智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向密度分布不均数据的加权逆近邻密度峰值聚类算法

吕莉, 陈威, 肖人彬, 韩龙哲, 谭德坤

引用本文:
吕莉,陈威,肖人彬,韩龙哲,谭德坤. 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.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202212015

您可能感兴趣的其他文章

面向不均衡数据的融合谱聚类的自适应过采样法
Spectral clustering–fused adaptive synthetic oversampling approach for imbalanced data processing
智能系统学报. 2020, 15(4): 732–739 https://dx.doi.org/10.11992/tis.201909062

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

公理化模糊共享近邻自适应谱聚类算法
Shared nearest neighbor adaptive spectral clustering algorithm based on axiomatic fuzzy set theory
智能系统学报. 2019, 14(5): 897–904 https://dx.doi.org/10.11992/tis.201810002

结合谱聚类的标记分布学习
Label distribution learning based on spectral clustering
智能系统学报. 2019, 14(5): 966–973 https://dx.doi.org/10.11992/tis.201809019

结合稀疏表示与约束传递的半监督谱聚类算法
A semi–supervised spectral clustering algorithm combined with sparse representation and constraint propagation
智能系统学报. 2018, 13(5): 855–862 https://dx.doi.org/10.11992/tis.201703013

一种改进的自适应快速AF–DBSCAN聚类算法
An improved adaptive and fast AF–DBSCAN clustering algorithm
智能系统学报. 2016, 11(1): 93–98 https://dx.doi.org/10.11992.tis.201410021

DOI: 10.11992/tis.202212015
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20230802.1133.007

面向密度分布不均数据的加权逆近邻密度峰值聚类算法

吕莉[1,2], 陈威[1,2], 肖人彬[3], 韩龙哲[1,2], 谭德坤[1,2]

（1. 南昌工程学院 信息工程学院, 江西 南昌 330099; 2. 南昌工程学院 南昌市智慧城市物联感知与协同计算重点实验室, 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）

摘　要: 针对密度分布不均数据, 密度峰值聚类算法易忽略类簇间样本的疏密差异, 导致误选类簇中心; 分配策略易将稀疏区域的样本误分到密集区域, 导致聚类效果不佳的问题, 本文提出一种面向密度分布不均数据的加权逆近邻密度峰值聚类算法。该算法首先在局部密度公式中引入基于 sigmoid 函数的权重系数, 增加稀疏区域样本的权重, 结合逆近邻思想, 重新定义了样本的局部密度, 有效提升类簇中心的识别率; 其次, 引入改进的样本相似度策略, 利用样本间的逆近邻及共享逆近邻信息, 使得同一类簇样本间具有较高的相似度, 可有效改善稀疏区域样本分配错误的问题。在密度分布不均、复杂形态和 UCI 数据集上的对比实验表明, 本文算法的聚类效果优于 IDPC-FA、FNDPC、FKNN-DPC、DPC 和 DPCSA 算法。

关键词: 密度峰值聚类; 密度分布不均; 逆近邻; 共享逆近邻; 样本相似度; 局部密度; 分配策略; 数据挖掘
中图分类号: TP301　　**文献标志码**: A　　**文章编号**: 1673−4785(2024)01−0165−11

中文引用格式: 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.
英文引用格式: 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.

Density peak clustering algorithm based on weighted reverse nearest neighbor for uneven density datasets

LYU Li[1,2], CHEN Wei[1,2], XIAO Renbin[3], HAN Longzhe[1,2], TAN Dekun[1,2]

(1. School of Information Engineering, Nanchang Institute of Technology, Nanchang 330099, China; 2. Nanchang Key Laboratory of IoT Perception and Collaborative Computing for Smart City, Nanchang Institute of Technology, Nanchang 330099, China; 3. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: For data with uneven density distribution, the density peak clustering algorithm disregards the sparsity difference among intercluster samples, causing an inaccurate selection of the cluster center. Moreover, the allocation strategy easily divides the samples in sparse areas into dense areas by mistake, leading to a poor clustering effect. Therefore, the density peak clustering algorithm based on the weighted reverse nearest neighbor (DPC-WR) against datasets with uneven density distribution is proposed in this paper. First, the weight coefficient based on the sigmoid function is introduced to the local density formula to increase the weight of samples in sparse areas. Combined with the concept of reverse nearest neighbor, the local density of samples is then redesigned to improve the recognition rate of cluster centers effectively. Second, an improved sample similarity strategy is introduced, which utilizes reverse nearest neighbors and shares this neighbor's information between samples to increase the similarity of samples in the same cluster. This effectively solves the problem of sample allocation error in sparse areas. Experiments on uneven density distribution, complex morphology, and UCI datasets show that the clustering effect of the DPC-WR algorithm outperforms that of IDPC-FA, FNDPC, FKNN-DPC, DPC, and DPCSA algorithms.

Keywords: density peak clustering; uneven density distribution; reverse nearest neighbor; shared reverse nearest neighbor; sample similarity; local density; distribution strategy; data mining

聚类是数据分析中一种重要的无监督学习方法, 致力于揭示看似杂乱无章的未知数据背后隐藏的内在属性和规律, 为决策提供支持, 并已成

收稿日期: 2022−12−13.　**网络出版日期**: 2023−08−02.
基金项目: 国家自然科学基金项目 (62066030); 江西省重点研发计划项目 (20192BBE50076, 20203BBGL73225); 江西省教育厅科技项目 (GJJ190958).
通信作者: 吕莉. E-mail: lvli623@163.com.

©《智能系统学报》编辑部版权所有

功应用于许多领域，如图像分析[1]、模式识别[2]、社会网络挖掘[3]、市场统计分析[4]和医学研究[5]等。

传统的聚类算法分为基于划分的[6]、基于层次的[7]、基于网格的[8]、基于模型的[9]和基于密度的[10]聚类算法。K-means[11]是最著名的划分聚类算法，通过多次迭代获得最优聚类中心。K-means 收敛速度快，对大规模数据集的处理效率高，但该算法的性能依赖于初始聚类中心的选择，且对噪声点和异常值敏感。BIRCH(balanced iterative reducing and clustering using hierarchies)[12]是一种基于层次的聚类算法，利用聚类特征树自底向上进行聚类。BIRCH 聚类速度快，能识别噪声点，但不适用于高维和非凸数据。CLIQUE (clustering in quest)[13]是一种基于网格的聚类算法，把数据空间分为不同的网格，将样本对应到网格中，并进行密度的计算。CLIQUE 适用于高维和大规模数据集，但该算法聚类的准确度较低。EM(expectation maximization)[14]是一种基于模型的聚类算法，根据极大后验概率估计寻找样本的概率模型参数进行聚类。该算法计算结果稳定、准确，但对初始化数据敏感。DBSCAN(density-based spatial clustering of applications with noise)[15]是典型的基于密度的聚类算法，它将样本分为核心点和噪声点，根据密度可达将核心点聚合到同一个集群中。该算法可以识别任意形状的稠密数据集且对数据集中的异常点不敏感，但不能处理密度差异过大的数据。

