智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

# 结合聚类边界采样的主动学习

胡峰, 李路正, 代劲, 刘群

引用本文:
胡峰,李路正,代劲,刘群. 结合聚类边界采样的主动学习[J]. 智能系统学报, 2024, 19(2): 482–492.
HU Feng, LI Luzheng, DAI Jin, et al. Active learning combined with clustering boundary sampling[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 482–492.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202205020

**DOI**: 10.11992/tis.202205020
**网络出版地址**: https://link.cnki.net/urlid/23.1538.TP.20231117.1722.010

# 结合聚类边界采样的主动学习

胡峰, 李路正, 代劲, 刘群
(重庆邮电大学 计算机科学与技术学院, 重庆 400065)

**摘　要:** 主动学习是一种机器学习方法, 需要选择最有价值的样本进行标注。目前, 主动学习在应用时面临着一些挑战, 其依赖分类器的先验假设, 这容易导致分类器性能意外下降, 同时需要一定规模的样本作为启动条件。聚类可以降低问题规模, 是主动学习的一种有效手段。为此, 结合密度聚类边界采样, 开展主动学习方法的研究。针对容易产生分类错误的聚类边界区域, 通过计算样本密度, 提出一种密度峰值聚类边界点采样方法; 在此基础上, 给出密度熵的定义, 并利用密度熵对聚类边界区域进行启发式搜索, 提出一种基于聚类边界采样的主动学习方法。试验结果表明, 与文献中的 5 种主动学习算法相比, 该算法能够以更少标记量获得同等甚至更高的分类性能, 是一种有效的主动学习算法; 在标记不足, 无标签样本总量 20% 的情况下, 算法在 Accuracy、F-score 等指标上取得较好的结果。

**关键词:** 主动学习; 机器学习; 聚类边界; 密度峰值聚类; 几何采样; 信息熵; 版本空间; 主动聚类
**中图分类号:** TP301　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)02−0482−11

**中文引用格式: 胡峰, 李路正, 代劲, 等. 结合聚类边界采样的主动学习 [J]. 智能系统学报, 2024, 19(2): 482–492.**
**英文引用格式: HU Feng, LI Luzheng, DAI Jin, et al. Active learning combined with clustering boundary sampling[J]. CAAI transactions on intelligent systems, 2024, 19(2): 482–492.**

# Active learning combined with clustering boundary sampling

HU Feng, LI Luzheng, DAI Jin, LIU Qun
(School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

**Abstract**: Active learning is a machine learning method that requires the selection of the most valuable samples for labeling. Currently, active learning encounters certain challenges in its practical application. It relies on prior assumptions of the classifier, which can lead to unexpected declines in classifier performance and requires a specific number of samples as an initial condition. Clustering, which can reduce the complexity of a problem, serves as an effective tool in active learning. Based on density clustering boundary sampling, this study focuses on active learning methods. First, a method of sampling boundary points in density peak clustering is introduced. This method calculates the sample density for a clustering boundary region that is prone to classification errors. Subsequently, with a specified definition of density entropy, an active learning method based on cluster boundary sampling is proposed. This method employs density entropy for the heuristic search of cluster boundary regions. The experimental results show that the proposed algorithm, compared with the five active learning algorithms referenced in the literature, can achieve equal or even higher classification performance with fewer markers. This proves that it is an effective active learning algorithm. When the number of labeled samples is less than 20% of the total number of unlabeled samples, the algorithm achieves better results in the accuracy and F-score metrics.

**Keywords**: active learning; machine learning; cluster boundary; density peak clustering; geometric sampling; entropy; version space; active clustering

**收稿日期:** 2022−05−17.　**网络出版日期:** 2023−11−20.
**基金项目:** 国家重点研发计划项目(2018YFC0832102); 重庆市教委重点合作项目(HZ2021008); 重庆市自然科学基金项目(cstc2021jcyj-msxmX0849).
**通信作者:** 胡峰. E-mail: hufeng@cqupt.edu.cn.

当前, 机器学习算法所需要处理的数据规模越来越大。传统的监督学习模型在训练时需要大量的已标记数据集, 然而在许多领域, 标记样本

依赖于特定领域的专家知识, 这使得标记的成本十分高昂, 如: 异常流量监测[1]、医学诊断[2]、图像分割[3]、金融交易领域的欺诈识别[4] 以及流体力学计算[5] 等。

作为一种机器学习方法, 主动学习能够在减小标记代价的同时, 以更少的样本训练得到一个满足预期指标的模型。以分类问题为例, 主动学习算法可以让模型选出最难区分的样本, 交付领域专家进行标记。主动学习的模型训练过程可以避免一些冗余样本的加入, 在降低标记成本的同时能够使分类器的精度快速达到预期值。主动学习的核心在于如何选取最有价值的样本进行标记, 根据场景可分为基于池的主动学习 (pool-based)、基于流的主动学习 (stream-based) 以及基于成员合成查询的主动学习 (membership query synthesis-based)。本研究讨论基于池的主动学习[6], 目前研究方法主要包括以下几种流行观点。

