的样本获得的自然邻居集都保持不变。

1.2　共享最近邻居

样本及其相邻样本通常属于同一聚类类别，从而可以基于其相邻信息更准确地评估样本的局部密度[32]。样本及其邻居的共享区域提供了丰富的局部信息，有助于更准确地描述样本的分布。

定义 3　(共享最近邻居)[33] 对于数据集 X 中的任何样本 x_i 和 x_j，它们的共享最近邻居表示为

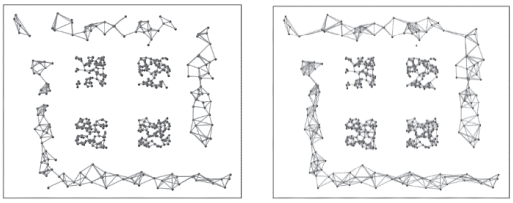
$$\mathrm{SNN}(x_i,x_j)=\mathrm{NN}_k(x_i)\cap \mathrm{NN}_k(x_j) \tag{3}$$

2　本文提出的算法

邻域图是表示数据集分布特征的一种重要方法。常见的邻域图识别复杂结构数据集的性能并不理想。在阅读了大量与聚类相关的论文后发现，现有的基于图的聚类方法在识别含噪声的复杂结构数据集时准确性不高。基于这种情况，本文提出了一种基于混合自然邻域图的复杂结构数据集层次聚类算法。

2.1　改进的自然邻域图

自然邻域图能够自适应地表达数据集的分布规律，因而获得广泛应用。许多研究者提出了基于自然邻域图的扩展算法[34-35]，然而，关于自然邻域图的改进研究却相对较少。改进前后自然域图对比如图 1。



(a) 原始自然邻域图　　(b) 改进后的自然邻域图

图 1　改进前后自然邻域图对比

Fig. 1　Comparison of original natural neighborhood graphs and improved natural neighborhood graphs

如图 1(a) 所示，自然邻域图的子图数量明显比实际的子图数量要多。虽然自然邻域图能自适应地表达数据集的分布，但容易将边界稀疏的同一子簇划分为多个子簇，从而影响后续的子图合并过程。为解决这个问题，Zhang 等[36] 提出了一种改进自然邻域图的定义。本研究提出一种类似方法，用于改进自然邻域图 (improved natural neighborhood graphs, INaNG)。首先，在自然邻居搜索过程中，若满足定义 1 则认为自然邻居状态已经达到稳定。然后，在自然邻居算法中增加了对子图数量 (N_{INaNG}) 的判断。由于构造共享自然邻域图时会将仅含 2 个样本的子图识别为异常值，故在 INaNG 中，样本总数为 2 的子图不计入 N_{INaNG} 中。在达到自然邻居算法的稳定状态后，如果 $N_{\mathrm{INaNG}}\leqslant n_{\mathrm{c}}$($n_{\mathrm{c}}$ 为真实的子图数量)，则认为它是最终的稳定状态。但如果 $N_{\mathrm{INaNG}}>n_{\mathrm{c}}$，让 λ+1，并让自然邻居算法继续下一轮搜索，直到 $N_{\mathrm{INaNG}}\leqslant n_{\mathrm{c}}$。尽管改进后的自然邻域图引入了新参数，但 HCHNG 法在合并子图时也需使用参数 n_{c}，故整体算法未增加额外参数。改进后的算法描述如下。

算法 1　改进自然邻域图算法

输入　数据集 X，由用户输入的真实子图数量 n_{c}。

输出　改进后的自然邻域图 INaNG，自然邻居集合 NaN。

1) 基于自然邻居的概念寻找稳定状态。

2) 判断自然邻域图的子图数量 (N_{INaNG}) 是否

大于 n_c。如果子图中有由两个样本点组成的子图, N_{INaNG}的数量减 1。如果$N_{\mathrm{INaNG}}>n_c$, λ+1, 继续搜寻下一轮稳定的自然邻居状态, 直到$N_{\mathrm{INaNG}}\leqslant n_c$。

3) 在具有稳定自然邻居关系的样本之间添加边。

4) 返回改进后的自然邻域图和自然邻居集合。

2.2　共享自然邻域图

自然邻居的思想是, 如果两个样本有双向的关系, 那这对邻居的关系更加紧密。共享邻居的思想是, 如果样本点之间有许多公共的样本, 它们的关系就越紧密。共享自然邻居结合了自然邻居和共享邻居的特性, 更有效地表达复杂数据集的分布结构。其相关定义如下。

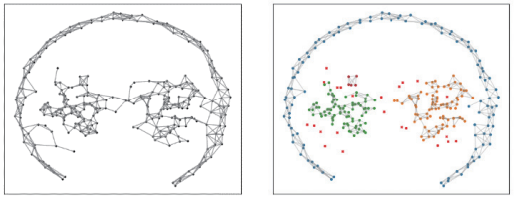
定义 4　(共享自然邻居) 对于任意两个自然邻居样本n_i和n_j, $\mathrm{NaN}(n_i)$是n_i的自然邻居的集合, $\mathrm{NaN}(n_j)$是n_j的自然邻居集合。因此, n_i和n_j的共享自然邻居为

$$\mathrm{SNaN}\left(n_i,n_j\right)=\mathrm{NaN}\left(n_i\right)\cap\mathrm{NaN}\left(n_i\right) \tag{4}$$