2014 年，Science 发表了通过快速搜索和寻找密度峰值聚类[16](clustering by fast search and find of density peaks, DPC)算法。由于其新颖的设计理念和强大的性能，使得基于密度的聚类算法受到更广泛的关注和应用。DPC 算法基于两点假设：聚类中心周围的样本的局部密度相对较低；不同聚类中心间的距离相对较远。DPC 算法计算过程无需迭代，只需预先设定一个参数来识别聚类中心，但 DPC 算法仍有一些缺点：1）算法局部密度无法准确识别各类簇间样本的疏密差异，易造成类簇中心的误判；2）虽然 DPC 中的分配规则非常有效，但是当聚类过程出现某一个样本被错误分配，就会出现多米诺骨牌效应。

针对 DPC 算法易出现类簇中心选取错误的问题，吕莉等[17]提出二阶 K 近邻和多簇合并的密度峰值聚类算法 (density peaks clustering with second-order k-nearest neighbors and multi-cluster merging, DPC-SKMM)。DPC-SKMM 算法提出最小二阶 K 近邻的概念，根据 K 近邻和二阶 K 近邻信息重新定义局部密度，凸显聚中心与非聚类中心的密度差异。Sun 等[18]提出了基于最近邻优化分配策略的自适应密度峰值聚类算法 (nearest neighbors-based adaptive density peaks clustering with optimized allocation strategy, NADPC)。NADPC 算法提出了候选簇心和相对密度的概念，根据候选聚类中心的相对密度和高密度最近邻距离，计算聚类中心的可信度，从而选择聚类中心。赵嘉等[19]提出了 K 近邻和加权相似性的密度峰值聚类算法 (density peaks clustering algorithm with k-nearest neighbors and weighted similarity, DPC-KWS)。DPC-KWS 算法从样本的 K 近邻信息出发，重新定义了局部密度，调整了不同类簇中局部密度的大小。

针对分配规则出现的问题，吴润秀等[20]提出基于相对密度估计和多簇合并的密度峰值聚类算法 (density peaks clustering based on relative density estimating and multi cluster merging, DPC-RD-MCM)。DPC-RD-MCM 算法重新定义了微簇间相似性度量准则，通过多簇合并策略得到最终聚类结果，避免了分配错误连带效应。Ding 等[21]提出了基于中心和邻居的社区检测算法 (community detection by propagating the label of center, DCN)。DCN 算法根据样本的邻居传播标签，提出了标签传播的多重策略，有效解决了 DPC 分配策略的多米诺效应。赵嘉等[22]提出面向流形数据的测地距离与余弦互逆近邻密度峰值聚类算法 (density peaks clustering algorithm based on geodesic distance and cosine mutual reverse nearest neighbors for manifold datasets, DPC-GDCN)。DPC-GDCN 算法将互逆近邻和余弦相似性相结合，得到基于余弦互逆近邻的样本相似度矩阵，为流形类簇准确分配样本。

上述算法均有效提高了 DPC 算法的聚类效果，但忽略了样本间的分布特征，无法对密度分布不均等特定数据集取得较好的聚类效果。因此，本文提出了面向密度分布不均数据的加权逆近邻密度峰值聚类算法 (density peak clustering algorithm based on weighted reverse nearest neighbor for uneven density datasets, DPC-WR)。DPC-WR 算法充分利用了逆近邻和共享逆近邻信息，算法的主要创新点如下：1）结合 sigmoid 函数及逆近邻思想重新定义了局部密度，平衡了样本间疏密程度的差异，提高了类簇中心的识别率；2）在样本分配策略中，引入逆近邻及共享逆近邻信息，避

免了稀疏区域样本的错误分配, 提高了聚类效果。

1 DPC 算法

DPC 是一种高效的密度峰值聚类算法, 可以快速找到聚类中心, 对多种聚类任务具有良好的适应性。该算法基于聚类中心密度大于邻域密度, 聚类中心间的距离相对较远的思想, 提出了两种描述样本$\boldsymbol{x}_i$的密度和距离的方法, 即局部密度ρ_i和相对距离δ_i。

设有数据集$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$。对数据集$\boldsymbol{X}$中的每个样本$\{\boldsymbol{x}_i\}_{i=1}^n$, 样本间的欧氏距离为

$$d_{ij}=\|\boldsymbol{x}_i-\boldsymbol{x}_j\| \tag{1}$$

局部密度ρ_i有两种定义方式:

$$\rho_i=\sum_{i\neq j}\chi(d_{ij}-d_{\mathrm{c}}) \tag{2}$$

其中: $\chi(x)=\begin{cases}1, & x<0\\ 0, & x\geqslant 0\end{cases}$, d_{c}是截止距离。

$$\rho_i=\sum_{i\neq j}\exp\left(-\frac{d_{ij}^2}{d_{\mathrm{c}}^2}\right) \tag{3}$$

式 (2) 为截断核, 式 (3) 为高斯核。相对距离δ_i的定义如下: 对于每个样本$\boldsymbol{x}_i$, 找到所有比样本$\boldsymbol{x}_i$密集的样本$\boldsymbol{x}_j$, 选择最小的d_{ij}; 如果情况相反, 则选择最大的d_{ij}。δ_i的计算公式如下:

$$\delta_i=\begin{cases}\min\limits_j(d_{ij}), & \exists j\ \mathrm{s.t.}\ \rho_i>\rho_j\\ \max\limits_j(d_{ij}), & \text{其他}\end{cases} \tag{4}$$

类簇中心由决策图确定, 以局部密度ρ_i为横坐标, 相对距离δ_i为纵坐标, 建立决策图。理想情况下, 聚类中心选取为密度较高且相距较远的样本。定义如下:

$$\gamma_i=\rho_i\cdot\delta_i \tag{5}$$

最后, 选取前n个较大的值作为聚类中心, n为最终类簇数。

2 DPC-WR 算法

在聚类算法中, K 近邻和逆近邻在表征密度时起着重要作用。K 近邻能准确反映样本在空间中的局部分布特征。而逆近邻基于全局视角检查它的邻域, 数据分布的变化会对样本的逆近邻造成影响, 使得算法更容易识别聚类中心和提升算法聚类性能。因此, 本文引入逆近邻和共享逆近邻信息, 重新定义了局部密度, 设计了样本相似度策略, 充分考虑了样本的总体分布, 使样本的局部一致性和全局一致性得到较好的均衡。

2.1 加权逆近邻的局部密度

定义 1　逆近邻[23]。设样本$\boldsymbol{x}_i,\boldsymbol{x}_j\in\boldsymbol{X}$, $\boldsymbol{x}_i$在$\boldsymbol{x}_j$的 K 近邻集中, 那么$\boldsymbol{x}_j$是$\boldsymbol{x}_i$的逆近邻, 具体定义如下:

$$\mathrm{RNN}(\boldsymbol{x}_i)=\left\{\boldsymbol{x}_j\in\boldsymbol{X}\middle|\boldsymbol{x}_i\in\mathrm{KNN}(\boldsymbol{x}_j)\right\} \tag{6}$$

