山西大学学报(自然科学版)47(1):30—39,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023140

基于簇中心预选策略的三支决策密度峰值聚类算法

罗舒文[1,2],万仁霞[2*],苗夺谦[3]
(1.泉州信息工程学院 通识教育中心,福建 泉州 362000;
2.北方民族大学 数学与信息科学学院,宁夏 银川 750021;
3.同济大学 电子与信息工程学院,上海 201804)

摘 要:本文针对密度峰值聚类算法(CFSFDP)无法自动选取簇中心的不确定性问题,通过引入三支决策理论对其进行优化,提出了一种基于簇中心预选策略的三支决策密度峰值聚类算法(TDPC)。首先利用密度和距离两参数的统计特性将数据对象划分核心域、边界域与琐碎域,符合条件的聚类中心被置于核心域,难以判定的疑似聚类中心点则被置于边界域,然后通过定义的k-可达域和判别准则对疑似聚类中心进行分析,选取出实际聚类中心。所提出算法有效解决了密度峰值聚类算法聚类中心自动确定问题。在2个人工数据集和4个UCI(University of California, lrvine)公共数据集上对TDPC进行测试。与CFSFDP算法和DBSCAN(Density-Based Spatial Clustering of Applications with Noise)算法进行聚类性能比较,所提出算法TDPC在轮廓系数、DB(Davies-Bouldin)指数、调整互信息、调整兰德系数、FM(Fowlkes-Mallows)指数、同质性、完整性等聚类评价指标方面均达到最优或与最优算法结果相近,表明TDPC综合聚类性能优于比较算法,具有良好的聚类可行性与有效性。

关键词:聚类算法;聚类中心;边界域;三支聚类;密度聚类;k-可达域
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)01-0030-10

Three-way Decision-based Density Peak Clustering Algorithm with Clustering Centers Preselection

LUO Shuwen[1,2], WAN Renxia[2*], MIAO Duoqian[3]
(1. *General Education Center, Quanzhou University of Information Engineering, Quanzhou* 362000*, China;*
2. *College of Mathematics and Information Science, North Minzu University, Yinchuan* 750021*, China;*
3. *College of Electronic and Information Engineering, Tongji University, Shanghai* 201804*, China*)

Abstract: Aiming at the uncertainty problem that the CFSFDP (clustering by fast search and find of density peaks) algorithm cannot automatically select the clustering center, in this paper, we propose a three-way decision-based density peak clustering algorithm with clustering centers preselection (TDPC) by incorporating the three-way decision theory. Firstly, the statistical characteristics of density and distance are used to divide the data objects into core region, boundary region and trivial region. The qualified cluster centers are assigned to the core region, and the suspected cluster centers that are difficult to determine are placed in the boundary region. Then the defined k-reachable region and discriminant criterion are used to analyze the suspected cluster centers, and the actual cluster centers are subsequently selected. The proposed algorithm can effectively solve the problem of automatic determination of cluster centers in density peak clustering algorithm. The proposed algorithm is evaluated on two synthetic datasets and four UCI (University

收稿日期:2023-06-08;**接受日期**:2023-09-04
基金项目:国家自然科学基金(61662001);中央高校基本科研业务费专项资金(FWNX04);宁夏自然科学基金(2021AAC03203)
作者简介:罗舒文(1997-),女,黑龙江佳木斯人,硕士研究生,研究方向为三支决策与粗糙集、数据挖掘与模式识别。E-mail:luoshuwen187@163.com
* **通信作者**:万仁霞(WAN Renxia),E-mail: wanrx1022@nmu.edu.cn
引文格式:罗舒文,万仁霞,苗夺谦.基于簇中心预选策略的三支决策密度峰值聚类算法[J].山西大学学报(自然科学版),2024,47(1):30-39. DOI:10.13451/j.sxu.ns.2023140

of California, lrvine) public datasets. Comparing to the CFSFDP algorithm and the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, TDPC demonstrated clustering performance that is on par with or superior to the optimal algorithm across various clustering evaluation indexes, including silhouette coefficient, DB (Davies-Bouldin) index, adjusted mutual information, adjusted rand index, FM (Fowlkes-Mallows) index, homogeneity, and completeness. These results indicate that TDPC outperforms the comparison algorithms in terms of comprehensive clustering performance, and highlight its good clustering feasibility and effectiveness.

Key words: clustering algorithm; clustering center; boundary region; three-way clustering; density clustering; *k*-reachable region

0 引言

聚类分析属于机器学习中的非监督学习[1]。聚类目的是使得属于同一簇内的对象之间是彼此相似的,不同簇的对象之间是不相似的。2014年,Rodriguez等在《Science》上发布了一种基于密度的聚类算法——快速搜索和寻找密度峰值聚类算法[2](Clustering by fast search and find of density peaks,CFSFDP)。该算法根据聚类中心本身局部密度大且与其他密度更大点相对距离较远的特征来绘制出决策图,进而确定聚类中心点。由于CFSFDP算法具有原理简单、聚类效率高和无须迭代就能实现对数据集处理等优点,其一经提出,便受到学界的广泛关注,王飞等[3]将数据网格化,在网格中进行密度峰值聚类并利用网格边界将结果合并从而减小了复杂度和计算量。谢娟英等[4]引入K近邻信息度量样本密度,提出两种基于K近邻的样本分配策略,避免了密度峰值聚类算法样本分配策略的缺陷。高润峰等[5]利用罚处思想使高密度簇非聚类中心的局部密度变小,令低密度族中的聚类中心得以显现,同时采用迭代阈值法将数据分为普通点和簇中心点。朱庆峰等[6]通过K近邻相似度函数判断密度峰值聚类结果中的密度与距离是否正确,从而有效减少了错误分配。Zhao等[7]结合K近邻和模糊邻域得到一个最近邻模糊核函数。然后,通过最近邻模糊核函数重新定义局部密度,更好地表征样本的分布特征。以上针对CFSFDP算法的改进主要措施在于优化决策图,但仍然不可避免地要人工通过决策图判断识别聚类中心,而通过其他迭代算法对决策图进行识别划分又存在效率不高的问题。