定义 5　(共享自然邻域图, shared natural neighborhood graph, SNaNG) 共享自然邻域图是通过连接自然邻居n_i和自然邻居n_j的共享邻居而形成的。SNaNG 是一个有向图, G=(V, E)。

$$\begin{cases} V=X \\ E=\{((n_i,n_j),\mathrm{SNaN}(n_i,n_j,X)),n_i,n_j\in X\} \end{cases} \tag{5}$$

如图 2(a) 所示, 自然邻域图将数据集识别为一个子图。图 2(b) 表示的共享自然邻域图可以将子图划分为几个子图和异常值。



(a) 改进后的自然邻域图　　(b) 共享自然邻域图

图 2　两种邻域图在 Path 数据集上的结果

Fig. 2　Results of two neighborhood graphs on the Path dataset

2.3　子图相似性

将数据集划分为多个子图并合并相似的子图是帮助识别复杂结构化数据集的有效策略[37-39]。然而, 如何确定子图之间的相似性是一个关键问题。本研究提出了一种新的方法来计算子图的相似性, 并将其定义如下。

定义 6　(子图间的最短距离) 计算子图G_i和G_j中所有样本点之间的距离, 然后按升序对这些距离进行排序。此外, 为了减轻异常值的影响, 参考了 Ding 的研究[40], 只选择G_i和G_j中最短距离的最小值 5% 来计算$G_n(n_i,n_j)$, 表示为

$$D_{\min}\left(G_i,G_j\right)=\frac{\sum_{1}^{G_n\left(n_i,n_j\right)}\sum_{n_i\in G_i,n_j\in G_j}d_{\mathrm{sort}}\left(n_i,n_j\right)}{G_n\left(n_i,n_j\right)} \tag{6}$$

定义 7　(子图间的桥) 对于子图G_i和G_j, 计算两个子图中所有λ最近邻居的交集。当两个子图之间的 λ 最近邻居的交集不为空时, 它表示一定程度的连通性。子图间的桥梁表示为

$$B_{\mathrm{cluster}}(G_i,G_j)=\sum_{n_i\in G_i}\sum_{n_j\in G_j}(\mathrm{NaN}(n_i)\cap\mathrm{NaN}(n_j)) \tag{7}$$

定义 8　(子图间的相似性) 当子图间的桥的数量越多, 子图间的最短距离越短, 子图间的相似性越大。子图间的相似性表示为

$$G_{\mathrm{Sim}}(G_i,G_j)=\frac{B_{\mathrm{cluster}}(G_i,G_j)+\lambda}{D_{\min}(G_i,G_j)} \tag{8}$$

2.4　基于混合邻域图的层次聚类算法

改进后的自然邻域图在处理边界清晰的流形数据集时展现出较强的表达能力, 然而在识别变密度数据集时, 其性能显著下降。共享自然邻域图通过综合考虑单个样本与两个样本之间共享的关系, 能更有效地描述变密度数据集中样本分布。因此, 本研究提出的算法同时利用这两种邻域图来揭示数据集中的结构规律, 使得算法能够适应不同类型的数据分布, 提高了对复杂结构数据集进行准确聚类的能力。

HCHNG 的主要步骤包括: 1) 构造改进自然邻域图 (INaNG)。2) 判断子图的数量: 如果改进 INaNG 的子图数量等于真实的聚类数量 (n_c), 输出为最终聚类结果; 如果改进后 INaNG 子图数量小于 n_c, 则构造共享自然邻域图 (SNaNG) 对 INaNG 进行分割。3) 基于 INaNG 构造 SNaNG, 将数据集划分为子图和异常样本: 划分子图需要设置一个切边阈值 (G_{cut}), 其中G_{cut}为自然邻居共享的邻居数量。构造 SNaNG 后, 初始化G_{cut}=0。循环判断 SNaNG 的子图数量 (N_{SNaNG}) 是否大于 n_c。如果$N_{\mathrm{SNaNG}}<n_c$, G_{cut}+1, 直至$N_{\mathrm{SNaNG}}\geqslant n_c$。将不小于$G_{\mathrm{cut}}$的共享自然邻域边连接, 得到 SNaNG, 小于$G_{\mathrm{cut}}$的样本为异常样本。4) 合并最相似的子图并分配异常样本: 将子图相似性从高到低排序, 重复合并两个相似性高的子图, 直到达到所需的子图数量; 将异常样本分配给它们最近的已聚类的样本中。本文提出的基于混合邻域图的层次聚类算法流程如算法 2 所示。

算法 2　HCHNG 算法

输入　数据集 X, 由用户输入的真实子图数量 n_c。

输出　返回最终的聚类结果。

1) 初始化分割图的阈值G_{cut}=0; 初始化异常样本集合I_{m}=∅; 初始化子图相似性矩阵$G_{\mathrm{Sim}}\left(G_i,G_j\right)$=∅;