定义 2　隶属度。样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$的隶属度μ_{ij}定义如下:

$$\mu_{ij}=\exp\left[-\frac{d_{ij}^2}{k\cdot[1+|R(i)|]}\right] \tag{7}$$

其中: k为样本的近邻数; $|R(i)|$表示样本$\boldsymbol{x}_i$的逆近邻数, 该值越大, 该点的隶属度越大。

定义 3　加权逆近邻的局部密度。局部密度定义如下:

$$\rho_i=\sum_{j\in\mathrm{RNN}(i)}\lambda_{ij}\cdot\mu_{ij} \tag{8}$$

权重系数:

$$\lambda_{ij}=\frac{\mathrm{RNN}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{1+\exp(-|R(i)|)} \tag{9}$$

$$\mathrm{RNN}(x_i,x_j)=\begin{cases}1, & x_i\in\mathrm{RNN}(x_j)\ \text{且}\ x_i\neq x_j\\ 0, & \text{其他}\end{cases}$$

其中: $\dfrac{1}{1+\exp(-x)}$为 sigmoid 函数, x 为实数。

类簇密度不同时, 数据稠密区域与数据稀疏区域的样本对聚类中心选取的贡献程度是不同的。因此, 处理密度分布不均数据时, 通过引入权重对样本的贡献进行处理, 可以达到良好的均衡效果。本文以样本的逆近邻数作为衡量密度的重要指标, 引入 sigmoid 函数, 对不同类簇中的样本进行权重调整。

式 (9) 中λ_{ij}为权重系数, 它在 sigmoid 函数的基础上进行重构, 分母部分以样本的逆近邻数替代了原函数的变量x值, 分子部分采用逆近邻代替实数值, 使密度分布不均数据在不同区域具有辨识度。从函数可知, 随着逆近邻数逐渐增加, 其函数值趋近于 1, 说明位于高密度区域的样本所加的权重近似于 1。对于较高密度的样本, 被选为聚类中心的概率较大, 此时逆近邻数起到关键的作用。当逆近邻数不断减少直至为 0 时, 样本的权重将会从 1 发生非线性变化减少到 0.5, 这不仅考虑到各样本间细微的影响, 还提高了聚类中心与非聚类中心的区分, 使式 (7) 的隶属度定义更为合理。

2.2 逆近邻和共享逆近邻的分配策略

定义 4　共享逆近邻。设样本$\boldsymbol{x}_i$的逆近邻集为$\mathrm{RNN}(\boldsymbol{x}_i)$, $\boldsymbol{x}_j$的逆近邻集为$\mathrm{RNN}(\boldsymbol{x}_j)$, 样本$\boldsymbol{x}_i$与$\boldsymbol{x}_j$的共享逆近邻定义如下所示:

$$\psi(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left\{\boldsymbol{x}_i\in\boldsymbol{X},\boldsymbol{x}_j\in\boldsymbol{X}\middle|\mathrm{RNN}(\boldsymbol{x}_i)\cap\mathrm{RNN}(\boldsymbol{x}_j)\right\} \tag{10}$$

定义 5　逆近邻和共享逆近邻的样本邻近度。通过样本间的逆近邻信息, 定义了邻近度ω_{ij}, 其定义如下:

$$\omega_{ij}=\begin{cases}\mathrm{e}^{d_{ij}}, & \boldsymbol{x}_j\in \mathrm{RNN}(\boldsymbol{x}_i)\\ \dfrac{\mathrm{e}^{d_{ij}}}{\max(d)}, & \boldsymbol{x}_j\notin \mathrm{RNN}(\boldsymbol{x}_i)\end{cases}\tag{11}$$

其中$\max(d)$表示数据集$\boldsymbol{X}$中样本间欧氏距离的最大值。

式 (11) 中第一行表示当样本$\boldsymbol{x}_j$位于样本$\boldsymbol{x}_i$的逆近邻范围内时所赋予的邻近度；第二行表示当样本$\boldsymbol{x}_j$不处于样本$\boldsymbol{x}_i$的逆近邻范围时，由于样本间的紧密程度低，若将值赋 0，易忽略未在范围内的样本的细微影响，故其邻近度在逆近邻范围的基础上除以最大距离所得。

定义 6　样本相似度。基于逆近邻和共享逆近邻，得到样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$的相似度：

$$S(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left[\left|\psi(\boldsymbol{x}_i,\boldsymbol{x}_j)\right|+\mathrm{RNN}(\boldsymbol{x}_i,\boldsymbol{x}_j)\right]\cdot\beta(\boldsymbol{x}_i,\boldsymbol{x}_j)\tag{12}$$

式中：$\left|\psi(\boldsymbol{x}_i,\boldsymbol{x}_j)\right|$表示$\psi(\boldsymbol{x}_i,\boldsymbol{x}_j)$集合中样本的个数，$\beta(\boldsymbol{x}_i,\boldsymbol{x}_j)$的定义如下：

$$\beta(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{1}{k}\left[\sum_{i,j=1}^{n}\omega(\boldsymbol{x}_i,\boldsymbol{x}_j)\right]\tag{13}$$

$\beta(\boldsymbol{x}_i,\boldsymbol{x}_j)$反映了样本所处空间的紧密程度，分子部分为每个样本的相似度之和，分母部分为归一化参数。式 (12) 考虑了样本本身及其共享逆近邻样本在定义样本间相似度方面起着重要的作用，因此，只有当样本之间存在逆近邻或共享逆近邻时，才存在相似性。

2.3　算法步骤

输入　数据集$\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^{n}$，近邻数$k$

输出　聚类结果C

1）数据归一化；

2）计算数据集样本间的欧氏距离；

3）根据式 (8) 和式 (4) 分别计算各样本的局部密度ρ_i和相对距离δ_i；

4）根据式 (5) 计算各样本的决策值γ_i并选取聚类中心；

5）根据式 (12) 计算基于逆近邻和共享逆近邻的样本相似度并构建相似度矩阵；

6）对于所有已分配的样本，找到相似度最高的未分配样本并将其分配到已分配样本所在的簇中；

7）若所有已分配样本与未分配样本间的相似度为 0，转至步骤 8），否则转至步骤 6）；

8）若还存在未分配的样本，则按 DPC 算法分配策略分配；

9）输出聚类结果。

2.4　算法复杂度分析

设样本规模为n，k为近邻数。DPC 算法的时间复杂度为$O(n^2)$[24]。DPC-WR 算法的时间复杂度主要由以下 6 个部分组成：1）计算样本间距离矩阵的复杂度$O(n^2)$；2）计算样本的局部密度，包括计算样本间的 K 近邻和样本间的逆近邻与逆近邻数，前者复杂度为$O(n)$，后者为$O(kn)$和$O(n^2)$；3）计算样本相对距离的复杂度$O(n^2)$；4）计算样本决策值的复杂度$O(n^2)$；5）计算样本的共享逆近邻与邻近度的复杂度$O(n^2)$；6）计算样本最坏分配情况的复杂度$O(n^2\log n)$。综上，DPC-WR 算法的时间复杂度为$O(n^2\log n)$。