2010年,加拿大学者姚一豫提出了三支决策(Three-way Decision,TWD)的思想与概念[8-9],通过决策粗糙集理论扩充,将论域划分为正域、边界域和负域三个部分,能较好地处理实际决策过程中出现的不确定性问题[10]。由于三支决策符合人类思维和认知特点,国内外学者对其进行深入研究[11-12]。贾修一等[13]对三支决策当中最重要的阈值计算给出了一种新的方法,能够将原有的决策风险有效的降低。刘盾等[14]从三支决策的角度对粗糙集进行了系统的研究和梳理。苗夺谦等[15]总结了三支决策的粒计算方法,为三支决策粒计算模型展示了应用前景。杨霁琳等[16]将阈值对引入模糊粗糙集模型,用模糊近似的三支决策理论给出了代价函数的计算方法。李波等[17]将多类别三支决策应用到意识别中,对未识别出意图的目标进行延迟决策,更新目标信息后重新进行识别。于洪[18]受三支决策理论启发,提出三支决策聚类算法,该算法的思路是用核心域和边界域两个集合表示一个类,其中核心域中的数据肯定属于该类,边界域的数据可能属于也可能不属于该类。Yao等[19]基于认知科学,重新讨论了三支决策和粒计算的关系,并提出了TAO模型(Trisecting-Acting-Outcome model)为三支决策的发展提供新的思路。

由于CFSFDP算法需要在决策图中人工选取聚类中心,若数据集分布相对复杂,对于聚类中心的选取人工无法做到高效精确,而三支决策理论是在实际决策过程中处理不确定性问题的有效工具,因此本文针对密度峰值聚类算法无法自动确定簇中心的缺陷,提出一种基于簇中心预选策略的三支决策密度峰值聚类算法(Three-Way Decision-Based Density Peak Clustering Algorithm with Clustering Centers Preselection,TDPC)。通过引入三支决策理论,构建预选策略,使得算法既能自动划分聚类中心点,而且也不需要引入额外的算法辅助决策,在时

间不增加的基础上,有效解决密度峰值聚类算法在选取簇中心时的不确定性的问题,并提高了聚类准确度。

1 相关理论

1.1 密度峰值聚类算法(CFSFDP)

CFSFDP算法选取聚类中心基于以下假设[20]:

(1)聚类中心有较高的密度,围绕聚类中心的邻近样本点的密度比它低;

(2)聚类中心与密度比其高的数据点之间的距离相对较大。

在密度峰值聚类模型中,对数据集中的每个数据点都需要计算两个变量:该数据点的局部密度ρ,以及密度比该数据点大的高密度点之间的最小距离δ。数据点x_i的局部密度ρ_i计算公式如下:

$$\rho_i=\sum_j \chi(d_{ij}-d_c),\chi(x)=\begin{cases}1,x<0,\\0,\text{其他},\end{cases} \tag{1}$$

其中d_{ij}为数据点x_i与其他数据点x_j之间的欧式距离;d_c为全局的截断距离。数据点x_i的相对距离δ_i计算公式为:

$$\delta_i=\min_{j:\rho_j>\rho_i}(d_{ij})。 \tag{2}$$

当数据点是数据集中密度最大的点时,局部密度选取为数据集所有数据点中与其相距最远的距离作为它的δ值,计算公式为:

$$\delta_i=\max_j(d_{ij})。 \tag{3}$$

密度峰值聚类算法以ρ为横轴表示数据点的密度,δ为纵轴表示数据点的距离,通过ρ与δ绘制出决策图。因聚类中心数据点同时具有较大的密度和距离,将出现在二维坐标系中的右上部分的数据点标识为聚类中心,求出聚类簇数。对于某些分布较为复杂的数据集,计算出每个数据点密度ρ与距离δ的乘积γ,将γ值按照从大到小进行降序排列,数据点的γ值越大越有可能成为聚类中心。密度峰值聚类算法中对剩余数据点的分配规则是,首先按数据点密度降序排序,依次将数据点分配到局部密度比自身大且相距最近的数据点所属的类中。

1.2 三支决策

在粗糙集理论中,基于上下近似的概念,将论域U中的某对象集A区分成正域$POS(A)$、边界域$BND(A)$和负域$NEG(A)$三个区域[21]。三支决策中,正域代表接受决策、边界域代表延迟决策,负域代表拒绝决策。根据贝叶斯决策过程,决策粗糙集由两元素组成,分别为两类状态集与三种行动集。状态集记为$\Omega=(X,\neg X)$,其中状态X表示对象满足某个条件C,$\neg X$表示不满足条件C。两状态形成互补关系。行动集记为$A=\{a_P,a_B,a_N\}$,其中a_P代表接受决策,a_B代表延迟决策,a_N代表拒绝决策。当对象满足条件C时,采取行动a_P,a_B,a_N对应的损失值分别表示为$\lambda_{PP},\lambda_{BP},\lambda_{NP}$。当对象不满足条件$C$时,采取行动$a_P,a_B,a_N$对应的损失值分别表示为$\lambda_{PN},\lambda_{BN},\lambda_{NN}$。对象满足$C$的条件概率记为$P\left(X|[X]\right)$,以该概率作为评价函数。

由贝叶斯公式及损失值间不等关系得到以下决策准则:

接受(P):当$P\left(X|[X]\right)\geqslant\alpha$;

延迟(B):当$\beta\leqslant P\left(X|[X]\right)<\alpha$;

拒绝(N):当$P\left(X|[X]\right)\leqslant\beta$。

在基于决策粗糙集的三支决策中,最佳α、β值为:

$$\alpha=\frac{(\lambda_{PN}-\lambda_{BN})}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BN}-\lambda_{PP})}, \tag{4}$$

$$\beta=\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}。 \tag{5}$$

2 基于簇中心预选策略的三支决策密度峰值聚类算法(TDPC)