2) 根据算法 1 获得 INaNG, 自然邻居集合 NaN;

3) 判断 INaNG 的子图数量, 如果 INaNG 的子图数量为 n_c, 直接跳到 11), 如果 INaNG 的子图数量<n_c, 进入 4);

4) 根据定义 4 计算共享自然邻居 SNaN (n_i, n_j), 在共享自然邻居添加边并连接, 得到共享自然邻域图 SNaNG;

5) 判断 SNaNG 的子图数量。循环判断, 如果 SNaNG 的子图数量小于 n_c, G_{cut}+1, 直至 SNaNG 的子图数量不小于 n_c;

6) 将共享自然邻居数不小于G_{cut}的共享自然邻域边连接, 得到 SNaNG。共享自然邻居数小于G_{cut}的共享自然邻域边加入I_m中;

7) 根据定义 8 计算子图之间的相似性;

8) 选择相似性最高的两个子图进行合并, 并更新子图相似性;

9) 重复 7) ~ 8), 直至合并至 n_c 个子图;

10) 使用 KD-Tree 找到距离I_m最近的样本 p。如果 p 已被聚类, 将I_m分配至 p 所属的子图中;

11) 返回最终的聚类结果。

2.5　时间复杂度分析

由于自然邻居搜索算法中应用了 KD-tree, 因此改进后的自然邻域图的时间复杂度为$O(n\cdot\log(n))$, 其中 n 是数据集的大小。获得改进自然邻域图后, 如果需要构造共享自然邻域图, 构造共享自然邻域图的时间复杂度为 $(n\cdot\log(n))$。子图合并的时间复杂度为$O(n-n_c)^n$。因此, 如果需要构造共享自然邻域图, 则 HCHNG 的总体时间复杂度为 $O(n\cdot\log n+2n)$。假设n特别大, 则 HCHNG 算法的时间复杂度是$O(n^2)$。

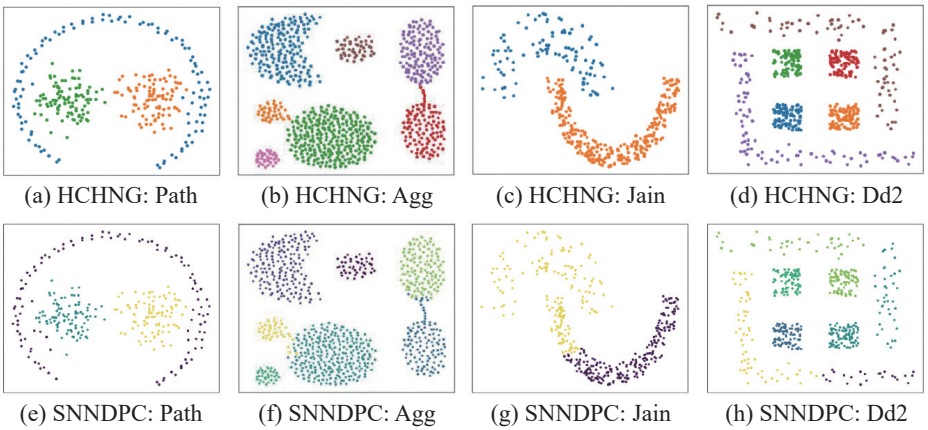
3　实验结果与分析

基于图的聚类算法需要通过可视化方法分析实验结果。为了展示所提出的 HCHNG 算法在识别复杂结构数据集方面的优势, 选取几种经典算法和现有的较新算法, 包括 K-means、DPC、NDP-Kmeans、SNNDPC、LCCV[41] 和 GBDPC 在合成数据集和真实数据集上进行了实验。

各算法参数设置: DPC 算法需设定截断距离 d_c 和在决策图中选择 m_c 个聚类中心。NDP-Kmeans 需要 n_c 个聚类中心和噪声比例 α。K-means 需确定 K 个聚类中心, 为保证准确性, 在使用的数据集上进行 10 次实验并取平均值作比较。SNND-PC 需设 K 值以计算共享邻居关系并选择 m_c 个聚类中心。GBDPC 需在决策图中选定 m_c 个聚类中心。LCCV 为无参算法。HCHNG 仅需一个 m_c 参数。

3.1　合成数据集上的实验结果与分析

为验证 HCHNG 算法处理复杂结构数据集的有效性, 本研究在 8 个不同特性的数据集上进行了实验[42-43]。这些数据集包括 Path、Agg、Jain、Dd2、Atom、Chainlink、T2 和 T4。其中, Path 和 Agg 为二维球形变密度数据集, 前者包含 299 个样本及 3 个子簇, 后者由 787 个样本和 7 个子簇构成。Jain 与 Dd2 是两个流形二维数据集, Jain 包含 373 个样本, 2 个子簇分布不均; Dd2 有 499 个样本, 6 个边缘稀疏的子簇。Atom 和 Chainlink 是两个三维数据集, 它们同样由 499 个样本和 2 个子簇组成。T2 和 T4 是大型二维含噪声数据集, T2 含 7936 个样本及 6 个子簇, T4 有 4 199 个样本和 6 个子簇。由于 SNNDPC 采用共享近邻进行聚类, 故图 3 和图 4 给出了 HCHNG 与 SNNDPC 在合成数据集的聚类结果对比。实验结果总结在表 1 中, 其中调整互指数(adjusted mutual information, AMI)、调整兰德指数(adjusted Rand index, ARI)、福尔克斯-马洛斯指数(Fowlkes-Mallows index, FMI)。