3　实验结果与分析

3.1　实验设置

为验证 DPC-WR 算法的性能，本文在密度分布不均数据集、复杂形态数据集和 UCI 真实数据集上进行实验。将 DPC-WR 算法与 IDPC-FA[25]、FNDPC[26]、FKK-DPC[20]、DPC[16] 和 DPCSA[27] 算法进行比较。其中，IDPC-FA、DPCSA 和 DPC 算法由原作者提供，FNDPC 和 FKNN-DPC 算法参照原文献编程实现。除了 DPCSA 和 IDPC-FA 无需对参数调优外，其余算法均需要调整参数。DPC-WR 和 FKNN-DPC 算法参数k值的选取是 1~50 之间的最优值；DPC 算法的截断距离d_c的选取在 0.1%~5%，步长为 0.1%；FNDPC 算法参数ε的选取在 0.01~1，步长为 0.01。实验环境为Win10 64 bit操作系统，AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz 处理器，16.0GB 内存。

本文采用调整互信息（adjusted mutual information, AMI）[28]、Fowlkes-Mallows 指数（fowlkes-mallows index, FMI）[28] 和调整兰德系数（adjusted rand index, ARI）[29] 对聚类效果进行评价，其中，3 个指标的最佳结果都为 1，各指标值接近 1 的程度越高，表明聚类结果越好。

3.2　密度分布不均数据集的实验结果与分析

本文选取了 6 个不同规模的密度分布不均数据集进行实验，其基本特征如表 1 所示。

表 1　密度分布不均数据集的基本特征

Table 1　Basic characteristics of datasets with uneven density distribution

数据集	样本规模	维度	类簇数
Jain	373	2	2
Twomoons	1 502	2	2
Cmc	1 002	2	3
Ring	1 200	2	2
LineBlobs	266	2	3
Ls	1 741	2	6

表 2 给出了 6 种算法在密度分布不均数据集上的聚类结果，其中最优结果以粗体表示，“Arg-”

表示各算法的最优参数取值。“—”表示不含参数。DPC-WR 算法在 6 个数据集上均获得最佳的聚类效果。IDPC-FA 算法对 Jain 和 LineBlobs 具有较好的聚类效果, 对其他数据集的聚类效果较差。FKNN-DPC 算法对 Cmc 和 LineBlobs 数据集聚类效果较好, 对其他数据集聚类效果不佳。DPCSA 算法仅对 LineBlobs 数据集具有较好的聚类效果。FNDPC 和 DPC 算法在 6 个数据集上的聚类性能均低于 DPC-WR 和 FKNN-DPC 算法。

表 2　6 种算法在密度分布不均数据集上的聚类结果

Table 2　Clustering results of six algorithms on datasets with uneven density distribution

算法	Jain			
	AMI	ARI	FMI	Arg-
DPC-WR	**1.0000**	**1.0000**	**1.0000**	24
IDPC-FA	**1.0000**	**1.0000**	**1.0000**	—
FNDPC	0.5961	0.7257	0.9051	0.47
FKNN-DPC	0.7092	0.8224	0.9359	43
DPC	0.6183	0.7146	0.8819	0.8
DPCSA	0.2167	0.0442	0.5924	—

算法	Twomoons			
	AMI	ARI	FMI	Arg-
DPC-WR	**1.0000**	**1.0000**	**1.0000**	39
IDPC-FA	0.5171	0.6106	0.8458	—
FNDPC	**1.0000**	**1.0000**	**1.0000**	0.12
FKNN-DPC	**1.0000**	**1.0000**	**1.0000**	77
DPC	0.6671	0.7621	0.9005	4.7
DPCSA	0.3647	0.2746	0.6607	—

算法	Cmc			
	AMI	ARI	FMI	Arg-
DPC-WR	**1.0000**	**1.0000**	**1.0000**	20
IDPC-FA	0.8093	0.8421	0.9027	—
FNDPC	0.8093	0.8421	0.9027	0.28
FKNN-DPC	**1.0000**	**1.0000**	**1.0000**	49
DPC	0.3857	0.2661	0.5377	5
DPCSA	0.6656	0.5761	0.7454	—

算法	Ring			
	AMI	ARI	FMI	Arg-
DPC-WR	**1.0000**	**1.0000**	**1.0000**	4
IDPC-FA	0.1333	0.0886	0.6362	—
FNDPC	0.0276	0.0104	0.6566	0.01
FKNN-DPC	0.5702	0.5900	0.8005	24
DPC	0.2073	0.1815	0.6431	0.06
DPCSA	0.6362	0.6721	0.8387	—

续表 2

算法	LineBlobs			
	AMI	ARI	FMI	Arg-
DPC-WR	**1.0000**	**1.0000**	**1.0000**	4
IDPC-FA	**1.0000**	**1.0000**	**1.0000**	—
FNDPC	0.7794	0.7179	0.8148	0.11
FKNN-DPC	**1.0000**	**1.0000**	**1.0000**	7
DPC	0.8375	0.8237	0.8842	4.2
DPCSA	**1.0000**	**1.0000**	**1.0000**	—

算法	Ls			
	AMI	ARI	FMI	Arg-
DPC-WR	**1.0000**	**1.0000**	**1.0000**	40
IDPC-FA	0.7076	0.6274	0.7325	—
FNDPC	0.7564	0.6898	0.7808	0.37
FKNN-DPC	0.8719	0.8179	0.8735	48
DPC	0.7665	0.6894	0.7779	0.91
DPCSA	0.7252	0.5999	0.7129	—

Friedman 检验[30] 是利用秩实现对多个总体分布是否存在显著差异的非参数检验方法。将对比算法进行检验可以更准确地反映算法间评价指标的差异, 秩均值越高则算法的聚类效果越优。从表 3 可以发现, 在密度分布不均数据集上聚类评价指标 AMI、ARI 和 FMI 的秩均值排名中, DPC-WR 算法都位列第 1, 且秩均值都大于 5.4。

表 3　6 种算法在密度分布不均数据集上的秩均值

Table 3　Rank mean of the six algorithms on the unevenly distributed density datasets

AMI		ARI		FMI	
算法	秩均值	算法	秩均值	算法	秩均值
DPC-WR	**5.42**	DPC-WR	**5.42**	DPC-WR	**5.42**
IDPC-FA	3.08	IDPC-FA	3.25	IDPC-FA	3.08
FNDPC	2.58	FNDPC	2.92	FNDPC	3.25
FKNN-DPC	4.67	FKNN-DPC	4.67	FKNN-DPC	4.67
DPC	2.67	DPC	2.33	DPC	2.17
DPCSA	2.58	DPCSA	2.42	DPCSA	2.42