1) 基于信息量的观点。此类方法通过不确定性实现样本的选择。不确定性的度量有多种方法, Lewis 等[7] 利用信息熵刻画每个样本的不确定性, 优先选取熵最大的样本进行标记。Kee 等[8] 在批处理模式下构建多个分类器构成的委员会, 选择委员会的预测分歧最大的样本进行标记。Shao 等[9] 将基于委员会查询的思想应用于迁移学习领域, 通过维护来自源域与目标域的不同委员会成员, 以提高分类准确率。

2) 基于代表性的观点。此类方法主要通过密度计算或者聚类实现。基于密度的方法从空间中高密度区域选择具有代表性的样本, 这样可以避免离群点问题。Density-Weighted 方法[10] 考虑某个未标记样本点与其他未标记样本之间的平均相似度, 用于描述该未标记样本点的代表性。基于聚类的方法先对输入空间进行聚类, 然后在各簇中选取代表性实例。如 Min 等[11] 在文中基于三支决策思想[12] 提出了 TACS(three-way active learning through clustering selection) 算法, 其用聚类将数据集层次二分为数据块, 在此过程中根据块中样本的特点又分为 3 种操作: 如果块中没有足够的已标记样本, 则查询代表性实例; 如果块中有足够的同标签的已标记样本, 则对块中的其他实例进行分类; 如果块中有不同标签的已标记样本, 则对块进一步聚类。

3) 基于信息量和代表性的观点。一些研究人员试图结合信息量和代表性开展相关研究, 但两者需要权衡, 当实例的信息量较高时, 其代表性通常会降低。基于 min-max 框架[13], Huang 等[14] 在 2014 年提出了 QUIRE(querying informative and representative examples), 该方法使用已标记样本的预测精度度量信息量, 用未标记样本的预测精度度量代表性。

4) 其他观点。一些方法不是简单的基于信息量或代表性, 而是基于新的观点或理论。Dong[15] 提出了基于改进的支持向量机 (cost-sensitive SVM, CSSVM) 的主动学习算法, 以解决网络流量识别中的不平衡问题。Siddiqui 等[16] 提出一种用于语义分割的多视图主动学习策略 (ViewAL), 通过结合不同视图预测的不一致性来评估模型的不确定性, 该方法能有效降低语义分割的模型训练时间和标注代价。聚类边界的几何采样是最近出现的一种观点, 认为聚类边界点是潜在的标记样本, 这些样本对于分类器的提升具有真正的价值。从版本空间理论来看, 基于不确定性和代表性的策略可以看成是近似为超球体的版本空间上的内部体积采样和外部体积采样[17]。Cao[18] 在论文中提出一种基于骑士巡游的几何边界主动学习方法——GAL(geometric active learning), 将主动学习的不确定性采样问题转为了聚类边界点的几何采样问题, 摆脱了分类器的假设依赖。GAL 计算各样本距离其$k$个近邻点的概率路径转移长度, 并通过该值的大小区分聚类核心点与边界点。

密度峰值聚类算法 (density peak clustering, DPC)[19] 是 Rodriguez 和 Laio 提出的一种聚类方法, 根据数据的结构和层次关系, 可以从中发现高密度点和密度更高但距离较远的点。利用聚类中心由局部较低密度点包围的基本假设, 该方法可识别任意形状的数据结构, 也利于发现离群点。利用密度峰值聚类算法, 既可以通过密度峰值点将数据快速划分为多个密度区域, 又可以在聚类边界区域获取易错分的点。如果能兼顾二者, 则将是进行主动学习的有效途径。

针对主动学习算法启动缓慢、未标记样本严重依赖初始假设的问题。本研究结合密度峰值聚类的局部密度概念和聚类边界的几何采样思想, 开展了主动学习方法的研究。首先, 提出了一种基于密度峰值的聚类边界点采样方法; 在此基础上, 提出一种主动学习算法——基于聚类边界采样的主动学习方法 (boudary sampling with density entropy, BSDE), 该算法使用密度熵对聚类边界区域进行启发式搜索, 进而获取兼具信息量和代表性的聚类边界样本点进行标记。文中的算法与 5 种先进的主动学习算法进行了对比, 试验结果表明, 本研究提出的算法是一种有效的主动学