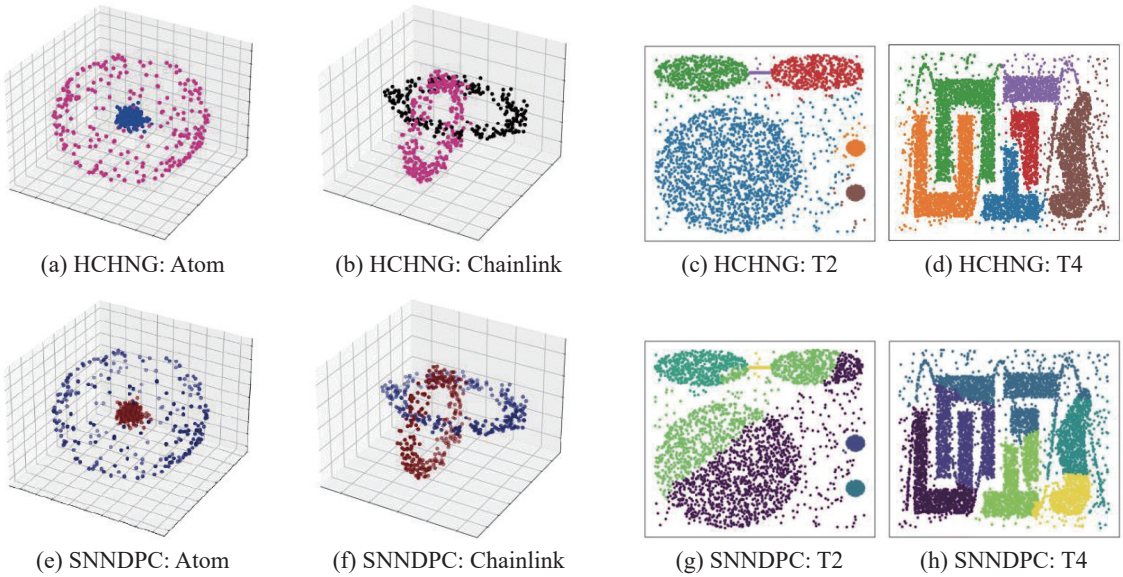


(a) HCHNG: Path　(b) HCHNG: Agg　(c) HCHNG: Jain　(d) HCHNG: Dd2

(e) SNNDPC: Path　(f) SNNDPC: Agg　(g) SNNDPC: Jain　(h) SNNDPC: Dd2

图 3　HCHNG 和 SNNDPC 算法在 Path、Agg、Jain 和 Dd2 数据集的聚类结果

Fig. 3　Clustering results of HCHNG algorithm and SNNDPC algorithm on Path, Agg, Jain and Dd2 datasets



(a) HCHNG: Atom　(b) HCHNG: Chainlink　(c) HCHNG: T2　(d) HCHNG: T4

(e) SNNDPC: Atom　(f) SNNDPC: Chainlink　(g) SNNDPC: T2　(h) SNNDPC: T4

图 4　HCHNG 和 SNNDPC 算法在 Atom、Chainlink、T2 和 T4 数据集的聚类结果

Fig. 4　Clustering results of HCHNG algorithm and SNNDPC algorithm on Atom, Chainlink, T2 and T4 datasets

表 1　7 种聚类算法在合成数据集上的比较

Table 1　Comparison of seven algorithms on synthetic datasets

数据集	评价指标	DPC	GBDPC	LCCV	K-means	NDP-Kmeans	SNNDPC	HCHNG
Path	AMI	0.504	0.493	0.470	0.545	0.5173	0.901	**1**
	ARI	0.420	0.434	0.269	0.464	0.309	0.929	**1**
	FMI	0.649	0.660	0.373	0.663	0.376	0.953	**1**
	Arg-	3,3.8	3	—	3	3,0.1	3,9	3
Agg	AMI	0.934	0.656	0.889	0.839	0.925	0.955	**0.979**
	ARI	0.920	0.536	0.679	0.711	0.778	0.959	**0.986**
	FMI	0.938	0.636	0.403	0.772	0.895	0.968	**0.989**
	Arg-	7,0.9	7	—	7	7,0.3	7,15	7
Jain	AMI	1	0.219	0.506	0.368	**1**	0.435	**1**
	ARI	1	0.059	0.360	0.324	**1**	0.406	**1**
	FMI	1	0.595	0.470	0.701	**1**	0.740	**1**
	Arg-	2,2 09	2	—	2	2,0.1	2,12	2
Dd2	AMI	0.629	0.708	0.906	0.732	0.777	0.848	**1**
	ARI	0.371	0.570	0.750	0.631	0.429	0.735	**1**
	FMI	0.521	0.662	0.868	0.702	0.691	0.785	**1**
	Arg-	6,2.62	6	—	6	6,0.1	6,2	6
T2	AMI	0.693	0.447	0.040	0.655	0.898	0.706	**0.982**
	ARI	0.469	0.288	0.006	0.484	0.759	0.507	**0.978**
	FMI	0.595	0.447	0.494	0.595	0.897	0.624	**0.980**
	Arg-	6,3.59	6	—	6	6,0.1	6,16	6
T4	AMI	0.646	0.186	0.773	0.561	0.873	0.544	**0.837**
	ARI	0.482	0	0.798	0.452	0.765	0.603	**0.872**
	FMI	0.581	-0.001	0.647	0.522	0.889	0.463	**0.860**
	Arg-	6,3.59	6	—	6	6,0.1	6,21	6

