山西大学学报(自然科学版)48(1):169—179,2025
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024136

基于阴影集的三支核均值漂移聚类算法

马云洁[1],万仁霞[1,2*],岳晓冬[3]
(1.北方民族大学 数学与信息科学学院,宁夏 银川 750021;
2.宁夏智能信息与大数据处理重点实验室,宁夏 银川 750021;
3.上海大学 计算机工程与科学学院,上海 200444)

摘 要:均值漂移属于硬划分的聚类算法,在处理不确定性数据时可能导致决策风险的提高和聚类精度的降低等问题。为此,本文引入阴影集理论来处理三支聚类的数据对象分类问题,提出了一种基于阴影集的三支核均值漂移聚类算法。算法采用类归属概率来刻画阴影集的隶属度概念。通过优化算法来获得阴影集划分的最优阈值,有效减少了人为干预带来的不确定性。最后基于最优阈值,形成了以阴影集隶属为依据的三支聚类。在2个人工数据集和8个UCI公共数据集上算法进行测试。相较于均值漂移算法、带宽自适应均值漂移算法(Adaptive Bandwidth Mean Shift Algorithm,ABMS)以及核均值漂移算法(Kernel Mean Shift Algorithm,KMS),本文所提出的基于阴影集的三支核均值漂移聚类算法(Three-way Kernel Mean Shift Algorithm Based on Shadow Sets,TKMSSS)不仅可以对数据进行有效划分,而且可以很好地刻画类簇的边界域,在戴维森堡丁指数、轮廓系数、准确率、调整兰德系数、同质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TKMSSS综合聚类性能优于比较算法。

关键词:阴影集;三支聚类;类归属概率;优化算法;类簇
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2025)01-0169-11

Three-way Kernel Mean Shift Algorithm Based on Shadow Sets

MA Yunjie[1], WAN Renxia[1,2*], YUE Xiaodong[3]
(1. *School of Mathematics and Information Science, North Minzu University, Yinchuan* 750021*, China;*
2. *Ningxia Key Laboratory of Intelligent Information and Data Processing, Yinchuan* 750021*, China;*
3. *School of Computer Engineering and Science, Shanghai University, Shanghai* 200444*, China*)

Abstract: Mean shift is a hard clustering algorithm. When dealing with uncertain data, it may lead to increased decision-making risks and reduced clustering accuracy. This paper introduces the shadow sets theory to address the data object classification problem in three-way clustering, and proposes a three-way mean shift clustering algorithm based on shadow sets. The proposed algorithm uses class belonging probability to represent the membership degree in the shadow sets. An optimization algorithm is employed to determine the optimal threshold for dividing the shadow sets, which effectively reduces the uncertainty caused by human intervention. Subsequently, a three-way clustering approach based on shadow sets membership is developed. The proposed algorithm is evaluated on 2 artificial datasets and 8 UCI public datasets. Compared to mean shift algorithm, adaptive bandwidth mean shift algorithm (ABMS) and kernel mean shift algorithm (KMS), the proposed algorithm TKMSSS can not only effectively divide the data, but also well describe the boundary region of the cluster. In terms of clustering evaluation indexes such as Divers-boding index, silhouette co-

收稿日期:2024-06-13;**接受日期**:2024-10-24
基金项目:国家自然科学基金(62066001);宁夏科技领军人才项目(2022GKLRLX08);宁夏自然科学基金(2021AAC03203)
作者简介:马云洁(1998—),女,山西晋城人,硕士研究生,研究方向为数据挖掘、三支决策和大数据分析。E-mail:myj436063@163.com
* **通信作者**:万仁霞(WAN Renxia),E-mail:wanrx1022@nmu.edu.cn
引文格式:马云洁,万仁霞,岳晓冬.基于阴影集的三支核均值漂移聚类算法[J].山西大学学报(自然科学版),2025,48(1):169-179. DOI:10.13451/j.sxu.ns.2024136.

efficient, accuracy, adjusted index coefficient and homogeneity, TKMSSS achieves better or similar results to the optimal algorithm, which indicates that the comprehensive clustering performance of TKMSSS is better than that of the comparison algorithms.
Key words: shadow sets; three-way clustering; class membership probability; optimization algorithm; cluster

0 引言

近年来,聚类分析作为数据挖掘中的重要技术,广泛应用于市场细分、社交网络分析等多个领域。其中,均值漂移(Mean Shift,MS)算法因其无参数特性和在多种应用场景中的良好聚类效果,逐渐引起了研究者的广泛关注。该算法最早由Fukunaga等[1]于1975年提出,核心思想是通过跟踪密度增大的最快速度方向来寻找每个样本,以找到密度最高的地方,即均值漂移方向。密度较高的区域被认为对应着分布的最大值,所有收敛到同一个局部最大值的样本点都被视为一个聚类的成员。随着研究的深入,相关学者对该算法进行了多方面的改进与扩展。Cheng[2]在1995年对此进行了改进,引入了一种新的内核函数,并加入了权重因子,使每个样本的重要程度有所区别,扩大了MS的应用范围。魏颖等[3]将Hessian矩阵滤波信息引入均值漂移聚类特征空间,使得特征空间差异变大,有利于聚类分割。郝茜茜等[4]利用半监督学习的方法对MS进行了扩展,利用已知的先验信息来限制聚类过程从而优化聚类数目。陈立伟等[5]使用光谱角距离作为MS聚类算法的相似性准则,减少了预估类别数带来的误差,提升了聚类结果的准确性。向俊伟等[6]设计了一种融合主成分分析(Principal Component Analysis,PCA)降维和均值漂移聚类的协同过滤推荐算法(Ollaborative Filtering Recommendation Algorithm Combining PCA Dimension Reduction and Mean Shift Clustering,PMCF),以缓解评分矩阵稀疏问题并提高近邻搜索效率,实验结果显示该算法在推荐准确性和时间效率方面均有显著提升。温柳英等[7]提出的基于覆盖树的自适应均值漂移聚类算法(MeanShift Based on Cover-Tree,MSCT),通过结合覆盖树数据集自适应生成带宽参数,解决了传统均值漂移算法依赖主观带宽选择和处理密度变化大数据集时精度问题。