CFSFDP算法需要根据决策图人工选取聚类中心点,具有强烈的主观性和不确定性,不利于算法的实现与应用。相比于非聚类中心,聚类中心往往具有较高的ρ值和δ值,转变过程中通常有一个明显的跳跃阈值,自动获取聚类中心方法正是基于这种思想,选取突增的点作为聚类中心。但对于一些数据集,聚类中心点与非聚类中心点变化可能并不显著,可能会出现遗漏或错选现象。

针对上述情况,本文引入三支决策理论对密度峰值聚类算法进行优化,提出一种预选策略的簇中心确定方法。算法分为两个部分,首

先利用密度和距离的统计特性将符合严格条件的密度峰值点作为聚类中心放入核心域,将难以判定的疑似聚类中心数据点放入边界域,然后通过判别规则对疑似聚类中心进行明确判别,得到实际聚类中心。

2.1 疑似聚类中心的选取

定义1 簇中心权值[5]作为划分聚类中心点的依据:

$$\gamma_i=\rho_i\cdot\delta_i。\tag{6}$$

由定义1知γ_i值较大的点为聚类中心点,将所有的γ_i值从大到小降序排列,在聚类中心到非聚类中心的过渡中,通过γ_i的最大跳跃下降作为判断拐点的依据[22],本文自动获取聚类中心的方法正是利用该思想进行实现的。

定义2 阈值点α:

$$\alpha=\arg\left(\max\frac{|\gamma_i-\gamma_{i+1}|}{|\gamma_1-\gamma_i|}\right),i=1,2,\cdots,n-1。\tag{7}$$

对降序排列后的γ_i值相邻两数做差,找出与$|\gamma_1-\gamma_i|$比值的最大值对应点作为阈值点α。

定义3 计算出阈值点α前所有点γ的差值平均值$\overline{\gamma_\alpha}$:

$$\overline{\gamma_\alpha}=\frac{\sum_{i=1}^{\alpha}|\gamma_i-\gamma_{i+1}|}{\alpha}。\tag{8}$$

定义4 计算出数据集中所有γ值的差值平均值$\overline{\gamma_\beta}$:

$$\overline{\gamma_\beta}=\frac{\sum_{i=1}^{n-1}|\gamma_i-\gamma_{i+1}|}{(n-1)},i=1,2,\cdots,n-1。\tag{9}$$

定义5 对于点i预划分规则如下:

若$\gamma_i>\overline{\gamma_\alpha}$则:$i\in POS(X),i=1,2,\cdots,n-1$;

若$\overline{\gamma_\alpha}\geqslant\gamma_i\geqslant\overline{\gamma_\beta}$则:$i\in BND(X),i=1,2,\cdots,n-1$;

若$\gamma_i<\overline{\gamma_\beta}$则:$i\in NEG(X),i=1,2,\cdots,n-1$。

若数据点簇中心权值大于阈值点α前所有点的差值平均值,则将该点作为聚类中心点放入核心域(即正域)。

若数据点簇中心权值小于阈值点α前的点的差值平均值且大于所有数据点的差值平均值,则将该点作为疑似聚类中心点放入边界域(即中间域),需要进一步判别该点是否为聚类中心点。

若数据点簇中心权值差小于所有数据点的差值平均值,则将该点作为非聚类中心点放入琐碎域(即负域)。

2.2 实际聚类中心的筛选

密度峰值聚类算法最为重要的内容是对于聚类中心点的计算和选取,聚类中心点是那些本身密度较大,同时与比它密度更高点距离较远的点。因此若疑似聚类中心距离更高密度点极近,则说明该点并非密度峰值点。由于密度峰值聚类算法的性质,密度更高的点更可能为聚类中心。

定义6 邻域$N(P)$[23]:与点P距离小于截断距离d_c的点组成的集合称为P的邻域$N(P)$。

定义7 k-可达域:与点P距离小于kd_c的点组成的集合称为P的k-可达域$kN(P)$,k称为可达因子。

对边界域中的疑似聚类中心点按照下列规则依次筛选,判别规则如下:

(P1)若该疑似聚类中心点的k-可达域中存在核心域中的点,则说明该点距离更高密度点极近,故该点并非密度峰值点,将其划入琐碎域;否则进行(P2)。

(P2)若该疑似聚类中心点的k-可达域中不存在核心域中的点,且也不存在其他疑似中心点,则将其认定为聚类中心;否则进行(P3)。

(P3)若该疑似聚类中心点的k-可达域中不存在核心域中的点,但存在其他疑似中心点,则需要判断该点与其k-可达域中包含的其他疑似中心点的密度大小,若存在密度比它大的点,则该点并非密度峰值点;若不存在密度比它大的点,则认定该点为聚类中心。

通过以上规则对边界域中每个疑似聚类中心点进行判别,最终筛选出实际聚类中心。

2.3 TDPC算法步骤

输入:数据集$X=\{x_1,x_2,\cdots,x_i,\cdots,x_N\}$,截断距离$d_c$,可达因子$k$;

输出:聚类中心$M=\{m_1,m_2,\cdots,m_n\}$,n为聚类中心个数。

Step1:计算每个数据点的局部密度ρ,相对距离δ;

Step2:根据定义1计算每个数据点的γ值并降序排列;

Step3:根据定义2计算阈值点α;

Step4:根据定义3和定义4计算$\overline{\gamma_\alpha}$与$\overline{\gamma_\beta}$;

Step5:根据$\overline{\gamma_\alpha}$值与$\overline{\gamma_\beta}$值对每个数据点进行划分:(1)若$\gamma_i>\overline{\gamma_\alpha}$,则将该点作为聚类中心点放入核心域,即$i\in POS(X)$;(2)若$\overline{\gamma_\alpha}\geqslant\gamma_i\geqslant\overline{\gamma_\beta}$,则将该点作为疑似聚类中心点放入边界域,即$i\in BND(X)$;(3)若$\gamma_i<\overline{\gamma_\beta}$,则将该点之作为非聚类中心点放入琐碎域,即$i\in NEG(X)$;

Step6:通过下列判别准则(P1—P3)对边界域$BND(X)$中的疑似聚类中心点进行明确分析:

(P1)对于$A\in BND(X)$,若$\exists B\in POS(X)$,且$B\in kN(A)$,则点A划入琐碎域中;否则进行(P2),如图1所示。

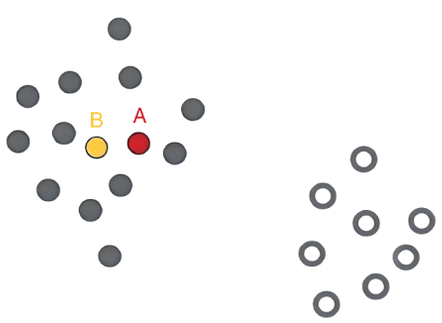

图1 P1规则判别示例

Fig. 1 Example of P1 rule discrimination

(P2)对于$A\in BND(X)$,若$\forall B\in POS(X)$,都有$B\notin kN(A)$,且$\forall C\in BND(X)$,都有$C\notin kN(A)$,则将点A划入核心域中;否则进行(P3),如图2所示。

(P3)对于$A\in BND(X)$,若$\forall B\in POS(X)$,都有$B\notin kN(A)$,但是$\exists C\in BND(X)$,使得$C\in kN(A)$,则对比点A与点C的密度ρ,若不存在密度比点A大的点,则将点A划入核心域中;若存在,则将其划入琐碎域中,如图3所示。

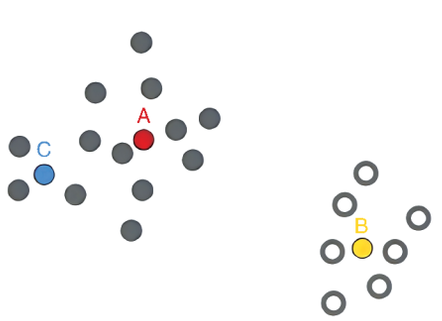

图2 P2规则判别示例

Fig. 2 Example of P2 rule discrimination

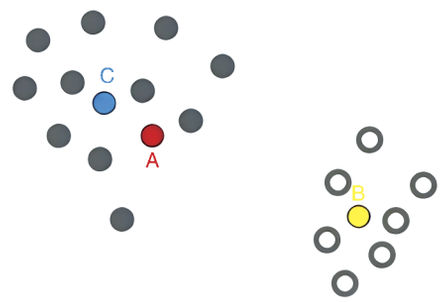

图3 P3规则判别示例

Fig. 3 Example of P3 rule discrimination

Step7:算法结束,输出聚类中心。

可以看出,对于样本量为n的数据集,该算法只需计算该点与剩余点的距离即可,其他步骤只是进行对比判断,因此,TDPC的时间复杂度与CFSFDP的时间复杂度相同,均为$O(n^2)$。

3 实验与结果分析

实验使用Windows 10-64位操作系统,实验在Inter(R)core(TM)i7-7700hQCPU,2.40 GHz,4.00 GB,Anaconda Python 3.8环境中进行。实验采用人工数据集和UCI(University of California,lrvine)数据集,如表1所示,其中D1、D2为人工数据集,其余4个数据集来自UCI数据库。人工数据集主要用以直观展示TDPC、CFSFDP、DBSCAN(Density-Based Spatial Clustering of Applications with Noise)[24]的聚类效果,UCI数据集用于测试算法的聚类评价指标。

算法的参数设置会对算法结果产生较大影响,为了呈现各算法的最佳性能,本文实验时,

表1 实验数据集

Table 1 Experimental datasets

数据集	样本总数	属性	类簇
D1	400	2	2
D2	2 000	2	5
Ecoil	336	7	8
Iris	150	4	3
vehicle	846	18	4
Wine	178	13	3

DBSCAN、CFSFDP 和 TDPC 三个算法对每个实验数据集进行 100 次实验,选取最优聚类结果对应的参数。为了描述方便,各算法的参数依次用 par1,par2,par3,…表示,其中,DBSCAN 算法中,Par1 代表 Eps 邻域半径大小,Par2 代表邻域中样本点个数;CFSFDP 算法中,Par1 代表全局截断距离 d_c,表示样本点近邻数目占总数目的百分比;TDPC 算法中,Par1 的含义和设置与 CFSFDP 保持一致,Par2 为可达因子。

3.1 聚类人工数据集

为了直观显示算法的聚类效果,选取人工数据集 D1、D2 进行比较实验。TDPC 算法、CFSFDP 算法和 DBSCAN 算法在人工数据集的参数设置如表 2 所示。

聚类结果如图 4—图 11 所示。

表 2 算法在人工数据集上的参数设置

Table 2 Parameter setting of algorithm on synthetic datasets

数据集	参数	DBSCAN	CFSFDP	TDPC
D1	Par1	0.3	2%	2%
	Par2	10		8
D2	Par1	0.6	2%	2%
	Par2	10		5.65

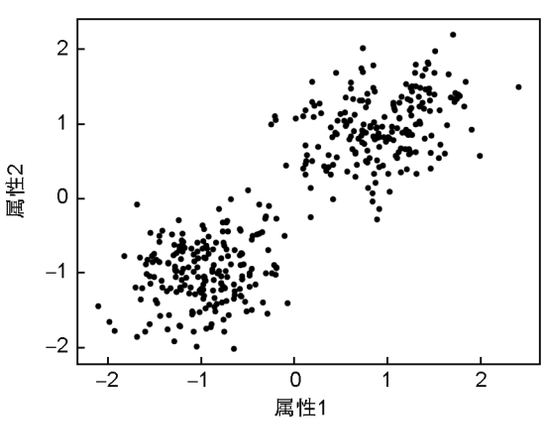

图 4 数据集 D1 的原始点分布图

Fig. 4 Original point distribution of dataset D1

人工数据集 D1 的样本数量较少,类别个数也较少。图 4 是未分类的 D1 数据集,从图中可以看出数据点比较明显地分成了两类。图 5 是 DBSCAN 算法聚类结果,其中黑色表示噪声点。图 6 是 CFSFDP 算法聚类结果。图 7 为本文 TDPC 算法的实验效果图,从两个图中可看出明显看出对于两类的划分基本一致,相比于 DBSCAN 得到了更好的聚类效果,说明运用本文算法最终可得正确的聚类结果。