由于篇幅所限, 本文选取了 1 个典型的密度分布不均数据集。图 1 给出了 DPC-WR、IDPC-FA、FNDPC、FKNN-DPC、DPC 和 DPCSA 算法在 Jain 数据集上的聚类结果。图中不同的颜色代表不同的类簇, 类簇中心用“六角星”表示。

Jain 数据集由 2 个稠密程度不同的新月形类簇构成。从图 1 可知, DPC-WR 和 IDPC-FA 算法充分考虑了样本间的密度差, 能准确地找到类簇中心; FNDPC 和 FKNN-DPC 算法虽然找到了正确的类簇中心, 但样本分配策略存在错误, 导致稀疏类簇样本的错误分配; DPC 和 DPCSA 算法没有找到正确的聚类中心, 导致聚类效果不佳。

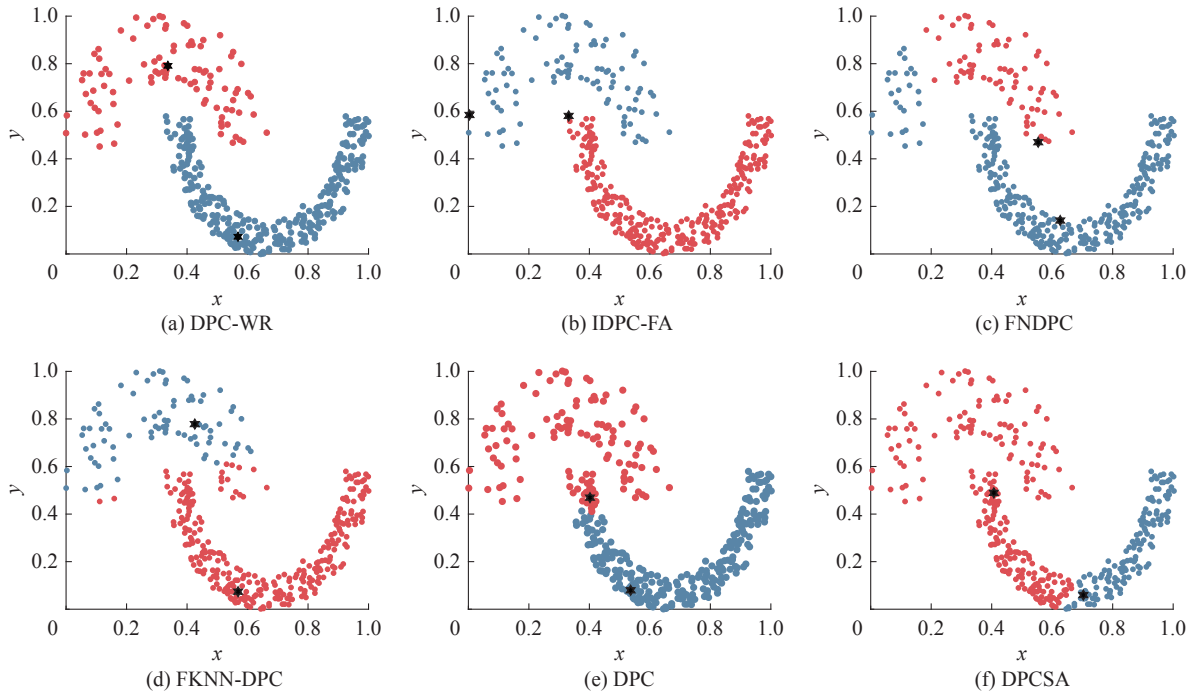

图 1　6 种算法在 Jain 数据集上的聚类结果

Fig. 1　The clustering results of 6 algorithms on Jain dataset

3.3　复杂形态数据集的实验结果与分析

复杂形态数据集是指具有多尺度、簇类形状多样等结构的数据集。本文选取了 6 个复杂形态的数据集，其基本特征如表 4 所示。表 5 给出了 6 种算法在复杂形态数据集上的聚类结果。从表 5 可知，DPC-WR 和 IDPC-FA 算法比其他对比算法的聚类结果更优，都存在 4 个聚类效果较好的数据集。从整体来看，DPC-WR 算法的聚类效果最佳，具体表现在 Flame、R15、Sticks 和 Pathbased 数据集。

表 4　复杂形态数据集的基本特征

Table 4　Basic characteristics of complex

数据集	样本规模	维度	类簇数
Flame	240	2	2
R15	600	2	15
Aggregation	788	2	7
D31	3 100	2	31
Sticks	512	2	4
Pathbased	300	2	3

表 6 为 6 种算法在 6 个复杂形态数据集上评价指标的秩均值。从表 6 可以发现，DPC-WR 算法在 AMI、ARI 和 FMI 评价指标的秩均值中位列第一，其次是 IDPC-FA 算法，然后是 FNDPC 算法。

表 5　6 种算法在复杂形态数据集上的聚类结果

Table 5　Clustering results of six algorithms on complex morphological datasets

算法	Flame			
	AMI	ARI	FMI	Arg-
DPC-WR	**1.0000**	**1.0000**	**1.0000**	1
IDPC-FA	**1.0000**	**1.0000**	**1.0000**	—
FNDPC	**1.0000**	**1.0000**	**1.0000**	0.13
FKNN-DPC	0.9267	0.9667	0.9845	5
DPC	**1.0000**	**1.0000**	**1.0000**	2.8
DPCSA	**1.0000**	**1.0000**	**1.0000**	—

算法	R15			
	AMI	ARI	FMI	Arg-
DPC-WR	**0.9938**	**0.9928**	**0.9933**	32
IDPC-FA	**0.9938**	**0.9928**	**0.9933**	—
FNDPC	**0.9938**	**0.9928**	**0.9933**	0.03
FKNN-DPC	**0.9938**	**0.9928**	**0.9933**	27
DPC	**0.9938**	**0.9928**	**0.9933**	0.6
DPCSA	0.9885	0.9857	0.9866	—

算法	Aggregation			
	AMI	ARI	FMI	Arg-
DPC-WR	0.9922	0.9956	0.9966	12
IDPC-FA	**1.0000**	**1.0000**	**1.0000**	—
FNDPC	0.9864	0.9913	0.9932	0.02
FKNN-DPC	0.9905	0.9949	0.9960	20
DPC	0.9922	0.9956	0.9966	4
DPCSA	0.9537	0.9581	0.9673	—

续表 5

算法	D31			
	AMI	ARI	FMI	Arg-
DPC-WR	0.961 7	0.946 5	0.948 2	50
IDPC-FA	0.957 5	0.940 2	0.942 1	—
FNDPC	0.955 5	0.936 4	0.938 5	0.04
FKNN-DPC	**0.965 8**	**0.952 2**	**0.953 7**	23
DPC	0.955 4	0.936 5	0.938 5	0.7
DPCSA	0.955 2	0.935 3	0.937 4	—