习算法，能够以更少标记量使分类器取得更高的性能。

## 1　相关工作

### 1.1　密度峰值聚类

密度峰值聚类是一种基于密度的聚类算法，该算法建立在 2 个基本假设[19] 之上：一是聚类中心由其局部密度较低的近邻点包围；二是这些聚类中心相比于其密度更高的点，距离较远。对于每个样本点，密度峰值算法要求计算 2 个量：局部密度 ($\rho_i$) 和较该点密度更大的最近点距离 ($\delta_i$)。

每个样本点$i$的数据密度$\rho_i$是指小于截断距离的近邻样本点数，计算公式为

$$\rho_i = \rho(i) = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

式中：$d_{ij}$为样本点$i$和样本点$j$的距离，通常使用欧氏距离；$d_c$为截断距离，可以人为设置或者根据样本点对距离的分布自动设置；$\delta_i$是样本点$i$较其密度更大样本点的距离最小值，计算公式为

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

如果$i$为密度最大的样本点，那么$\delta_i$可取距离$i$最远样本点$j$的距离$d_{ij}$。

利用上述 2 个量，可以构建横轴为$\rho$，纵轴为$\delta$的决策图。根据决策图，可将样本点分为密度峰值点、正常点和离群点，密度峰值点就是聚类中心。聚类中心数量可以由用户根据决策图选取，也可以选取$k$个$\rho\cdot\delta$(简记为$\gamma$) 最大的点。当聚类中心确定后，再将其他点一次分配到最近的密度较高点所在的簇中。

### 1.2　主动学习

主动学习主要包括 2 个步骤：1）选择有价值的样本，将其交给专家进行标记，然后将专家标记样本与已标记样本集构成新的训练集；2）在新训练集上进行重新训练得到新模型，利用新模型对测试集进行重新测试，记录分类器的性能指标。上述 2 个过程不断迭代，直至分类器达到预设指标或是学习过程超出预设代价 (一般用标记数量刻画)。

定义数据集为$\mathcal{D}(\mathcal{D}=\{\mathcal{D}_{\text{train}},\mathcal{D}_{\text{test}}\})$，训练集为$\mathcal{D}_{\text{train}}=\{\mathcal{D}_l,\mathcal{D}_u\}$，其中，$\mathcal{D}_l$表示用于初始化分类器的已标记集，$\mathcal{D}_u$表示主动学习起始时的无标记集(也称作未标记池)。选择有价值的样本可以看成是一个采样过程，不确定性采样和聚类边界点采样是实现该过程的 2 种思路，下面进行简要介绍。

#### 1.2.1　不确定性采样

以多分类问题为例，不确定性采样常用信息熵[20] 刻画未标记样本的不确定性，即基于熵的不确定性采样 (Entropy)。在选择有价值样本标记前先使用已标记集训练一个初始分类器，随后分类器会对所有未标记样本进行预测，并选取后验概率熵值最大的样本进行标记。选取结束后更新已标记集和未标记池，再更新分类器并重复上述过程。基于熵的不确定性采样可描述为

$$x_{\text{sel}} = \underset{x}{\operatorname{argmax}} -\sum_i P_\theta(y_i|x)\log(P_\theta(y_i|x)) \tag{3}$$

式中：$x_{\text{sel}}$表示选择标记的样本；$\theta$表示分类器的一组参数；$p_\theta(y_i|x)$表示由参数$\theta$确定的分类器将样本$x$预测为类别$y_i$的概率。

#### 1.2.2　聚类边界点采样

聚类边界点是分布在每个聚类边缘区域的一组特殊对象，其标签由聚类结构给出，指导聚类划分。Xia 等[21] 给出了聚类边界点和聚类核心点的形式化描述。

**定义 1**　聚类边界点$b$

1）其位于一个稠密区域$\mathcal{R}$内。

2）存在一个$b$的近邻区域$\mathcal{R}'$，其满足以下任意一个条件：

$$\text{Density}(\mathcal{R}') \gg \text{Density}(\mathcal{R})$$
$$\text{Density}(\mathcal{R}') \ll \text{Density}(\mathcal{R})$$

**定义 2**　聚类核心点$c$

1）其位于一个稠密区域$\mathcal{R}$内。

2）存在一个基于$\mathcal{R}$的拓展区域$\mathcal{R}''$，满足

$$\text{Density}(\mathcal{R}'') - \text{Density}(\mathcal{R}) \to 0。$$

对于一个聚类结构良好的数据集，其聚类边界点能够刻画出分类器的目标决策边界。对此，Cao[18] 提出了一个几何观点，即分类器的性能是由聚类边界点决定的。

**定理 1**　设$\xi$和$\eta$分别表示聚类的核心点和边界点构成的集合，其并集$\Xi=\{\xi,\eta\}$；设$h^{\Xi}$、$h^{\eta}$分别表示$\Xi$和$\eta$对应的分类器解空间。则，对于泛化误差分歧$\Delta'$，满足

$$\Delta' = \text{error}\left(h^{\Xi}\right) - \text{error}(h^{\eta}) \to 0 \tag{4}$$