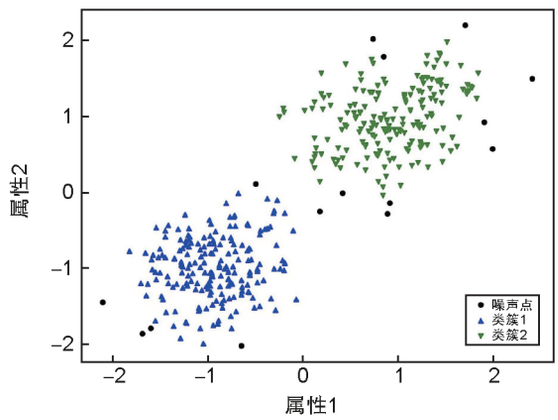

图 5 DBSCAN 算法在 D1 上的聚类结果

Fig. 5 Clustering results of DBSCAN algorithm in D1

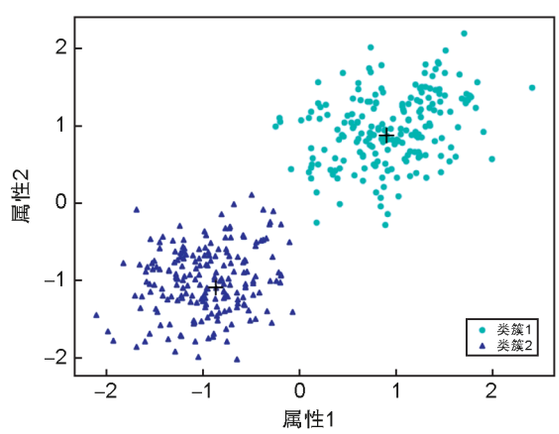

图 6 CFSFDP 算法在 D1 上的聚类结果

Fig. 6 Clustering results of CFSFDP algorithm in D1

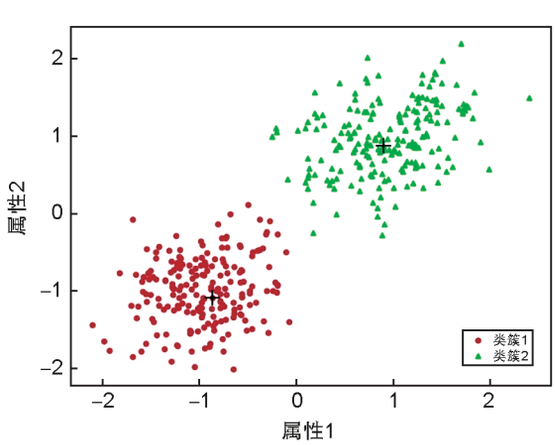

图 7 TDPC 算法在 D1 上的聚类结果

Fig. 7 Clustering results of TDPC algorithm in D1

人工数据集 D2 的样本数量较多,类别个数也相对较多。图 8 是未分类的 D2 数据集。图 9 是 DBSCAN 算法在 D2 上的聚类结果,其中黑色表示噪声点。图 10 是 CFSFDP 算法在 D2 上的聚