注: 加粗表示结果最优, “Arg-”是算法取得最优值时的参数。

表 1 数据表明，HCHNG 算法在 Path、Jain 和 Dd2 数据集上取得了最佳的聚类指标值 1，在 Agg 数据集上的指标也接近最优。面对结构复杂的噪声数据集 T2 和 T4 时，HCHNG 算法相较其他对比算法展现出更优的性能。DPC 在 Agg 和 Jain 这类有明确中心点的数据集上表现尚可，但在处理中心点不明显的 Path、Dd2、Atom 和 Chainlink 时效果欠佳，同时由于对噪声敏感，其在含噪数据集 T2 和 T4 上的表现并不理想。GBDP 通过粒球划分进行聚类，可能与其划分策略有关，在这些数据集上的效果受限。LCCV 在密度分布不均的 Chainlink、Atom 上取得最佳结果，在 Dd2 上也有不错表现，然而在噪声数据集 T2 和 T4 上性能较差。K-means 在非凸数据集和噪声数据集上的聚类效果较差。

如表 1 所示，NDP-Kmeans 算法由于提前处理了低密度点，因此在 T2 和 T4 噪声数据集上表现出色，对于 Agg 和 Jain 这类具有明确中心点的数据集，其性能同样优越。然而，在边缘稀疏的 Dd2 数据集和变密度的 Path 数据集中，NDP-Kmeans 的性能表现不佳。图 3(e)、(f) 与图 4(e)、(f) 显示，SNNDPC 的聚类效果优秀，因为该算法在计算密度时考虑了共享近邻的关系，故而在变密度数据集上表现出了较好的聚类性能，但由于利用共享近邻策略分配样本时，并未考虑噪声样本对聚类结果的影响，这可能导致算法在 T2 和 T4 噪声数据集上的聚类性能下降。如图 3(a) ~ (d) 和图 4(a) ~ (d) 所示，HCHNG 算法不仅考虑了自然邻域的关系，还考虑了共享自然邻域之间的关系，因此对于复杂结构的数据集都展现出了优秀的聚类性能。

3.2　真实数据集上的实验结果与分析

为进一步验证 HCHNG 算法的有效性，本研究在 UCI 机器学习库中选取了 8 个真实数据集[44-45]Wine、Semelon、BCW、Contraceptive-MC、Movement_libras、Dermatology、Mnist-r120 和 Page-blocks 进行了实验，比较了 7 种算法的 AMI、ARI 和 FMI 表现。

本研究涉及的数据集在规模和维度上各不相同。Wine 是 13 维数据集，含 178 个样本，分为 3 簇；Semeion 是 265 维数据集，含 2 693 个样本，分为 2 簇；BCW 是 10 维数据集，含 683 个样本，分为 2 簇；Contraceptive-MC 是 9 维数据集，含 1 473 个样本，分为 3 簇；Movement_libras 是 91 维数据集，含 360 个样本，分为 15 簇；Dermatology 是 34 维数据集，含 358 个样本，分为 6 簇；Mnist-r120 是 500 维数据集，含 810 个样本，分为 7 簇；Page-blocks 是 10 维数据集，含 5 473 个样本，分为 5 簇。表 2 给出了 HCHNG 算法与 6 种算法在这些数据集上的比较结果。

表 2　7 种聚类算法在真实世界数据集上的比较

Table 2　Comparison of seven algorithms on real-world datasets

数据集	指标	DPC	GBDPC	LCCV	NDP-Kmeans	K-means	SNNDPC	HCHNG
Wine	AMI	0.397	0.332	0.393	0.408	0.429	0.876	**0.947**
	ARI	0.293	0.224	0.203	0.167	0.280	0.898	**0.965**
	FMI	0.619	0.583	0.294	0.345	0.339	0.932	**0.977**
	Arg-	3,06	3	—	3,0.1	3	3,18	3
Semeion	AMI	0.013	−0.016	0.008	0.077	0.044	0.014	**0.606**
	ARI	0.028	0.001	0.035	0.236	−0.062	−0.063	**0.554**
	FMI	0.666	0.659	0.410	0.748	0.619	0.795	**0.901**
	Arg-	2,2.04	2	—	2,0.18	2	2,21	2
BCW	AMI	0.697	0.001	0.017	0.005	0.005	0.791	**0.809**
	ARI	0.803	0.003	0.006	0.016	0.017	0.858	**0.891**
	FMI	0.912	0.735	0.303	0.432	0.433	0.934	**0.950**
	Arg-	2,2 14	2	—	2,0.12	2	2,10	2
Contraceptive-MC	AMI	0.010	**0.245**	0.029	0.015	0.029	0.009	0.121
	ARI	0.003	**0.150**	0.006	−0.002	0.017	0.001	0.011
	FMI	0.434	0.478	0.298	0.585	0.364	0.441	**0.593**
	Arg-	2,1.3	3	—	3,0.2	3	3,21	3