尽管上述研究在均值漂移算法的改进与应用上取得了一定进展,但其在处理边界点和不确定性较大的数据时仍然表现出一定的局限性。这是由于均值漂移算法在聚类过程中通常采用硬聚类方式,即将每个数据点明确划分到某个类中或完全排除在外。对于存在边界模糊或数据不确定性的情况,这种严格的划分方式往往会导致决策风险增加,并降低聚类的整体准确性。为了解决这一问题,近年来的研究逐渐将多种决策理论与聚类算法相结合,以提高其处理不确定性数据的能力。其中,三支决策理论[8]由于其灵活处理边界数据的优势,受到广泛关注。Yu等[9-11]将三支决策理论引入聚类分析,提出了三支聚类算法,显著增强了算法在不确定性环境下的表现。然而,三支聚类方法在阈值选择上依然存在主观性,可能导致聚类结果的稳定性下降,并增加人为干预的风险。

在此背景下,为解决人为干预带来的不确定性,本文基于均值漂移算法的特性,提出一种基于阴影集的三支核均值漂移聚类算法(Three-way Kernel Mean Shift Algorithm Based on Shadow Sets,TKMSSS)。该算法旨在通过引入阴影集来解决聚类结果中边缘域对象的归属问题,并通过将数据属于各类的可能性(概率)作为评价函数,灵活划分数据集。

1 相关理论

1.1 核均值漂移算法

(1) 核函数

均值漂移是一种基于密度的无参数聚类算法,它试图通过追踪数据集中每个点处的密度梯度来发现密集区域。具体来说,对于给定的数据集$X=\{x_1,x_2,\cdots,x_n\}$,它会为每个数据点设定一个特定的核函数,然后将所有点的核函数相加,得到数据集的整体的核密度估计,点x处的核密度估计函数[12]为:

$$f_k(x)=\frac{1}{n_h}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right), \tag{1}$$

其中h为带宽参数,n为样本总数,$K(x)$为核函数,x_i表示第i个数据的样本值。

常用的核函数为高斯核函数:

$$K(u)=\mathrm{e}^{-\frac{u^2}{2}}。 \tag{2}$$

核函数满足以下两个约束条件:

$$\int K(u)\mathrm{d}u=1, \tag{3}$$

$$K(u)=K(|u|)。 \tag{4}$$

(2)均值漂移量[12]

均值漂移量$\boldsymbol{M}_h(x)$:

$$\boldsymbol{M}_h(x)=\frac{1}{n_x}\sum_{x_i\in S_h(x)}(x_i-x)=\frac{1}{n_x}\sum_{x_i\in S_h(x)}x_i-x。 \tag{5}$$

引入核函数后的均值漂移量为:

$$\boldsymbol{M}_h(x)=\frac{\sum_{x_i\in S_h(x)}x_iK\left(\left\|\frac{x-x_i}{h}\right\|\right)^2\omega(x_i)}{\sum_{x_i\in S_h(x)}K\left(\left\|\frac{x-x_i}{h}\right\|\right)^2\omega(x_i)}-x, \tag{6}$$

其中$S_h(x)$是一个以x为中心点,半径为h的高维球区域,n_x表示在这个球中有n_x个点,$\omega(x_i)$为样本权重。

$\omega(x_i)$的计算采用平均核函数$K_h(x,x_j)$:

$$K_h(x,x_j)=\begin{cases}1 & \text{if } \|x-x_j\|\leqslant h,\\ 0 & \text{otherwise}。\end{cases} \tag{7}$$

当x_j在x的邻域内时,权重为1,否则权重为0;$\boldsymbol{M}_h(x)$漂移量的方向总是指向密度梯度增加的方向,其收敛点即为局部密度极大值点。

(3)更新簇中心

由此更新中心坐标:

$$x'=\boldsymbol{M}_h(x)+x。 \tag{8}$$

然后中心点将根据公式(5)的密度梯度方向移动到一个局部最大值,通过这种不断重复的移动,使中心点逐步逼近到最佳位置,直至算法收敛。这意味着它们已经到达了其局部最大值。

均值漂移聚类算法[13]的具体步骤如下:

输入:样本集$D=\{x_1,x_2,\cdots,x_n\}$,带宽参数h,阈值ε;

Step1:随机选择数据集中一个点x作为初始中心点(称为搜索点);

Step2:对于搜索点x,在带宽h范围内找到其邻域内的所有数据点,根据公式(1)计算这些数据点在搜索点x处的核密度估计$f_K(x)$;

Step3:计算搜索点的密度梯度上升方向,即寻找密度增大的方向,根据公式(6)计算均值漂移量$\boldsymbol{M}_h(x)$;

Step4:更新中心点位置,使其沿着密度梯度上升方向移动到新的位置x',检查搜索点的移动距离。如果移动距离$\|x'-x\|$小于预设的阈值ε,则认为搜索点已收敛;否则,将搜索点更新为x',返回Step2,继续迭代;

Step5:对于数据集中的每一个数据点重复执行Step1至Step4的迭代过程,直至所有点都收敛到各自的密度峰值;

Step6:在所有搜索点收敛后,通过计算数据点与所有收敛点(即聚类中心)的距离,应用合并策略,将收敛到相同位置的点归为同一簇。

输出:聚类标签labels,聚类中心centers。

对于Step2中“其邻域内”是指搜索点的邻域。涉及的邻域通常使用带宽参数h决定搜索点周围的邻域范围。一种常见的做法是将h设置为数据集的标准差的一部分,例如标准差的十分之一,这样可以保证搜索点的邻域范围能够覆盖数据的主要分布区域。另一种方法是通过交叉验证来确定最佳的h值。对于Step4中的阈值ε,收敛阈值的选择通常基于数据的尺度和分布,常见方法包括将阈值设为数据范围的一小部分(如0.01%到1%)或数据标准差的一小部分(如1%),通过初步实验和多次验证,观察不同阈值下的收敛速度和聚类效果,逐步调整以在合理时间内获得高质量的聚类结果。

1.2 三支聚类

为解决重叠数据集的聚类归属问题,Yu等[9]将三支决策[14-16]的思想引入聚类分析,提出了三支聚类理论。假设$U=\{x_1,x_2,\cdots,x_n\}$是非空有限集合,在三支聚类过程中,用$Co(C)$、$Fr(C)$、$Tr(C)$分别表示核心域、边缘域和琐碎域,$Co(C)$中的对象一定属于类C,$Fr(C)$中的对象可能属于类C,$Tr(C)$中的对象一定不属于类C。三支聚类结果表示为:

$$(Co(C_1),Fr(C_1)),(Co(C_2),Fr(C_2)),\cdots,(Co(C_k),Fr(C_k))$$

且满足如下性质:

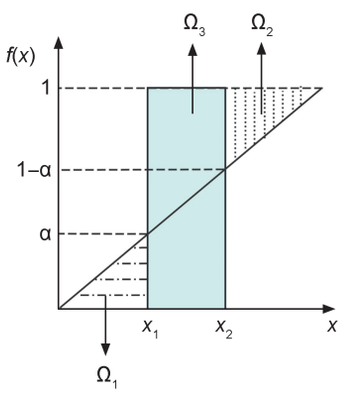
①$Co(C_i)\neq\varnothing$;

②$\bigcup_{i=1}^{k}(Co(C_i)\cup Fr(C_i))=U$;

③$Co(C_i)\cap Co(C_j)=\varnothing, i\neq j$。

1.3 阴影集

阴影集[17-18]是在模糊集的基础之上演化而来,其通过模糊隶属度来处理类簇边界模糊的现象。阴影集有三种量化级别,分别是:1、0、[0,1],即将模糊集转换成阴影集的形式,依次对应隶属度为1、0、隶属度不确定的区间[19],阴影集结构如图1所示。



注:Ω_1表示对象x隶属度的减少区域,Ω_2表示对象x隶属度的增大区域,Ω_3表示阴影区域;α是阴影集的划分阈值。

图1 阴影集结构

Fig. 1 Structure of shadow sets

在论域U中,对于模糊概念A,假设阈值为α_0,且$0\leqslant\alpha_0<0.5$,阴影集将论域U中所有对象根据隶属度$\mu_A(x)$映射到集合$\{0,[0.1],1\}$中[20],有

$$f(\mu_A(x))=\begin{cases}0 & \mu_A(x)\leqslant\alpha_0,\\ [0,1] & \alpha_0<\mu_A(x)<1-\alpha_0,\\ 1 & \mu_A(x)\geqslant1-\alpha_0。\end{cases}\quad(9)$$

其中隶属度$\mu_A(x)$表示对象x隶属于某一模糊概念A的程度,f为模糊映射。

阴影集是连接模糊集与粗糙集的桥梁,其基于平衡总的不确定性构造[21]。若$\mu_A(x)\geqslant1-\alpha_0$,则将对象$x$的隶属度$\mu_A(x)$设法提升到1;若$\mu_A(x)\leqslant\alpha_0$,则将对象$x$的隶属度$\mu_A(x)$法降低到0,若$\alpha_0<\mu_A(x)<1-\alpha_0$,则将对象$x$划分至阴影区域[22]。虽然以上操作会减少阴影区域,但会改变不确定性存在的规律,为使总的变化量整体平衡(即不确定性平衡),使其达到完全平衡,则要满足以下公式:

$$\Omega_1+\Omega_2=\Omega_3。\quad(10)$$

通过最小化目标函数求解最优阈值α_0,目标函数为:

$$V_{\alpha_0}=|\Omega_1+\Omega_2-\Omega_3|。\quad(11)$$

对于连续的隶属度函数,α_0应使得公式(11)取得最小值:

$$V=\left|\int_{-\infty}^{\alpha_0}f(x)\mathrm{d}x+\int_{1-\alpha_0}^{+\infty}(1-f(x))\mathrm{d}x-\int_{\alpha_0}^{1-\alpha_0}f(x)\mathrm{d}x\right|。\quad(12)$$

对于离散的隶属度函数,α_0应使得公式(12)取得最小值:

$$V=\left|\sum_{i:f(x_i)<\alpha_0}f(x_i)+\sum_{i:f(x_i)>1-\alpha_0}(1-f(x_i))-N_{\mathrm{card}}\{x_i\in X|\alpha_0<f(x)<1-\alpha_0\}\right|。\quad(13)$$

式中$|\bullet|$代表绝对运算;$N_{\mathrm{card}}\{\bullet\}$代表集合中对象的数量。

2 基于阴影集的三支核均值漂移聚类算法

2.1 算法思想

相较于传统最近邻技术划分数据对象归类的方法,数据对象隶属各类的程度更能真实反映数据点和类簇间的关系,因此,本文参考文献[23-24]的方法,将每个数据点属于各类的概率视为评价来构建三支聚类的划分依据,即将均值漂移的聚类问题通过阴影集理论转化为基于概率评价的三支聚类问题,其中,每个数据点属于各类的概率计算如下:

$$p_{ij}=\frac{\mathrm{e}^{-\frac{1}{2}\left\|\frac{x_i-x_i}{h}\right\|^2}}{\sum_{m=1}^{k}\mathrm{e}^{-\frac{1}{2}\left\|\frac{x-x_m}{h}\right\|^2}}。\quad(14)$$

这样,转化后的均值漂移对数据对象的三支划分将通过阴影集映射到集合$\{0,[0.1],1\}$中,1表示对象属于该类簇,对应于三支聚类的核心域;$[0,1]$于该类簇,对应于三支聚类的边缘域;隶属度为0表示不属于该类簇,对应于三支聚类的琐碎域。

2.2 基于阴影集的三支核均值漂移聚类算法

输入:样本集$D=\{x_1,x_2,\cdots,x_n\}$,带宽参数h,收敛阈值ε;

Step 1:在未被分类的数据点中随机选择一个点作为中心点;

Step 2:构建核函数,计算每个点的均值漂移向量$\boldsymbol{M}_h(x)$;

Step 3:更新中心点位置,直至在移动方向满足漂移向量的模小于收敛阈值ε,否则返回Step 1;

Step 4:计算概率值p_{ij}为模糊集;

Step 5:确定α_0值:$V_{\alpha_0}=|\Omega_1+\Omega_2-\Omega_3|$,结合优化算法寻找$\alpha_0$的全局最优解,使其满足$\alpha_0=\arg\min(V_i)$且$0\leqslant\alpha_0<0.5$,其中:

$$V_i=\left|\sum_{j:p_{im}<\alpha_0}p_{im}+\sum_{j:p_{im}>1-\alpha_0}(1-p_{im})-N_{\text{card}}\left\{x_i\in X\middle|\alpha_0<p_{im}<1-\alpha_0\right\}\right|,\quad(15)$$

其中p_{im}为对象x_i到所有类的最大概率值;

Step 6:对于每一类C_i取$x_j\in C_i$,根据α_0进行划分聚类:若$p_{ij}\geqslant1-\alpha_0$,则将$x_i$划分到$Co(i)$,若$1-\alpha_0>p_{ij}>\alpha_0$,则将$x_i$划分到$Fr(i)$,若$p_{ij}<\alpha_0$,则将$x_i$划分到$Tr(i)$;