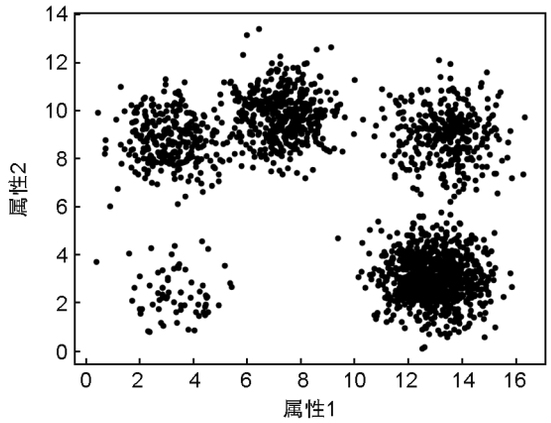

图8 数据集D2的原始点分布图

Fig. 8 Original point distribution of dataset D2

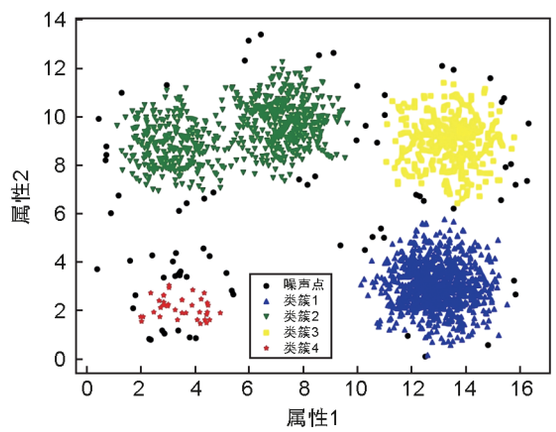

图9 DBSCAN算法在D2上的聚类结果

Fig. 9 Clustering results of DBSCAN algorithm in D2

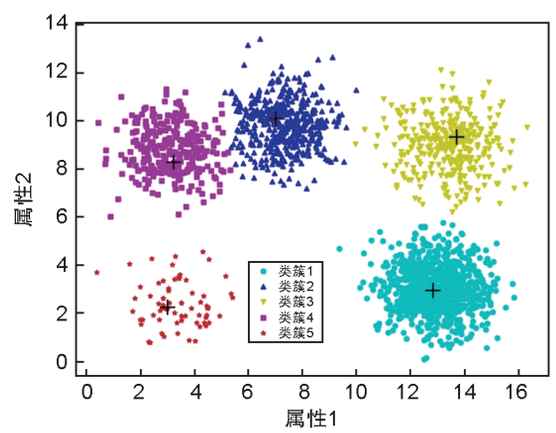

图10 CFSFDP算法在D2上的聚类结果

Fig. 10 Clustering results of CFSFDP algorithm in D2

类结果,可以看到该算法得到了5个类。图11为本文TDPC算法在D2上的实验效果图,由图可知运用本文算法最终可得正确的聚类结果。

3.2 聚类UCI标准数据集与评价指标

为进一步检验算法性能,本文选取UCI标准数

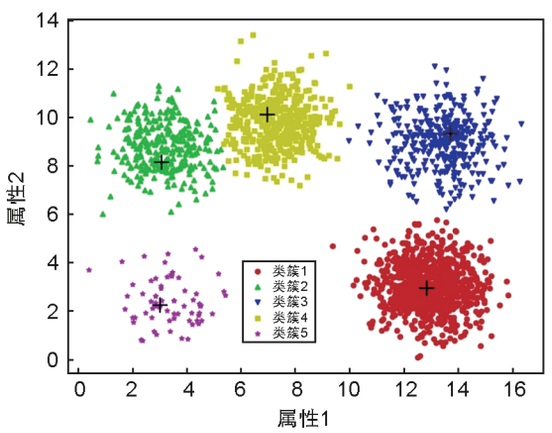

图11 TDPC算法在D2上的聚类结果

Fig. 11 Clustering results of TDPC algorithm in D2

据集中的4组数据进行实验,从七种聚类的评价指标来比较算法性能。实验时参数设置如表3。

表3 算法在UCI标准数据集上的参数设置

Table 3 Parameter setting of algorithm on UCI standard datasets

数据集	参数	DBSCAN	CFSFDP	TDPC
Ecoil	Par1	0.4	1%	1%
	Par2	4		3.05
Iris	Par1	0.4	1%	1%
	Par2	10		4
Vehicle	Par1	0.7	2%	2%
	Par2	6		3
Wine	Par1	5	2%	2%
	Par2	2		28

评价指标有如下几种:

轮廓系数[25]:

$$s=\frac{b-a}{\max(a,b)}, \tag{10}$$

轮廓系数是一组数据集的轮廓系数等于该数据集中每一个样本轮廓系数的平均值,其中a表示样本对象i与同一类簇当中其他样本对象j的平均距离;b表示样本对象i与最近的一个类簇当中所有样本对象n的平均距离。轮廓系数处于[−1,1]的范围内。

DBI(Davies-Bouldin 指数)[26]:

$$I_{\mathrm{DBI}}=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\bar{c}_i+\bar{c}_j}{\left\|w_i-w_j\right\|_2}\right)。 \tag{11}$$

上式之中,$\bar{c}_i$表示第i类中所有样本到聚类中心w_i的平均距离,$\left\|w_i-w_j\right\|_2$表示类i与类j

聚类中心之间的欧式距离,k表示聚类数。DB指数越小,聚类性能越好。

调整互信息(AMI)[27]:

$$I_{\mathrm{AMI}}=\frac{MI(U,V)}{\sqrt{H(U)\cdot H(V)}},\tag{12}$$

调整兰德系数(ARI)[28]:

$$I_{\mathrm{ARI}}=\frac{2(ad-bc)}{(b+d)(a+c)+(c+d)(a+b)},\tag{13}$$

FM指数(Fowlkes-Mallows Index)[28]:

$$I_{\mathrm{FMI}}=\frac{a}{\sqrt{(a+b)(a+c)}}。\tag{14}$$

U为数据的真实标签,V为数据的聚类标签。a表示在U中为同一簇且在V中也为同一簇的数据点个数。b表示在U中为同一类但在V中不为同一簇的数据点个数,c表示在U中不为同一簇但在V中为同一簇的数据点个数,d表示在U中不为同一簇且在V中也不为同一簇的数据点个数。AMI、ARI、FMI三者的取值范围分别为:$I_{\mathrm{AMI}}\in[-1,1]$,$I_{\mathrm{ARI}}\in[-1,1]$,$I_{\mathrm{FMI}}\in[0,1]$。三者都是越接近1越理想,即越大代表聚类效果越好。

同质性(homogeneity)指类簇都只包含单个类的数据对象。

完整性(completeness)指给定类中的所有数据对象都被划分到同一类簇中。

同质性与完整性取值范围为[0,1],取值越大代表聚类结果与真实结果越贴近。同质性和完整性对应公式如下[29]:

$$\begin{aligned}h&=1-\frac{H(C|K)}{H(C)},\\ c&=1-\frac{H(K|C)}{H(K)}。\end{aligned}\tag{15}$$

其中,$H(C|K)$是给定类簇条件下类别划分的条件熵,$H(C)$是类别划分熵,类簇划分的条件熵$H(K|C)$与类簇划分熵$H(K)$以对称的形式定义。n表示样本总数,n_c表示类别c下样本数,n_k表示类簇k下样本数,$n_{c,k}$表示类别c中被分配给类簇k的样本数。

3.3 实验结果

基于簇中心预选策略的三支决策密度峰值聚类算法在数据集Ecoli、iris、vehicle、wine均比其他算法表现更好。实验结果如表4所示。

表4 数据集上的实验结果对比

Table 4 Comparison of experimental results on datasets

数据集	聚类指标	DBSCAN	CFSFDP	TDPC
Ecoil	轮廓系数	0.229 1	0.218 7	0.268 7
	DBI	0.817 8	1.352 9	0.816 9
	AMI	0.118 0	0.405 0	0.629 6
	ARI	0.038 3	0.260 9	0.670 4
	FMI	0.531 0	0.503 3	0.770 7
	同质性	0.076 9	0.347 4	0.595 9
	完整性	0.720 1	0.517 7	0.702 6
Iris	轮廓系数	0.121 3	0.533 7	0.488 1
	DBI	2.075 8	0.726 1	0.469 1
	AMI	0.563 0	0.643 9	0.715 6
	ARI	0.456 9	0.442 8	0.558 4
	FMI	0.656 1	0.677 9	0.764 2
	同质性	0.566 6	0.579 4	0.579 4
	完整性	0.576 5	0.737 3	0.951 2
Vehicle	轮廓系数	0.068 5	0.149 2	0.240 3
	DBI	3.125 2	1.728 6	0.869 4
	AMI	0.035 1	0.109 1	0.124 3
	ARI	0.000 5	0.068 3	0.067 9
	FMI	0.453 3	0.388 7	0.424 3
	同质性	0.026 6	0.091 9	0.099 6
	完整性	0.094 2	0.143 1	0.199 5
Wine	轮廓系数	−0.609 6	0.570 8	0.570 8
	DBI	2.972 7	0.531 7	0.531 7
	AMI	0.066 7	0.413 1	0.413 1
	ARI	−0.007 4	0.371 5	0.371 5
	FMI	0.495 5	0.583 4	0.583 4
	同质性	0.129 9	0.419 8	0.419 8
	完整性	0.187 2	0.418 9	0.418 9