续表 2

数据集	指标	DPC	GBDPC	LCCV	NDP-Kmeans	K-means	SNNDPC	HCHNG
Movement_libras	AMI	0.215	0.215	0.071	0.082	−0.002	0.299	**0.218**
	ARI	0.341	0.343	0.030	0.067	0.242	−0.011	**0.334**
	FMI	0.485	0.619	0.233	0.400	0.473	0.559	**0.897**
	Arg-	15,0.3	15	—	15,0.15	15	15,7	15
Dermatology	AMI	0.595	0.017	0.257	0.380	0.185	**0.871**	0.842
	ARI	0.359	0.004	0.047	0.088	0.047	0.733	**0.782**
	FMI	0.567	0.199	0.208	0.184	0.110	0.788	**0.830**
	Arg-	6,0.7	6	—	6,0.1	6	6,9	6
Mnist-r120	AMI	0.897	0.040	—	0.018	0.872	0.923	**0.930**
	ARI	0.863	−0.025	—	−0.001	0.896	0.875	**0.877**
	FMI	0.909	0.287	—	0.588	0.896	0.917	**0.918**
	Arg-	7,2.73	7	—	7,0.1	7	7,18	7
Page-blocks	AMI	0.031	0.002	0.034	0.010	0.049	0.127	**0.128**
	ARI	0.027	0.008	0.047	0.004	−0.011	0.053	**0.205**
	FMI	**0.902**	0.895	0.787	0.810	0.651	0.053	0.900
	Arg-	5,0.55	5	—	5,0	5	5,19	5

注: 加粗表示结果最优, “Arg-”是算法取得最优值时的参数。

HCHNG 算法在 Wine、Movement_libras、Semeion、BCW 和 Mnist-r120 数据集上优于其他算法, 其中 AMI、ARI 和 FMI 等指标排名第 1, 其中一些指标显著超过其他算法。对于 Dermatology 数据集, HCHNG 的 ARI 和 FMI 指标最高, 超过了其他 6 种算法。在 Contraceptive-MC 数据集中, HCHNG 表现出最佳的 FMI 性能。在 Page-blocks 数据集中, HCHNG 的 AMI 和 ARI 指标排名第 1, 但 FMI 稍逊于 DP 算法。实验结果表明, HCHNG 在大多数 UCI 数据集上优于其他 6 种算法, 特别是在处理密集区域中的聚类任务方面表现出色。

4　结束语

针对复杂结构数据集, 本文提出一种基于邻域图的层次聚类算法 (HCHNG)。1)HCHNG 算法利用混合的邻域图表达数据集的分布规律, 拓宽了基于邻域图聚类的思路。2) 该算法将数据集划分为子图, 然后根据定义的子图相似性进行合并, 提高了算法对于复杂结构数据集的聚类性能。实验结果表明, HCHNG 算法在识别各类复杂结构数据集中具有较高的有效性。

然而, 所提出的方法有一定的局限性。1) 它需要多次邻居搜索, 这在处理高维数据时并不理想。2) 层次聚类算法由于其生成最终聚类结果需要分层处理而具有更高的时间复杂度。3) 该算法要求用户设置真实的集群的数量。如何自适应地确定聚类结果也是未来研究的一个关键方向。同时还计划探索将所提出的邻域图方法与其他经典聚类算法相结合的可能性。

参考文献:

[1] CHENG Zhanhong, TRÉPANIER M, SUN Lijun. Probabilistic model for destination inference and travel pattern mining from smart card data[J]. Transportation, 2021, 48(4): 2035−2053.

[2] YOON B, JEONG Y, KIM S. Detecting a risk signal in stock investment through opinion mining and graph-based semi-supervised learning[J]. IEEE access, 2020, 8: 161943−161957.

[3] CAI Shousong, ZHANG Jing. Exploration of credit risk of P2P platform based on data mining technology[J]. Journal of computational and applied mathematics, 2020, 372: 112718.

[4] TAVALLALI P, TAVALLALI P, SINGHAL M. K-means tree: an optimal clustering tree for unsupervised learning[J]. The journal of supercomputing, 2021, 77(5): 5239−5266.

[5] ZHU Qidan, TANG Xiangmeng, ELAHI A. Application of the novel harmony search optimization algorithm for DBSCAN clustering[J]. Expert systems with applications, 2021, 178: 115054.

[6] CHEN Yewang, ZHOU Lida, BOUGUILA N, et al.

BLOCK-DBSCAN: Fast clustering for large scale data[J]. Pattern recognition, 2021, 109: 107624.