算法	Sticks			
	AMI	ARI	FMI	Arg-
DPC-WR	**1.000 0**	**1.000 0**	**1.000 0**	3
IDPC-FA	**1.000 0**	**1.000 0**	**1.000 0**	—
FNDPC	**1.000 0**	**1.000 0**	**1.000 0**	0.22
FKNN-DPC	**1.000 0**	**1.000 0**	**1.000 0**	7
DPC	0.809 4	0.753 4	0.823 5	2
DPCSA	0.763 4	0.636 0	0.744 3	—

算法	Pathbased			
	AMI	ARI	FMI	Arg-
DPC-WR	**0.940 1**	**0.959 0**	**0.972 7**	5
IDPC-FA	0.844 2	0.859 3	0.906 7	—
FNDPC	0.575 1	0.506 7	0.706 5	0.01
FKNN-DPC	0.930 5	0.949 9	0.966 5	9
DPC	0.521 2	0.471 7	0.666 4	3.8
DPCSA	0.707 3	0.613 3	0.751 1	—

表 6　6 种算法在复杂形态数据集上的秩均值

Table 6　Rank mean of 6 algorithms on complex morphological datasets

AMI		ARI		FMI	
算法	秩均值	算法	秩均值	算法	秩均值
DPC-WR	**4.67**	DPC-WR	**4.67**	DPC-WR	**4.67**
IDPC-FA	4.42	IDPC-FA	4.42	IDPC-FA	4.42
FNDPC	3.25	FNDPC	3.08	FNDPC	3.17
FKNN-DPC	3.92	FKNN-DPC	3.92	FKNN-DPC	3.92
DPC	2.92	DPC	3.08	DPC	3.00
DPCSA	1.83	DPCSA	1.83	DPCSA	1.83

3.4　UCI 数据集的实验结果与分析

UCI 数据集又称真实数据集, 它是一个常用的标准测试数据集。为了进一步验证 DPC-WR 算法的有效性, 本文选取了 8 个真实数据集, 对 6 种算法进行实验。其中测试的数据集包括 Iris、Wine、Seeds、Ecoli、Inonsphere、Libras、Dermatology 和 Wdbc。表 7 给出了各数据集的基本特征。表 8 为 6 种算法在 UCI 数据集上的聚类效果。从表 8 可以发现, 处理 Seeds 数据集时, DPC-WR 算法的聚类效果不及 IDPC-FA、FKNN-DPC 和 DPC 算法。处理 Inonsphere 数据集时, DPC-WR 算法的聚类效果低于 FKNN-DPC 算法。处理 Dermatology 数据集时, DPC-WR 算法的聚类效果比 DPC-SA 算法好, 但略逊于其他算法。剩余的 Iris、Wine、Ecoli、Libras 和 Wdbc 数据集, DPC-WR 算法的聚类效果都优于其他算法。

表 7　UCI 数据集的基本特征

Table 7　Basic characteristics of UCI datasets

数据集	样本规模	维度	类簇数
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Ecoli	336	8	8
Inonsphere	351	34	2
Libras	360	90	15
Dermatology	366	33	6
Wdbc	569	30	2

表 8　6 种算法在 UCI 数据集上的聚类结果

Table 8　Clustering results of six algorithms on UCI datasets

算法	Iris			
	AMI	ARI	FMI	Arg-
DPC-WR	**0.897 1**	**0.909 3**	**0.935 6**	9
IDPC-FA	0.862 3	0.885 7	0.923 3	—
FNDPC	0.883 1	0.903 8	0.935 5	0.11
FKNN-DPC	0.883 1	0.903 8	0.935 5	22
DPC	0.860 6	0.885 7	0.923 3	0.2
DPCSA	0.883 1	0.903 8	0.935 5	—

算法	Wine			
	AMI	ARI	FMI	Arg-
DPC-WR	**0.871 6**	**0.897 5**	**0.931 9**	44
IDPC-FA	0.767 5	0.771 3	0.847 8	—
FNDPC	0.789 8	0.802 5	0.868 6	0.26
FKNN-DPC	0.848 1	0.883 9	0.922 9	8
DPC	0.706 5	0.672 4	0.783 5	2
DPCSA	0.748 0	0.741 4	0.828 3	—

算法	Seeds			
	AMI	ARI	FMI	Arg-
DPC-WR	0.716 1	0.763 5	0.841 6	7
IDPC-FA	0.729 9	0.767 0	0.844 4	—
FNDPC	0.713 6	0.754 5	0.836 1	0.07
FKNN-DPC	**0.775 7**	**0.802 4**	**0.868 2**	9
DPC	0.729 9	0.767 0	0.844 4	0.7
DPCSA	0.660 9	0.687 3	0.791 8	—

算法	Ecoli			
	AMI	ARI	FMI	Arg-
DPC-WR	**0.669 7**	**0.764 9**	**0.835 8**	1
IDPC-FA	0.663 8	0.756 1	0.828 4	—
FNDPC	0.483 3	0.561 8	0.717 8	0.35
FKNN-DPC	0.587 8	0.589 4	0.702 7	2
DPC	0.497 8	0.446 5	0.577 5	0.4
DPCSA	0.440 6	0.459 3	0.646 7	—

续表 8

算法	Inonsphere			
	AMI	ARI	FMI	Arg-
DPC-WR	0.2289	0.3122	0.6722	10
IDPC-FA	0.1355	0.2183	0.6432	—
FNDPC	0.1630	0.2483	0.6531	0.06
FKNN-DPC	**0.3485**	**0.4790**	**0.7716**	8
DPC	0.1504	0.2357	0.6491	0.5
DPCSA	0.1335	0.2135	0.6390	—

算法	Libras			
	AMI	ARI	FMI	Arg-
DPC-WR	**0.5757**	**0.3915**	**0.4453**	13
IDPC-FA	0.5733	0.3816	0.4247	—
FNDPC	0.5494	0.3290	0.3869	0.17
FKNN-DPC	0.5554	0.3459	0.4044	10
DPC	0.5358	0.3193	0.3717	0.3
DPCSA	0.5388	0.3095	0.3791	—

算法	Dermatology			
	AMI	ARI	FMI	Arg-
DPC-WR	**0.8654**	0.7398	0.7929	6
IDPC-FA	0.8308	**0.8464**	**0.8764**	—
FNDPC	0.7933	0.8029	0.8441	0.17
FKNN-DPC	0.8083	0.8361	0.8706	35
DPC	0.8354	0.8389	0.8715	1.5
DPCSA	0.7470	0.6099	0.6922	—

算法	Wdbc			
	AMI	ARI	FMI	Arg-
DPC-WR	**0.7208**	**0.8243**	**0.9190**	14
IDPC-FA	0.6237	0.7423	0.8829	—
FNDPC	0.6076	0.3645	0.8758	0.05
FKNN-DPC	0.6423	0.7613	0.8894	2
DPC	0.4366	0.4964	0.7941	0.5
DPCSA	0.3361	0.3771	0.7595	—