通过上述对比实验可以看出,与密度峰值聚类算法CFSFDP以及DBSCAN算法进行比较,本文提出的三支决策密度峰值聚类算法TDPC具有更好的聚类效果。即使在个别实验数据集上的个别性能不是最突出的,但与最佳效果也是极为接近的。这是因为TDPC算法是对标准CFSFDP算法上的改进,它继承了CFSFDP算法的优势,同时TDPC算法能够自动获取聚类中心,从而有效避免CFSFDP算法中人工选取聚类中心的不确定性问题。

4 结论

本文将三支决策思想引入密度峰值聚类算法,提出了能够自动获取聚类中心点的三支决策密度峰值聚类算法。算法通过“疑似聚类中

心的选取”策略,将满足一定统计特性的数据点放入边界域,有效避免了原密度峰值聚类算法遗漏或错选聚类中心的现象发生,而算法的“实际聚类中心的筛选”策略则保证了聚类中心的最大代表性。在不同数据集上的实验结果证明了所提出算法的有效性。未来工作将致力于不同“距离”度量下算法的有效性研究。

参考文献:

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine Learning[M]. Beijing: Tsinghua University Press, 2016.

[2] RODRIGUEZ A, LAIO A. Clustering by Fast Search and Find of Density Peaks[J]. *Science*, 2014, **344**(6191): 1492−1496. DOI: 10.1126/science.1242072.

[3] 王飞, 王国胤, 李智星, 等. 一种基于网格的密度峰值聚类算法[J]. 小型微型计算机系统, 2017, **38**(5):1034−1038. DOI: 10.3969/j.issn.1671-1122.2015.04.008.
WANG F, WANG G Y, LI Z X, *et al*. Clustering by Fast Search and Find of Density Peaks Based on Grid[J]. *J Chin Comput Syst*, 2017, **38**(5): 1034−1038. DOI: 10.3969/j.issn.1671-1122.2015.04.008.

[4] 谢娟英, 高红超, 谢维信. K近邻优化的密度峰值快速搜索聚类算法[J]. 中国科学(信息科学), 2016, **46**(2): 258−280. DOI: 10.1360/N112015-00135.
XIE J Y, GAO H C, XIE W X. K-nearest Neighbor Optimization Fast Search Clustering Algorithm for Density Peak[J]. *Sci Sin Informationis*, 2016, **46**(2): 258−280. DOI: 10.1360/N112015-00135.

[5] 高润峰, 苏一丹, 覃华. 罚处共享最近邻密度峰聚类算法[J]. 计算机工程与设计, 2021, **42**(12): 3407−3414. DOI: 10.16208/j.issn1000-7024.2021.12.015.
GAO R F, SU Y D, QIN H. Penalty Shared Nearest Neighbor Density Peak Clustering Algorithm[J]. *Comput Eng Des*, 2021, **42**(12): 3407−3414. DOI: 10.16208/j.issn1000-7024.2021.12.015.

[6] 朱庆峰, 葛洪伟. K近邻相似度优化的密度峰聚类[J]. 计算机工程与应用, 2019, **55**(2): 148−153. DOI: 10.3778/j.issn.1002-8331.1710-0059.
ZHU Q F, GE H W. Density Peaks Clustering Optimized by K Nearest Neighbor's Similarity[J]. *Comput Eng Appl*, 2019, **55**(2): 148−153. DOI: 10.3778/j. issn. 1002-8331.1710-0059.

[7] 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 Recognit*, 2023, **139**: 109406. DOI: 10.1016/j.patcog.2023.109406.

[8] YAO Y Y. Three-way Decisions with Probabilistic Rough Sets[J]. *Inf Sci*, 2010, **180**(3): 341−353. DOI: 10.1016/j.ins.2009.09.021.

[9] YAO Y Y. An Outline of a Theory of Three-way Decisions[M]//Rough Sets and Current Trends in Computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 1−17. DOI: 10.1007/978-3-642-32115-3_1.

[10] 刘盾, 李天瑞, 杨新, 等. 三支决策-基于粗糙集与粒计算研究视角[J]. 智能系统学报, 2019, **14**(6): 1111−1120. DOI: 10.11992/tis.201905039.
LIU D, LI T R, YANG X, *et al*. Three-way Decisions: Research Perspectives for Rough Sets and Granular Computing[J]. *CAAI Trans Intell Syst*, 2019, **14**(6): 1111−1120. DOI: 10.11992/tis.201905039.

[11] 索郎王青, 杨海龙, 姚一豫. 三元思维: 三支决策理论与实践[J]. 陕西师范大学学报(自然科学版), 2022, **50**(3): 7−16. DOI: 10.15983/j.cnki.jsnu.2022102.
LANGWANGQING S, YANG H L, YAO Y Y. Triadic Thinking: The Theory and Practice of Three-way Decision[J]. *J Shaanxi Norm Univ Nat Sci Ed*, 2022, **50**(3): 7−16. DOI: 10.15983/j.cnki.jsnu.2022102.

[12] 万仁霞, 王大庆, 苗夺谦. 基于三支决策的高斯混合聚类研究[J]. 重庆邮电大学学报(自然科学版), 2021, **33**(5): 806−815. DOI: 10.3979/j.issn.1673-825X.202105200168.
WAN R X, WANG D Q, MIAO D Q. Gaussian Mixture Clustering Based on Three-way Decision[J]. *J Chongqing Univ Posts Telecommun Nat Sci Ed*, 2021, **33**(5): 806−815. DOI: 10.3979/j.issn.1673-825X.202105200168.