[7] CHU Zhenyue, WANG Weifeng, LI Bangzhun, et al. An operation health status monitoring algorithm of special transformers based on BIRCH and Gaussian cloud methods[J]. Energy reports, 2021, 7: 253−260.

[8] 邵佳, 金百锁. 基于层次聚类的图像分割算法[J]. 计算机应用, 2022, 42(S2): 211−216.
SHAO Jia, JIN Baisuo. Image segmentation algorithm based on hierarchical clustering[J]. Journal of computer applications, 2022, 42(S2): 211−216.

[9] ZHANG Tian, RAMAKRISHNAN R, LIVNY M. BIRCH [J]. ACM SIGMOD record, 1996, 25(2): 103−114.

[10] GUHA S, RASTOGI R, SHIM K. Cure: an efficient clustering algorithm for large databases[J]. Information systems, 2001, 26(1): 35−58.

[11] KARYPIS G, HAN E H, KUMAR V. Chameleon: hierarchical clustering using dynamic modeling[J]. Computer, 1999, 32(8): 68−75.

[12] KARYPIS G, AGGARWAL R, KUMAR V, et al. Multilevel hypergraph partitioning: application in vlsi domain[C]// Proceedings of the 34th Design Automation Conference. Anaheim: IEEE, 1997: 526−529.

[13] 吕端端. 基于最近邻思想的 Chameleon 聚类算法研究[D]. 西安: 西安理工大学, 2020: 31−49.
LYU Duanduan. Research on Chameleon clustering algorithm based on nearest neighbor idea[D]. Xi'an: Xi'an University of Technology, 2020: 31−49.

[14] CHENG Dongdong, ZHU Qingsheng, HUANG Jinlong, et al. A hierarchical clustering algorithm based on noise removal[J]. International journal of machine learning and cybernetics, 2019, 10(7): 1591−1602.

[15] ARTHUR D, VASSILVITSKII S. k-means++: The advantages of careful seeding[C]//Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 2007: 1027−1035.

[16] CHENG Dongdong, HUANG Jinlong, ZHANG Sulan, et al. K-means clustering with natural density peaks for discovering arbitrary-shaped clusters[J]. IEEE transactions on neural networks and learning systems, 2024, 35(8): 11077−11090.

[17] TZORTZIS G, LIKAS A. The MinMax k-means clustering algorithm[J]. Pattern recognition, 2014, 47(7): 2505−2516.

[18] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996: 226−231.

[19] BRYANT A, CIOS K. RNN-DBSCAN: a density-based clustering algorithm using reverse nearest neighbor density estimates[J]. IEEE transactions on knowledge and data engineering, 2018, 30(6): 1109−1121.

[20] GHOLIZADEH N, SAADATFAR H, HANAFI N. K-DBSCAN: an improved DBSCAN algorithm for big data[J]. The journal of supercomputing, 2021, 77(6): 6214−6235.

[21] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492−1496.

[22] TONG Wuning, LIU Sen, GAO Xiaozhi. A density-peak-based clustering algorithm of automatically determining the number of clusters[J]. Neurocomputing, 2021, 458: 655−666.

[23] CHEN Di, DU Tao, ZHOU Jin, et al. A domain density peak clustering algorithm based on natural neighbor[J]. Intelligent data analysis, 27(2): 443−462.

[24] 冯骥. 自然邻居思想概念及其在数据挖掘领域的应用[D]. 重庆: 重庆大学, 2016: 31−89.
FENG Ji. Concept of natural neighbor and its application in data mining[D]. Chongqing: Chongqing University, 2016: 31−89.

[25] 张清华, 周靖鹏, 代永杨, 等. 基于代表点与 K 近邻的密度峰值聚类算法[J]. 软件学报, 2023, 34(12): 5629−5648.
ZHANG Qinghua, ZHOU Jingpeng, DAI Yongyang, et al. Density peaks clustering algorithm based on representative points and K-nearest neighbors[J]. Journal of software, 2023, 34(12): 5629−5648.

[26] 位雅, 张正军, 何凯琳, 等. 基于相对密度的密度峰值聚类算法[J]. 计算机工程, 2023, 49(6): 53−61.
WEI Ya, ZHANG Zhengjun, HE Kailin, et al. Density peak clustering algorithm based on relative density[J]. Computer engineering, 2023, 49(6): 53−61.

[27] CHENG Dongdong, LI Ya, XIA Shuyin, et al. A fast granular-ball-based density peaks clustering algorithm for large-scale data[J]. IEEE transactions on neural networks and learning systems, 2024, 35(12): 17202−17215.

[28] LIU Rui, WANG Hong, YU Xiaomei. Shared-nearest-neighbor-based clustering by fast search and find of density peaks[J]. Information sciences, 2018, 450: 200−226.

[29] CHENG Dongdong, ZHU Qingsheng, HUANG Jinlong, et al. A novel cluster validity index based on local cores [J]. IEEE transactions on neural networks and learning systems, 2019, 30(4): 985−999.

[30] NGUYEN X, EPPS J, BAILEY J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance[J]. Properties, normalization and correction for chance, 2009, 10: 2837−2854.