Step 7:输出划分结果

$$S=\left\{\left[Co(1),Fr(1)\right],\left[Co(2),Fr(2)\right],\cdots,\left[Co(k),Fr(k)\right]\right\}。$$

TKMSSS算法基于阴影集构造了一个目标函数,在Step5中,我们使用狮群优化算法[25-26]来求解α_0的全局最优解,以便更准确地对边界区域的对象进行划分,通过这种方式,TKMSSS算法可以提供比一般方法更有效的阈值选取,从而更准确地划分数据的边界区域,增强聚类的性能。

其中,狮群优化算法(Lion Optimization Algorithm,LOA)是一种基于群体智能的优化算法,灵感来自狮群的社会组织结构和捕猎行为。狮群由几只雄狮、若干雌狮和幼狮组成,雄狮负责保护领地,雌狮负责捕猎并共享食物,狮群优化算法通过模拟这一过程来实现全局搜索和局部搜索。具体来说,算法将候选解视为狮子,分为雄狮和雌狮两种角色。雄狮主要进行领地保护,避免陷入局部最优;雌狮则在固定区域内进行局部搜索,捕获最优解。通过雄狮的迁徙和雌狮的捕猎行为,算法不断更新种群中的解,从而逐步逼近全局最优。数学上,LOA的搜索过程通过一系列迭代和更新公式来实现,每一代通过评估适应度函数,选择优良个体进行繁殖和淘汰,以达到优化目标。

2.3 复杂度分析

2.3.1 空间复杂度分析

每个数据对象到所有数据对象的空间复杂度,即存储核矩阵为$\boldsymbol{O}(N^2)$,其中N为数据对象总个数,存储每个数据V_i,α_0的空间复杂度为$O(N+C+1)$,其中1代表存储α_0,C表示每次迭代时中间变量$\min V_i$,$\min\alpha_0$,$preV_i$(上一次迭代生成的V_i)的存储空间。因此,TKMSSS算法总体空间复杂度为$\boldsymbol{O}(N^2)$,同原始均值漂移算法一致。

2.3.2 时间复杂度分析

(1)创建高斯核矩阵的复杂度为$\boldsymbol{O}(N^2)$;

(2)均值漂移的时间复杂度主要由带宽和样本数量决定,通常为$O(TNd)$,其中T是迭代次数,d是样本的特征数量;

(3)计算每个数据点的V_i值、其时间复杂度为$\boldsymbol{O}(N^2)$;

(4)算法中的阈值α_0是通过优化算法来取得最优值的,本文采用狮群算法来搜寻α_0,其时间复杂度取为$\boldsymbol{O}(\max_iters\cdot num_lions)$,其中$\max_iters$是迭代次数,$num_lions$是每次搜索的数量;

考虑到,一般而言N的取值大于$\max_iters$、num_lions的取值,因而,本文算法TKMSSS的总体时间复杂为$\boldsymbol{O}(N^2)$。

3 实验与结果分析

3.1 数据集与实验环境

本文实验环境为python 3.10.7.操作系统为Windows 11,处理器为AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz。

本文使用8个UCI数据集和2个人工数据集(D1、D2),详见表1。

3.2 评价指标

本文选取戴维森堡丁指数、轮廓系数、准确率、调整兰德系数和同质性5个指标来衡量聚

表 1 实验数据集

Table 1 Experimental datasets

数据集	样本数	属性个数	类别
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Liver	345	6	2
Pima	768	8	2
Vehicle	846	18	4
Vote	435	16	2
Wpbc	198	34	2
D1	600	2	3
D2	2 000	2	5

类的结果。

① 戴维森堡丁指数[27](Davies-Bouldin Index,DBI)

$$I_{\mathrm{DBI}}=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\frac{s_i+s_j}{d_{ij}},\tag{16}$$

其中k表示簇的个数,s_i表示第i个簇内样本到簇中心的平均距离;d_{ij}表示第i个簇和第j个簇中心之间的距离。

DBI通过计算簇内距离和簇间距离的比值来度量聚类的紧密度和分离度。DBI指数越小,表示聚类质量越好,簇内距离越小,簇间距离越大。

② 轮廓系数[28](Silhouette Coefficient,SC)

$$I_{\mathrm{SC}}=\frac{b-a}{\max(a,b)},\tag{17}$$

其中a表示每个样本到同一类别中其他样本的平均距离(即类别内平均距离),b表示每个样本到其他类别的所有样本的平均距离(即类别间平均距离)。

I_{SC}基于样本点与所属簇内和最近邻簇之间的距离,相似性和紧密度等因素计算,轮廓系数越小,说明样本越相似,反之则越不相似。

③ 准确率[29](Accuracy,ACC)

$$I_{\mathrm{ACC}}=\frac{1}{N}\sum_{i=1}^{k}C_i,\tag{18}$$

其中N表示全部样本的数目,C_i表示样本被划分到对应聚类i正确的个数;k为聚类数。

I_{ACC}越大,聚类效果越好。

④ 调整兰德系数[30](Adjusted Rand Index,ARI)

$$I_{\mathrm{ARI}}=\frac{2(ad-bc)}{(b+d)(a+c)+(c+d)(a+b)},\tag{19}$$

其中a表示真实在同一类,预测也在同一类的样本数;b表示真实在同一类、预测在不同类的样本数;c表示真实在不同类、预测在同一类的样本数;d表示真实在不同类、预测也在不同类的样本数。

I_{ARI}的取值范围为$[-1,1]$,值越大表示两个聚类结果越相似。

⑤ 同质性[31](Homogeneity,HOM)

$$I_{\mathrm{Hom}}=1-\frac{H(C|K)}{H(C)},\tag{20}$$

其中$H(C)$是类别C的熵,$H(C|K)$是给定聚类结果K的条件下类别C的条件熵。熵在这里用于衡量类别的不确定性或混乱程度。

同质性衡量的是每个聚类簇中只包含单个真实类别的成员的程度。同质性得分越高,说明聚类结果越倾向于将相同类别的样本聚集在一起,聚类效果越好。

3.3 实验阈值选择

本文采用狮群优化算法计算获得各数据集的最优阈值,计算时设置迭代次数为100,初始狮群为10,由于$0\leqslant\alpha_0<0.5$,因此设置寻优范围为$(0,0.5)$,经过多次迭代搜索,我们得到了每个数据集的最优阈值。这些阈值是在给定的寻优范围内求解得到的全局最优解,具体阈值选取结果如下表2。