表 9 为 6 种算法在 UCI 数据集上评价指标的秩均值。从表 9 可知，DPC-WR 算法相较于其他对比算法评价指标的秩均值都是最高的，其次是 FKNN-DPC 和 IDPC-FA 算法。由此可以得出，在 UCI 真实数据集上，DPC-WR 算法的聚类效果明显优于 IDPC-FA、FNDPC、FKNN-DPC、DPC 和 DPCSA 算法。

表 9　6 种算法在 UCI 数据集上的秩均值

Table 9　Rank mean of six algorithms on UCI datasets

AMI		ARI		FMI	
算法	秩均值	算法	秩均值	算法	秩均值
DPC-WR	**5.38**	DPC-WR	**5.00**	DPC-WR	**5.00**
IDPC-FA	3.81	IDPC-FA	3.88	IDPC-FA	3.88
FNDPC	3.00	FNDPC	3.00	FNDPC	3.38
FKNN-DPC	4.63	FKNN-DPC	4.75	FKNN-DPC	4.63
DPC	2.56	DPC	2.63	DPC	2.38
DPCSA	1.63	DPCSA	1.75	DPCSA	1.75

3.5　算法的仿真时间与分析

本文中，仿真时间指完成单个数据集聚类所花费的时长，是评价算法聚类性能的重要指标。为计算聚类算法的聚类时间，本文针对前述 3 类数据集进行了相应的实验，结果详见表 10。

从表 10 可以发现，对于样本规模较小的数据集，如 Jain、LineBlobs 等，DPC-WR 算法可以与对比算法相媲美或者更优；对于样本规模较大的数据集，如 Twomoons、Cmc 等，DPC-WR 算法弱于 FNDPC、FKNN-DPC、DPC 和 DPCSA 算法，但是，相较于 IDPC-FA 算法，DPC-WR 算法加速效应明显。这主要是由于 DPC-WR 算法的时间复杂度为 $O(n^2 \log n)$，样本规模的增大，$\log n$ 的影响将变得更加明显。

表 10　6 种算法在 3 类数据集上的仿真时间

Table 10　Simulation time of six algorithms on three types of datasets　　s

数据集	DPC-WR	IDPC-FA	FNDPC	FKNN-DPC	DPC	DPCSA
Jain	0.07	11.23	0.1	0.1	**0.06**	**0.06**
Twomoons	1.57	341.54	0.23	0.31	**0.23**	**0.23**
Cmc	0.38	87.32	0.18	0.21	0.14	**0.13**
Ring	0.3	134.56	0.18	0.16	**0.13**	0.15
LineBlobs	**0.03**	6.47	0.1	0.06	0.06	0.06
Ls	2.42	672.19	0.32	0.35	**0.28**	0.3
Flame	**0.01**	5.53	0.09	0.05	0.05	0.06
R15	0.18	45.65	0.26	**0.13**	0.18	0.17
Aggregation	0.21	48.78	0.19	0.19	**0.13**	**0.13**
D31	15.43	7093.37	1.59	1.53	**1.07**	1.15
Sticks	**0.07**	18.42	0.14	0.09	0.08	0.1
Pathbased	**0.03**	7.76	0.12	0.07	0.06	0.06

续表 10

数据集	DPC-WR	IDPC-FA	FNDPC	FKNN-DPC	DPC	DPCSA
Iris	**0.02**	3.83	0.12	0.07	0.06	0.06
Wine	**0.04**	4.42	0.11	0.06	0.07	0.06
Seeds	**0.03**	5.23	0.11	0.07	0.07	0.06
Ecoli	**0.02**	14.23	0.17	0.08	0.1	0.11
Inonsphere	**0.06**	13.37	0.12	**0.06**	0.07	**0.06**
Libras	**0.09**	43	0.27	0.11	0.18	0.15
Dermatology	**0.06**	17.07	0.16	0.1	0.09	0.09
Wdbc	0.17	28.36	0.14	**0.07**	0.08	0.08

4　结束语

针对 DPC 算法在面对密度分布不均数据集时易出现误选类簇中心以及样本分配错误的问题, 本文提出了一种面向密度分布不均数据的加权逆近邻密度峰值聚类算法。DPC-WR 算法首先引入基于 sigmoid 函数的权重系数及逆近邻思想, 重新定义样本的局部密度; 随后, 利用逆近邻和共享逆近邻计算样本相似度; 最后, 按照样本相似度将剩余样本进行分配。实验结果表明, DPC-WR 算法有效改善了类簇间样本疏密差异导致误判聚类中心以及稀疏区域样本错误分配的问题, 在密度分布不均数据集、复杂形态数据集和 UCI 数据集上的聚类效果明显优于其对比算法。由于 DPC-WR 的聚类结果往往取决于参数k的选择, 如何快速选择k的最佳值是下一步的研究重点。同时, 群体智能[31-32] 的引入有望进一步提升聚类算法的性能, 特别是借助群智能进化[33-35] 的途径将使得聚类算法的自适应性更加趋于完善。

参考文献:

[1] SZALONTAI B, DEBRECZENY M, FINTOR K, et al. SVD-clustering, a general image-analyzing method explained and demonstrated on model and Raman microspectroscopic maps[J]. Scientific reports, 2020, 10: 4238.

[2] 程鹏宇, 赵嘉, 韩龙哲, 等. 双向多尺度 LSTM 的短时温度预测 [J]. 江西师范大学学报 (自然科学版), 2022, 46(2): 134–139.
CHENG Pengyu, ZHAO Jia, HAN Longzhe, et al. The short-term temperature prediction based on bidirectional multi-scale LSTM[J]. Journal of Jiangxi normal university (natural science edition), 2022, 46(2): 134–139.

[3] LI Chunlin, BAI Jingpan. Automatic content extraction and time-aware topic clustering for large-scale social network on cloud platform[J]. The journal of supercomputing, 2019, 75(5): 2890–2924.

[4] MIRZAL A. Statistical analysis of microarray data clustering using NMF, spectral clustering, kmeans, and GMM[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2022, 19(2): 1173–1192.

[5] ZHAO Yuan, FANG Zhaoyu, LIN Cuixiang, et al. RF-Cell: a gene selection approach for scRNA-seq clustering based on permutation and random forest[J]. Frontiers in genetics, 2021, 12: 665843.

[6] 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.

[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] BUREVA V, POPOV S, TRANEVA V, et al. Generalized net model of cluster analysis using CLIQUE: clustering in quest[J]. International journal bioautomation, 2019, 23(2): 131–138.

[9] REN Jie, WANG Zulin, XU Mai, et al. An EM-based user clustering method in non-orthogonal multiple access[J]. IEEE transactions on communications, 2019, 67(12): 8422–8434.

[10] 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.

[11] ZHANG En, LI Huimin, HUANG Yuchen, et al. Practical multi-party private collaborative k-means clustering[J]. Neurocomputing, 2022, 467: 256–265.

[12] ZHANG Tian, RAMAKRISHNAN R, LIVNY M. Birch[J]. ACM SIGMOD record, 1996, 25(2): 103–114.