[31] 冯骥, 冉瑞生, 魏延. 基于自然邻居邻域图的无参数离

群检测算法[J]. 智能系统学报, 2019, 14(5): 998−1006.
FENG Ji, RAN Ruisheng, WEI Yan. A parameter-free outlier detection algorithm based on natural neighborhood graph[J]. CAAI transactions on intelligent systems, 2019, 14(5): 998−1006.

[32] 周欢欢, 张征, 张琦. 结合共享近邻和共享逆近邻的密度峰聚类[J]. 西华师范大学学报 (自然科学版), 2022, 43(1): 108−115.
ZHOU Huanhuan, ZHANG Zheng, ZHANG Qi. Density peak clustering combining shared nearest neighbors and shared inverse neighbors[J]. Journal of China west normal university (natural sciences), 2022, 43(1): 108−115.

[33] JARVIS R A, PATRICK E A. Clustering using a similarity measure based on shared near neighbors[J]. IEEE transactions on computers, 1973, C-22(11): 1025−1034.

[34] ZHANG Jinghui, YANG Lijun, ZHANG Yong, et al. Non-parameter clustering algorithm based on saturated neighborhood graph[J]. Applied soft computing, 2022, 130: 109647.

[35] HUANG Jinlong, ZHU Qingsheng, YANG Lijun, et al. A novel outlier cluster detection algorithm without top-n parameter[J]. Knowledge-based systems, 2017, 121: 32−40.

[36] ZHANG Yuru, DING Shifei, WANG Yanru, et al. Chameleon algorithm based on improved natural neighbor graph generating sub-clusters[J]. Applied intelligence, 2021, 51(11): 8399−8415.

[37] 张辉. 密度峰值点快速搜索聚类算法的研究与改进[D]. 青岛: 山东科技大学, 2019: 32−37.
ZHANG Hui. Research and improvement of fast search clustering algorithm for density peak points[D]. Qingdao: Shandong University of Science and Technology, 2019: 32−37.

[38] 吕莉, 陈威, 肖人彬, 等. 面向密度分布不均数据的加权逆近邻密度峰值聚类算法[J]. 智能系统学报, 2024, 19(1): 165−175.
LYU Li, CHEN Wei, XIAO Renbin, et al. Density peak clustering algorithm based on weighted reverse nearest neighbor for uneven density datasets[J]. CAAI transactions on intelligent systems, 2024, 19(1): 165−175.

[39] 陈磊, 吴润秀, 李沛武, 等. 加权 K 近邻和多簇合并的密度峰值聚类算法[J]. 计算机科学与探索, 2022, 16(9): 2163−2176.
CHEN Lei, WU Runxiu, LI Peiwu, et al. Weighted K-nearest neighbors and multi-cluster merge density peaks clustering algorithm[J]. Journal of frontiers of computer science and technology, 2022, 16(9): 2163−2176.

[40] DING Shifei, DU Wei, XU Xiao, et al. An improved density peaks clustering algorithm based on natural neighbor with a merging strategy[J]. Information sciences, 2023, 624: 252−276.

[41] CHENG Bifang, BUNDROCK T, WILLIAMS D J. AAC oriental 200 oriental mustard[J]. Canadian journal of plant science, 2018, 98(4): 985−987.

[42] 王赢己. 基于自然邻居的密度峰值聚类算法研究[D]. 哈尔滨: 哈尔滨工程大学, 2021: 50−55.
WANG Yingji. Research on density peak clustering algorithm based on natural neighbors[D]. Harbin: Harbin Engineering University, 2021: 50−55.

[43] 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 Jia, MA Qing, XIAO Renbin, et al. Density peaks clustering based on shared nearest neighbor for manifold datasets[J]. CAAI transactions on intelligent systems, 2023, 18(4): 719−730.

[44] 谢文波. 基于互惠最近邻的层次聚类算法及其应用研究[D]. 成都: 电子科技大学, 2021: 36−57.
XIE Wenbo. Research on hierarchical clustering algorithm based on reciprocal nearest neighbors and its application[D]. Chengdu: University of Electronic Science and Technology of China, 2021: 36−57.

[45] 程东东. 基于局部核心点的聚类算法与度量研究[D]. 重庆: 重庆大学, 2018: 38−74.
CHENG Dongdong. Research on clustering algorithm and measurement based on local core points[D]. Chongqing: Chongqing University, 2018: 38−74.

作者简介:

陈仲尚, 硕士研究生, 主要研究方向为数据挖掘。E-mail: chenzhongshang@foxmail.com。

冯骥, 副教授, 博士, 计算机与信息科学院副院长, 主要研究方向为数据挖掘、人工智能。主持及参与国家自然科学基金、省部级项目等 10 余项。发表学术论文 10 余篇。E-mail: jifeng@cqnu.edu.cn。

杨德刚, 教授, 博士, 主要研究方向为智能算法、神经网络、复杂网络。主持及参与国家自然科学基金、省部级项目等 20 余项。发表学术论文 50 余篇。E-mail: yangdg@cqnu.edu.cn。