[13] 贾修一, 商琳. 一种求三支决策阈值的模拟退火算法[J]. 小型微型计算机系统, 2013, **34**(11): 2603−2606. DOI: 10.3969/j.issn.1000-1220.2013.11.038.
JIA X Y, SHANG L. A Simulated Annealing Algorithm for Learning Thresholds in Three-way Decision-theoretic Rough Set Model[J]. *J Chin Comput Syst*, 2013, **34**(11): 2603−2606. DOI: 10.3969/j.issn.1000-1220.2013.11.038.

[14] 刘盾, 李天瑞, 李华雄. 粗糙集理论: 基于三支决策视角[J]. 南京大学学报(自然科学版), 2013, **49**(5): 574−581.
LIU D, LI T R, LI H. Rough Set Theory: A Three-way Decisions Perspective[J]. *J Nanjing Univ Nat Sci*, 2013, **49**(5): 574−581.

[15] 苗夺谦, 张清华, 钱宇华, 等. 从人类智能到机器实现模型: 粒计算理论与方法[J]. 智能系统学报, 2016, **11**(6): 743−757. DOI: 10.11992/tis.201612014.
MIAO D Q, ZHANG Q H, QIAN Y H, *et al*. From Human Intelligence to Machine Implementation Model: Theories and Applications Based on Granular Comput-

ing[J]. *CAAI Trans Intell Syst*, 2016, **11**(6): 743−757. DOI: 10.11992/tis.201612014.

[16] 杨霁琳, 张贤勇, 唐孝, 等. 基于最优相似度三支决策的模糊粗糙集模型[J]. 计算机科学, 2018, **45**(10): 27−32. DOI: 10.11896/j.issn.1002-137X.2018.10.005.
YANG J L, ZHANG X Y, TANG X, *et al*. Fuzzy Rough Set Model Based on Three-way Decisions of Optimal Similar Degrees[J]. *Comput Sci*, 2018, **45**(10): 27−32. DOI: 10.11896/j.issn.1002-137X.2018.10.005.

[17] 李波, 越凯强, 范盘龙, 等. 基于序贯三支决策的目标作战意图识别方法[J]. 陕西师范大学学报(自然科学版), 2022, **50**(3): 17−23. DOI: 10.11809/bqzbgcxb2020.07.035.
LI B, YUE K Q, FAN P L, *et al*. Combat Intention Identification of Target Based on Sequential Three-way Decision[J]. *J Shaanxi Norm Univ Nat Sci Ed*, 2022, **50**(3): 17−23. DOI: 10.11809/bqzbgcxb2020.07.035.

[18] 于洪. 三支聚类分析[J]. 数码设计, 2016(1): 31−35.
YU H. Three-way Cluster Analysis[J]. *Peak Data Sci*, 2016(1): 31−35.

[19] YAO J T, YAO Y Y, CIUCCI D, *et al*. Granular Computing and Three-way Decisions for Cognitive Analytics[J]. *Cogn Comput*, 2022, **14**(6): 1801−1804. DOI: 10.1007/s12559-022-10028-0.

[20] 杨洁, 王国胤, 王飞. 基于密度峰值的网格聚类算法[J]. 计算机应用, 2017, **37**(11): 3080−3084. DOI: 10.11772/j.issn.1001-9081.2017.11.3080.
YANG J, WANG G Y, WANG F. Grid Clustering Algorithm Based on Density Peaks[J]. *J Comput Appl*, 2017, **37**(11): 3080−3084. DOI: 10.11772/j. issn. 1001-9081.2017.11.3080.

[21] PAWLAK Z. Rough Sets[J]. *Int J Comput Inf Sci*, 1982, **11**(5): 341−356. DOI: 10.1007/BF01001956.

[22] 吴斌, 卢红丽, 江惠君. 自适应密度峰值聚类算法[J]. 计算机应用, 2020, **40**(6): 1654−1661. DOI: 10.11772/j.issn.1001-9081.2019111881.
WU B, LU H L, JIANG H J. Adaptive Density Peaks Clustering Algorithm[J]. *J Comput Appl*, 2020, **40**(6): 1654−1661. DOI: 10.11772/j.issn.1001-9081.2019111881.

[23] 袁英, 陈立潮, 任姚鹏, 等. 结合引力的模糊C-值聚类算法研究[J]. 计算机应用与软件, 2010, **27**(8): 271−272. DOI: 10.3969/j.issn.1000-386X.2010.08.084.
YUAN Y, CHEN L C, REN Y P, *et al*. Research on Fuzzy C-means Clustering Algorithm Combining Gravity[J]. *Comput Appl Softw*, 2010, **27**(8): 271−272. DOI: 10.3969/j.issn.1000-386X.2010.08.084.

[24] 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 Second International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1996: 226−231. DOI: 10.5555/3001460.3001507.

[25] FAHAD A, ALSHATRI N, TARI Z, *et al*. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis[J]. *IEEE Trans Emerg Top Comput*, 2014, **2**(3): 267−279. DOI: 10.1109/TETC.2014.2330519.

[26] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008, **19**(1):48−61. DOI: 10.3724/SP.J.1001.2008.00048.
SUN J G, LIU J, ZHAO L Y. Clustering Algorithms Research[J]. *J Softw*, 2008, **19**(1):48−61. DOI: 10.3724/SP.J.1001.2008.00048.

[27] VINH N X, EPPS J, BAILEY J. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?[C]//Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009: 1073−1080. DOI: 10.1145/1553374.1553511.

[28] FOWLKES E B, MALLOWS C L. A Method for Comparing Two Hierarchical Clusterings: Rejoinder[J]. *J Am Stat Assoc*, 1983, **78**(383): 584. DOI: 10.2307/2288123.

[29] ROSENBERG A, HIRSCHBERG J. V-measure: A Conditional Entropy-based External Cluster Evaluation measure[C]//Proceedings of the 2007 Joint Conference On Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP−CoNLL). 2007: 410−420.