表 2 各数据集最优阈值

Table 2 Optimal thresholds for each dataset

数据集	阈值α_0	$1-\alpha_0$
Iris	0.2	0.8
Wine	0.427	0.573
Seeds	0.469	0.531
Liver	0.292	0.708
Pima	0.369	0.631
Vehicle	0.366	0.634
Vote	0.301	0.616
Wpbc	0.384	0.699
D1	0.466	0.534
D2	0.219	0.781

注:阈值α_0和$1-\alpha_0$用于阴影集的划分,代表不同数据集在聚类过程中确定核心域和边缘域的最优边界。

3.4 实验结果分析

3.4.1 基于人工数据集实验

为了直观展示文中提出的TKMSSS算法与其他算法聚类结果的不同,本节选取mean shift

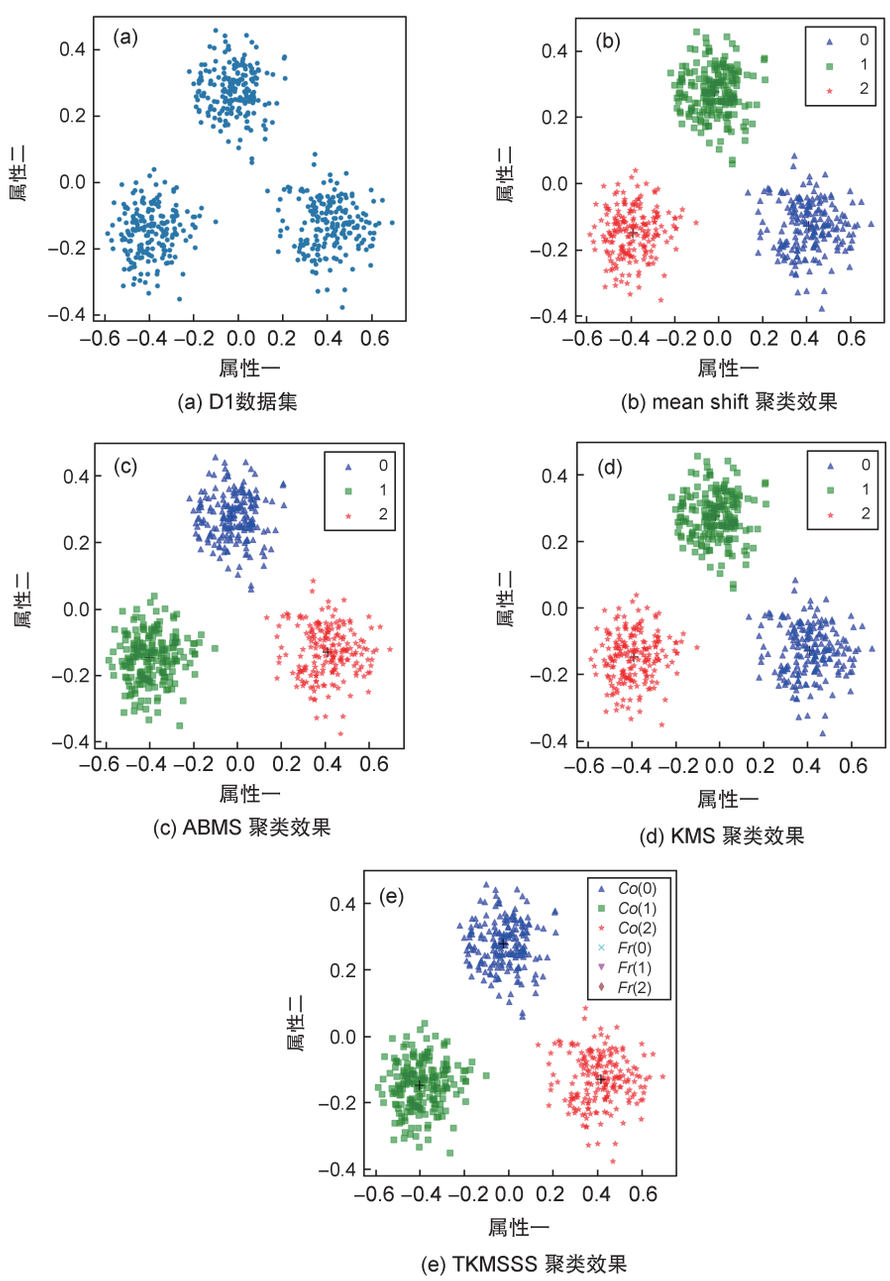
算法、带宽自适应均值漂移算法(Adaptive Bandwidth Mean Shift Algorithm,ABMS)[32]以及核均值漂移算法(Kernel Mean Shift Algorithm,KMS)[23]和本文提到的算法相对比,给出在二维人工数据集D1、D2上的实验结果展示。聚类结果如图2—图3。

图2(a)是未分类的D1数据集,观察几种算法在D1数据集上的聚类结果,可以看出数据比较明显地分成了三类,几种算法对于类的划分是一致的,说明本算法可以得到正确的聚类结果。

图3(a)是未分类的D2数据集,观察几种算法在D2数据集上的聚类结果,可以看出有两个类之间存在重叠,本文算法找到了类的重叠部分,并将其标记出来,使得类间对象的关系得到了更真实地展现。