[13] AGRAWAL R, GEHRKE J, GUNOPULOS D, et al. Automatic subspace clustering of high dimensional data for data mining applications[C]//Proceedings of the 1998 ACM SIGMOD international conference on Management of data. Seattle: ACM, 1998: 94−105.

[14] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the royal statistical society:series B (methodological), 1977, 39(1): 1–22.

[15] ESTER M. A density-based Algorithm for discovering

clusters in large spatial databases with noise[C]//KDD '96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Oregon: AAAI Press, 1996: 226−231.

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

[17] 吕莉, 朱梅子, 康平, 等. 二阶 K 近邻和多簇合并的密度峰值聚类算法 [J/OL]. 吉林大学学报 (工学版). (2023−01−31)[2023−07−13]. https://kns.cnki.net/kcms/detail/22.1341.t.20230131.1100. 005. html.
LYU Li, ZHU Meizi, KANG Ping, et al. Density peaks clustering with second-order K-nearest neighbors and multi-cluster merging[J/OL]. Journal of Jilin University (engineering and technology edition). (2023−01−31)[2023−07−13]. https://kns.cnki.net/kcms/detail/22.1341.t.20230131.1100.005.html.

[18] SUN Lin, QIN Xiaoying, DING Weiping, et al. Nearest neighbors-based adaptive density peaks clustering with optimized allocation strategy[J]. Neurocomputing, 2022, 473: 159–181.

[19] 赵嘉, 陈磊, 吴润秀, 等. K 近邻和加权相似性的密度峰值聚类算法 [J]. 控制理论与应用, 2022, 39(12): 2349–2357.
ZHAO Jia, CHEN Lei, WU Runxiu, et al. Density peaks clustering algorithm with K-nearest neighbors and weighted similarity[J]. Control theory & applications, 2022, 39(12): 2349–2357.

[20] 吴润秀, 尹士豪, 赵嘉, 等. 基于相对密度估计和多簇合并的密度峰值聚类算法 [J]. 控制与决策, 2023, 38(4): 1047–1055.
WU Runxiu, YIN Shihao, ZHAO Jia, et al. Density peaks clustering based on relative density estimating and multi cluster merging[J]. Control and decision, 2023, 38(4): 1047–1055.

[21] DING Jiajun, HE Xiongxiong, YUAN Junqing, et al. Community detection by propagating the label of center[J]. Physica A:statistical mechanics and its applications, 2018, 503: 675–686.

[22] 赵嘉, 王刚, 吕莉, 等. 面向流形数据的测地距离与余弦互逆近邻密度峰值聚类算法 [J]. 电子学报, 2022, 50(11): 2730–2737.
ZHAO Jia, WANG Gang, LYU 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.

[23] 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, 2017, 30(6): 1109–1121.

[24] 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.

[25] ZHAO Jia, TANG Jingjing, SHI Aiye, et al. Improved density peaks clustering based on firefly algorithm[J]. International journal of bio-inspired computation, 2020, 15(1): 24.

[26] DU Mingjing, DING Shifei, XUE Yu. A robust density peaks clustering algorithm using fuzzy neighborhood[J]. International journal of machine learning and cybernetics, 2018, 9(7): 1131–1140.

[27] YU Donghua, LIU Guojun, GUO Maozu, et al. Density peaks clustering based on weighted local density sequence and nearest neighbor assignment[J]. IEEE access, 2019, 7: 34301–34317.

[28] 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: 2837–2854.

[29] 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.

[30] ZIMMERMAN D W, ZUMBO B D. Relative power of the wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks[J]. The journal of experimental education, 1993, 62(1): 75–86.

[31] 肖人彬. 面向复杂系统的群集智能 [M]. 北京: 科学出版社, 2013.

[32] 肖人彬, 冯振辉, 王甲海. 群体智能的概念辨析与研究进展及应用分析 [J]. 南昌工程学院学报, 2022, 41(1): 1–21.
XIAO Renbin, FENG Zhenhui, WANG Jiahai. Collective intelligence: conception, research progresses and application analyses[J]. Journal of Nanchang institute of technology, 2022, 41(1): 1–21.

[33] 张曦, 李璠, 付雪峰, 等. 随机学习萤火虫算法优化的模糊软子空间聚类算法 [J]. 江西师范大学学报 (自然科学版), 2021, 45(2): 137–144.
ZHANG Xi, LI Fan, FU Xuefeng, et al. The fuzzy soft subspace clustering algorithm optimized by random learning firefly algorithm[J]. Journal of Jiangxi normal university (natural science edition), 2021, 45(2): 137–144.

[34] 肖人彬, 陈峙臻. 从群智能优化到群智能进化 [J]. 南昌工程学院学报, 2023, 42(1): 1–10.
XIAO Renbin, CHEN Zhizhen. From swarm intelligence optimization to swarm intelligence evolution[J]. Journal

of Nanchang institute of technology, 2023, 42(1): 1–10.

[35] 赵嘉, 谢智峰, 吕莉, 等. 深度学习萤火虫算法 [J]. 电子学报, 2018, 46(11): 2633–2641.
ZHAO Jia, XIE Zhifeng, LYU Li, et al. Firefly algorithm with deep learning[J]. Acta electronica sinica, 2018, 46(11): 2633–2641.

作者简介:

吕莉, 教授, 博士, 主要研究方向为智能计算与计算智能、目标跟踪与检测、大数据与人工智能。 主持国家自然科学基金项目 2 项, 发表学术论文 80 余篇。E-mail: lvli623@163.com。

陈威, 硕士研究生, 主要研究方向为数据挖掘。E-mail: chenwei9801@163.com。

肖人彬, 教授, 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。主持国家自然科学基金项目 11 项, 主持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。

第十届中国数据挖掘会议
The 10th China Conference on Data Mining

第十届中国数据挖掘会议(The 10th China Conference on Data Mining, CCDM 2024)由中国计算机学会人工智能与模式识别专业委员会、中国人工智能学会机器学习专委会和济南大学承办, 将于 2024 年 7 月 28—30 日在山东泰安举行。会议现公开征集优秀学术论文(中文), 会议录用的论文将被推荐到合作中文期刊, 欢迎相关领域的研究者和学生积极投稿。

重要日期

征文截止日期: 2024 年 3 月 10 日

录用通知日期: 2024 年 5 月 19 日

注意事项

1) 论文须未公开发表, 会议仅接收中文论文, 采用《计算机研究与发展》格式排版, 一般不超过 6000 字。

2) 论文应包括题目、作者姓名、作者单位、摘要、关键词、正文和参考文献。另附作者通信地址、邮编、电话及 E-mail 地址。

3) 学生(不包括博士后和在职博士生)第一作者的论文稿件请在首页脚注中注明, 否则将不具有参选“优秀学生论文”的资格。

投稿地址: https://conf.ccf.org.cn/CCDM2024/paper

大会官网: https://ccf.org.cn/CCDM2024

联系人: 牛老师, 周老师

联系电话: 13806409965, 17705315838

邮箱: ccdm2024@126.com