3.4.2 UCI数据集结果分析

为了进一步验证算法性能,本文选取了UCI数据集中的通过8组数据进行试验,并采



(a) D1数据集 (b) mean shift 聚类效果

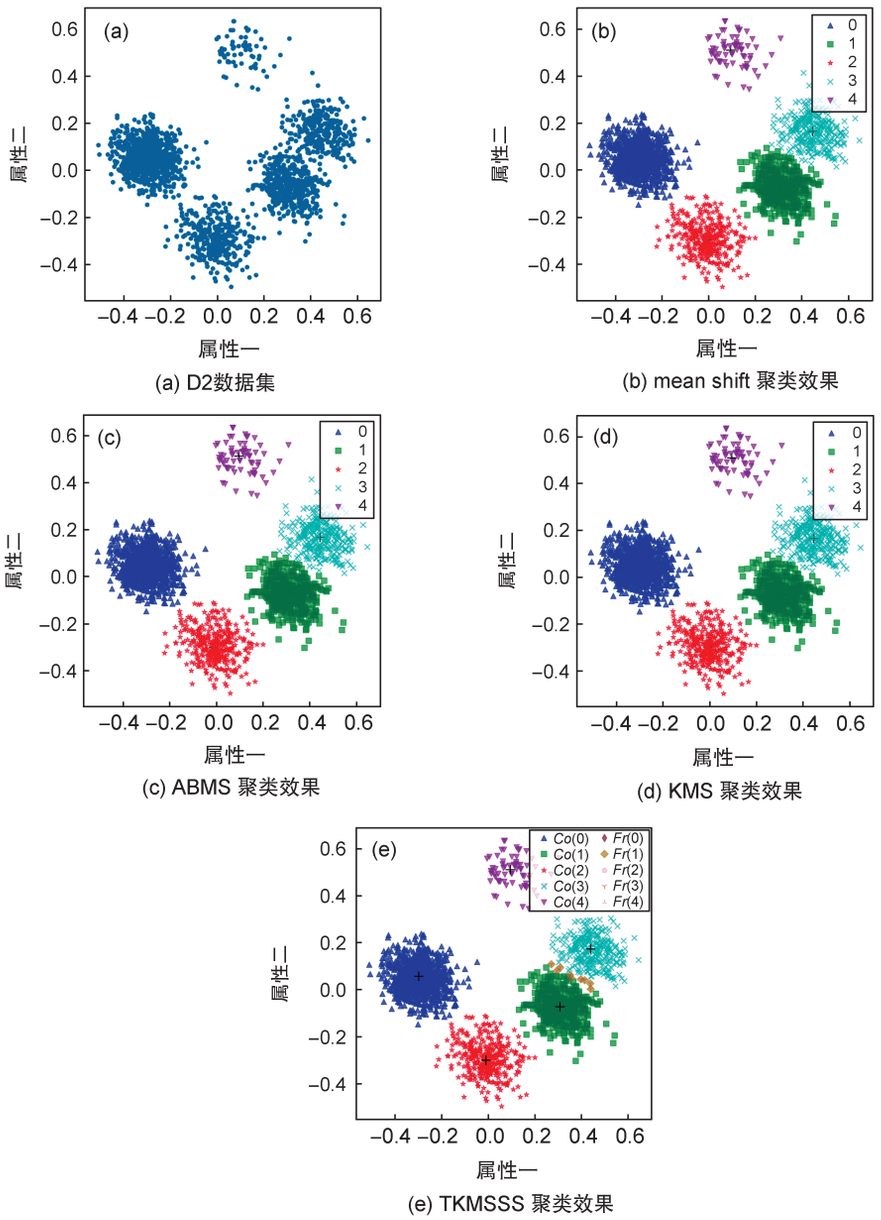
(c) ABMS 聚类效果 (d) KMS 聚类效果

(e) TKMSSS 聚类效果

注:0、1、2分别表示聚类结果中的类簇编号,$Co(i)$表示三支聚类的核心域,$Fr(i)$表示三支聚类的边缘域。

图2 数据集D1上的实验结果

Fig. 2 Experimental results on D1 dataset



(a) D2数据集 (b) mean shift 聚类效果

(c) ABMS 聚类效果 (d) KMS 聚类效果

(e) TKMSSS 聚类效果

注:1、2、3、4、5分别表示聚类结果中的类簇编号,$Co(i)$表示三支聚类的核心域,$Fr(i)$表示三支聚类的边缘域。

图 3 数据集D2上的实验结果

Fig. 3 Experimental results on D2 dataset

用DBI、SC、ACC、ARI和HOM5个聚类评价指标来进行对比分析。实验结果如表3。

通过表3各评价指标的对比,我们可以观察到TKMSSS算法的各项指标在Iris、Wine、Liver、Pima、Vehicle和Vote均表现最优。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,如在seeds数据集中,TKMSSS的同质性略低于取得最优的mean shift和ABMS;在Wpbc数据集上,TKMSSS的轮廓系数略低于KMS。这是因为TKMSSS沿用核函数方法对mean shift上的改进,继承了KMS的优势,同时TKMSSS算法最优阈值寻优的策略有效避免引入阴影集后人工选取阈值的不确定性问题,而边缘域的刻画不仅展现了数据对象与类间的关系细节,也为进一步分类决策提供了重要依据。

4 结论

本文通过引入阴影集理论和三支聚类的思想来改进均值漂移算法,提出了一种基于阴影集的三支核均值漂移聚类算法。算法结合核函

表 3 UCI数据集评价指标结果

Table 3 Results of evaluation indicator on UCI dataset

数据集	评价指标	mean shift	ABMS	KMS	TKMSSS	数据集	评价指标	mean shift	ABMS	KMS	TKMSSS
Iris	DBI	0.637	0.638	0.652	**0.622**	Pima	DBI	3.021	3.021	3.052	**2.882**
	SC	0.511	0.511	0.507	**0.535**		SC	0.125	0.125	0.124	**0.126**
	ACC	0.860	0.867	0.860	**0.881**		ACC	0.879	0.879	0.882	**0.904**
	ARI	0.041	0.045	0.041	**0.063**		ARI	0.120	0.120	0.120	**0.134**
	HOM	0.150	0.185	0.197	**0.205**		HOM	0.019	0.041	0.019	**0.159**
Wine	DBI	1.296	1.372	1.471	**1.180**	Vehicle	DBI	1.522	1.510	1.581	**1.417**
	SC	0.194	0.198	0.202	**0.213**		SC	0.204	0.197	0.192	**0.204**
	ACC	0.887	0.888	0.893	**0.901**		ACC	0.764	0.798	0.748	**0.837**
	ARI	0.001	0.001	0.001	**0.006**		ARI	0.063	0.061	0.061	**0.167**
	HOM	0.094	0.094	0.105	**0.258**		HOM	0.198	0.186	0.196	**0.201**
Seeds	DBI	1.686	1.030	1.041	**0.974**	Vote	DBI	1.419	1.706	1.418	**1.402**
	SC	0.167	0.374	0.347	**0.421**		SC	0.296	0.305	0.296	**0.342**
	ACC	0.995	0.990	0.995	**1.000**		ACC	0.710	0.683	0.713	**0.988**
	ARI	0.003	0.003	0.003	**0.154**		ARI	0.179	0.191	0.038	**0.226**
	HOM	**0.207**	**0.207**	0.027	0.200		HOM	0.155	0.171	0.047	**0.174**
Liver	DBI	0.882	1.279	0.955	**0.866**	Wpbc	DBI	1.518	1.904	1.518	**1.507**
	SC	0.624	0.360	0.614	**0.636**		SC	0.268	0.104	**0.268**	0.257
	ACC	0.951	0.817	0.939	**0.959**		ACC	0.944	0.712	0.944	**0.968**
	ARI	0.004	0.003	0.004	**0.006**		ARI	−0.001	−0.001	−0.001	**0.180**
	HOM	**0.019**	0.041	**0.019**	**0.019**		HOM	**0.032**	**0.032**	**0.032**	**0.032**

数方法,采用数据对象的类归属概率来刻画阴影集的隶属度,构建了以阴影集隶属为依据的三支聚类。而通过优化算法获得阴影集划分的最优阈值的策略则有效减少了人为参数选择带来的不确定性,在不同数据集上的对比实验结果表明本文所提出的算法能够有效改进均值漂移算法的聚类性能。

参考文献:

[1] FUKUNAGA K, HOSTETLER L. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition[J]. *IEEE Trans Inf Theory*, 1975, **21**(1): 32−40. DOI: 10.1109/TIT.1975.1055330.

[2] CHENG Y Z. Mean Shift, Mode Seeking, and Clustering [J]. *IEEE Trans Pattern Anal Mach Intell*, 1995, **17**(8): 790−799. DOI: 10.1109/34.400568.

[3] 魏颖,徐陆,李翔, 等. 结合多尺度圆形滤波与MS聚类的疑似结节分割[J]. 仪器仪表学报, 2016, **37**(1): 192−199. DOI: 10.19650/j.cnki.cjsi.2016.01.026.
WEI Y, XU L, LI X, *et al*. Segmentation for Suspected Nodules by Multi-scale Circular Filtering Combined with MS Clustering[J]. *Chin J Sci Instrum*, 2016, **37**(1): 192−199. DOI: 10.19650/j.cnki.cjsi.2016.01.026.

[4] 郝茜茜, 周亚同, 任婷婷. 基于半监督核均值漂移聚类的地震相识别研究[J]. 河北工业大学学报, 2017, **46**(6): 6−12. DOI: 10.14081/j.cnki.hgdxb.2017.06.002.
HAO X X, ZHOU Y T, REN T T. Research on Seismic Facies Identification Based on Semi-supervised Kernel Mean Shift Clustering[J]. *J Hebei Univ Technol*, 2017, **46**(6): 6−12. DOI: 10.14081/j.cnki.hgdxb.2017.06.002.

[5] 陈立伟, 邱艳芳, 朱海峰, 等. 基于均值漂移聚类的端元束提取[J]. 应用科技, 2020, **47**(1): 21−30. DOI: 10.11991/yykj.201909015.
CHEN L W, QIU Y F, ZHU H F, *et al*. Endmember Bundle Extraction Based on Mean Shift Clustering[J]. *Appl Sci Technol*, 2020, **47**(1): 21−30. DOI: 10.11991/yykj.201909015.

[6] 向俊伟, 李玲娟. 融合PCA降维和均值漂移聚类的协同过滤推荐算法[J]. 南京邮电大学学报(自然科学版), 2023, **43**(3): 90−95. DOI: 10.14132/j.cnki.1673-5439.2023.03.010.
XIANG J W, LI L J. A Collaborative Filtering Recommendation Algorithm Combining PCA Dimension Reduction and Mean Shift Clustering[J]. *J Nanjing Univ Posts Telecommun Nat Sci Ed*, 2023, **43**(3): 90−95. DOI: 10.14132/j.cnki.1673-5439.2023.03.010.

[7] 温柳英, 庞柯. 基于覆盖树的自适应均值漂移聚类算法[J]. 计算机工程与设计, 2024, **45**(2): 452−458. DOI: 10.16208/j.issn1000-7024.2024.02.017.
WEN L Y, PANG K. Adaptive Mean Shift Clustering

Based on Cover-tree[J]. *Comput Eng Des*, 2024, **45**(2): 452−458. DOI: 10.16208/j.issn1000-7024.2024.02.017.

[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] YU H, ZHANG C, WANG G Y. A Tree-based Incremental Overlapping Clustering Method Using the Three-way Decision Theory[J]. *Knowl Based Syst*, 2016, **91**: 189−203. DOI: 10.1016/j.knosys.2015.05.028.

[10] YU H. A Framework of Three-way Cluster Analysis[C]. International Joint Conference on Rough Sets. Cham: Springer. 2017, 300−312. DOI: 10.1007/978-3-319-60840-2_22.

[11] YU H, CHANG Z H, WANG G Y, *et al*. An Efficient Three-way Clustering Algorithm Based on Gravitational Search[J]. *Int J Mach Learn Cybern*, 2020, **11**(5): 1003−1016. DOI: 10.1007/s13042-019-00988-5.

[12] 聂倩, 赵艳福. 结合IR-MAD与均值漂移算法的密集城区遥感影像变化检测[J]. 测绘通报, 2020(6): 57−62.
NIE Q, ZHAO Y F. Change Detection of Remote Sensing Imagein Dense Urban Area Based on IR-MAD and Mean Shift Algorithm[J]. *Bull Surv Mapp*, 2020(6): 57−62.

[13] ZHAO M Y, JHA A, LIU Q, *et al*. Faster Mean-shift: GPU-accelerated Clustering for Cosine Embedding-based Cell Segmentation and Tracking[J]. *Med Image Anal*, 2021, **71**: 102048. DOI: 10.1016/j.media.2021.102048.

[14] 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.

[15] 李京阳, 刘三民, 张匡燕. 基于三支决策的数据流主动学习分类研究[J]. 天津理工大学学报, 2023, **39**(3): 21−26. DOI: 10.3969/j.issn.1673-095X.2023.03.004.
LI J Y, LIU S M, ZHANG K Y. Research on Active Learning Classification of Data Stream Based on Three-way Decision[J]. *J Tianjin Univ Technol*, 2023, **39**(3): 21−26. DOI: 10.3969/j.issn.1673-095X.2023.03.004.

[16] 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.

[17] PEDRYCZ W. Shadowed Sets: Representing and Processing Fuzzy Sets[J]. *IEEE Trans Syst Man Cybern B Cybern*, 1998, **28**(1): 103−109. DOI: 10.1109/3477.658584.

[18] PEDRYCZ W. From Fuzzy Sets to Shadowed Sets: Interpretation and Computing[J]. *Int J Intell Syst*, 2009, **24**(1): 48−61. DOI: 10.1002/int.20323.

[19] 董雪, 万仁霞, 苗夺谦, 等. 基于阴影集的三支高斯混合聚类算法[J]. 广西大学学报(自然科学版), 2023, **48**(4): 958−971. DOI: 10.13624/j. cnki. issn. 1001-7445.2023.0958.
DONG X, WAN R X, MIAO D Q, *et al*. Three-way Gaussian Mixture Clustering Algorithm Based on Shadow Set[J]. *J Guangxi Univ Nat Sci Ed*, 2023, **48**(4): 958−971. DOI: 10.13624/j.cnki.issn.1001-7445.2023.0958.

[20] 雒僖, 范九伦, 于海燕, 等. 基于阴影集的截集式可能性C-均值聚类截集门限的选取[J]. 计算机科学, 2019, **46**(8): 249−254. DOI: 10.11896/j. issn. 1002-137X. 2019.08.041.
LUO X, FAN J L, YU H Y, *et al*. Selection of Cutset Threshold for Cutset-type Possibilistic C-means Clustering Based on Shadowed Set[J]. *Comput Sci*, 2019, **46**(8): 249−254. DOI: 10.11896/j. issn. 1002-137X. 2019.08.041.

[21] 王丽娜, 王建东, 李涛, 等. 集成粗糙集和阴影集的簇特征加权模糊聚类算法[J]. 系统工程与电子技术, 2013, **35**(8): 1769−1776. DOI: 10.3969/j. issn. 1001-506X.2013.08.31.
WANG L N, WANG J D, LI T, *et al*. Cluster's Feature Weighting Fuzzy Clustering Algorithm Integrating Rough Sets and Shadowed Sets[J]. *Syst Eng Electron*, 2013, **35**(8): 1769−1776. DOI: 10.3969/j. issn. 1001-506X.2013.08.31.

[22] 郭晋华, 苗夺谦, 周杰. 基于阴影集的粗糙聚类阈值选择[J]. 计算机科学, 2011, **38**(10): 209−210. DOI: 10.3969/j.issn.1002-137X.2011.10.048.
GUO J H, MIAO D Q, ZHOU J. Shadowed Sets Based Threshold Selection in Rough Clustering[J]. *Comput Sci*, 2011, **38**(10): 209−210. DOI: 10.3969/j. issn. 1002-137X.2011.10.048.

[23] HINTON G. Visualizing Data Using T-SNE[J]. *J Mach Learn Res*, 2008, **9**(2): 2579−2605.

[24] 鲜焱, 吕佳. 基于核均值漂移聚类的改进局部协同训练算法[J]. 重庆师范大学学报(自然科学版), 2020, **37**(4): 106−113. DOI: 10.11721/cqnuj20200408.
XIAN Y, (LÜ/LV/LU/LYU) J. Improved Partial Co-training Algorithm Based on Kernel Mean Shift[J]. *J Chongqing Norm Univ Nat Sci*, 2020, **37**(4): 106−113. DOI: 10.11721/cqnuj20200408.

[25] 万仁霞, 高艳龙. 基于粗糙集约简与狮群优化算法的机器人路径规划研究[J]. 郑州大学学报(理学版), 2022, **54**(2): 32−38. DOI: 10.13705/j. issn. 1671-6841.2021218.
WAN R X, GAO Y L. Robot Path Planning Based on Rough Set Reduction and Lion Swarm Optimization[J]. *J Zhengzhou Univ Nat Sci Ed*, 2022, **54**(2): 32−38. DOI: 10.13705/j.issn.1671-6841.2021218.

[26] 黄志锋, 刘媛华. 基于改进狮群算法的城市无人机低空路径规划[J]. 信息与控制, **2023**(6): 747−757+772. DOI: 10.13976/j.cnki.xk.2023.2372.
HUANG Z F, LIU Y H. Low Altitude Path Planning of Urban UAV Based on Improved Lion Swarm Optimization[J]. *Inf Contr*, **2023**(6): 747−757+772. DOI: 10.13976/j.cnki.xk.2023.2372.

[27] 凡嘉琛, 王平心, 杨习贝. 基于三支决策的密度敏感谱聚类[J]. 山东大学学报(理学版), 2023, **58**(1): 59−66. DOI: 10.6040/j.issn.1671-9352.0.2021.671.
FAN J C, WANG P X, YANG X B. Density-sensitive Spectral Clustering Based on Three-way Decision[J]. *J Shandong Univ Nat Sci Ed*, 2023, **58**(1): 59−66. DOI: 10.6040/j.issn.1671-9352.0.2021.671.

[28] 徐天杰, 王平心, 杨习贝. 基于人工蜂群的三支k-means聚类算法[J]. 计算机科学, 2023, **50**(6): 116−121. DOI: 10.11896/jsjkx.220800150.
XU T J, WANG P X, YANG X B. Three-way K-means Clustering Based on Artificial Bee Colony[J]. *Comput Sci*, 2023, **50**(6): 116−121. DOI: 10.11896/jsjkx.220800150.

[29] 陈玉洪, 张清华, 杨洁. 基于区间阴影集的密度峰值聚类算法[J]. 模式识别与人工智能, 2019, **32**(6): 531−544. DOI: 10.16451/j.cnki.issn1003-6059.201906006.
CHEN Y H, ZHANG Q H, YANG J. Density Peak Clustering Algorithm Based on Interval Shadowed Sets[J]. *Pattern Recognit Artif Intell*, 2019, **32**(6): 531−544. DOI: 10.16451/j.cnki.issn1003-6059.201906006.

[30] WALLACE D L. A Method for Comparing Two Hierarchical Clusterings: Comment[J]. *J Am Stat Assoc*, 1983, **78**(383): 569. DOI: 10.2307/2288118.

[31] 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.
LUO S W, WAN R X, MIAO D Q. Three-way Decision-based Density Peak Clustering Algorithm with Clustering Centers Preselection[J]. *J Shanxi Univ Nat Sci Ed*, 2024, **47**(1): 30−39. DOI: 10.13451/j.sxu.ns.2023140.

[32] 周芳芳, 樊晓平, 叶榛. 基于自适应带宽均值漂移聚类算法设计传递函数[J]. 信息与控制, 2007, **36**(5): 585−591. DOI: 10.3969/j.issn.1002-0411.2007.05.011.
ZHOU F F, FAN X P, YE Z. Designing Transfer Function Based on Adaptive Bandwidth Mean Shift Clustering Algorithm[J]. *Inf Contr*, 2007, **36**(5): 585−591. DOI: 10.3969/j.issn.1002-0411.2007.05.011.