边界距离表示一个点到分类器超平面的距离。由于聚类核心点的边界距离较聚类边界点的距离更远，因此上述定义易证。

聚类边界采样算法期望获取类簇的边界点，这些点真正决定了分类器的版本空间。在主动学习中，根据此特点可以选择出最有标记价值的样本。Cao[18] 最早将聚类边界点采样运用到主动学习领域，提出了一种基于概率路径转移长度的聚类边界点采样算法 GAL，该算法的灵感来源于图论中的 Knight's tour 问题。在 GAL 算法中，需要

计算样本点$x_i$距离其$\varepsilon$个近邻点$M_i^j$的概率路径转移长度$\mathcal{M}_i$

$$\mathcal{M}_i = \sum_{j=1}^{\varepsilon} \frac{\left\| r_{1\times 1}^{x_i \to M_i^j} \right\|_2^2}{\sum_{v=1}^{\varepsilon} \left\| r_{1\times 1}^{x_i \to M_i^v} \right\|_2} \tag{5}$$

一般说来, 聚类核心点转移到其$\varepsilon$个近邻的路径长度之和较大, 即概率路径转移长度较大; 而聚类边界点则相反。GAL 算法通过计算各样本点的概率路径转移长度值, 之后进行降序排列, 依次选择前$t$的样本 ($t$一般大于 30%)[22] 作为聚类边界点。

## 2　结合聚类边界采样的主动学习

首先介绍一种基于密度峰值的聚类边界点采样算法; 在此基础上, 提出一种主动学习算法。

### 2.1　基于密度峰值的边界点采样算法

本研究的采样方法基于密度峰值聚类的 2 个假设, 满足这类假设的数据集具有核心区域比较稠密、边界区域比较稀疏的特点。假设数据集被密度峰值聚类算法划分成$k$个簇, 且各个簇样本点的密度分布相似, 那么可以根据决策图直接筛选若干最小密度$\rho$样本点作为边界点, 这是一种基于全局的策略。GAL 算法也采取了这种策略, 但该方式未考虑数据局部结构。由于数据局部区域性质可能存在区别 ( 例如某些簇的密度偏大, 某些簇的密度偏小 ), 单一的全局选择方式容易导致边界点的选取效果变得很不稳定。

为了解决以上问题, 本研究首先使用密度峰值对数据进行聚类。密度聚类之后, 可以得到不同密度的聚类簇。再对聚类结果进行边界采样。

**定义 3**　边界离群点$o$

如果$o$是边界样本集中的离群点, 那么满足$o = \left\{ \left| Z(b^C) \right| > Z_{\text{th}}, C \in \{1,2,\cdots,k\} \right\}$, 其中, $Z(\cdot)$表示 $z$ 分数, $b^C$表示属于簇$C$的边界点, $Z_{\text{th}}$表示密度偏离阈值, 用于刻画离群点的密度相对于所属簇的密度均值偏移了多少个标准差, 一般大于 2.5。

**算法 1**　基于密度峰值的边界点采样算法。

**输入**　数据集$\mathcal{D}$, 离群点的密度偏离阈值$Z_{\text{th}}$, 簇数$k$, 边界样本占总样本的比例$\lambda$。

**输出**　聚类边界点的集合$S$

1) 数据集$\mathcal{D}$进行密度峰值聚类, 记录所有样本点的密度$\rho$以及簇标记 label。

2) $i \leftarrow 1$。

3) $S \leftarrow \varnothing$。

4) WHILE $i \leqslant k$。

5)　记录当前簇的样本点个数 cnt。

6)　$N \leftarrow \lfloor \text{cnt} \cdot \lambda \rfloor$。

7)　$\text{seq} \leftarrow \text{top}_N(\text{sort}_\rho(C_i))$。

8)　$S \leftarrow S \cup \text{seq}$。

9)　根据定义 3 选取离群点 $o$。

10)　$S \leftarrow S \setminus o$。

11)　$i \leftarrow i+1$。

12) END WHILE。

13) 输出$S$。

算法 1 描述了聚类边界样本的采样过程。1) 表示密度峰值聚类过程, 需要记录密度和样本的类簇标签; 5) −6) 表示计算边界样本点的采样个数; 7) −8) 表示按照$\rho$的大小对各簇的样本点进行降序排序, 并取前$N$个加入到边界样本集中; 9) −10) 表示去除边界样本的离群点。

### 2.2　基于聚类边界点采样与密度熵的主动学习算法 (BSDE)

分类器的作用是将特征空间划分为多个类别的决策区域, 这些区域的边界称为决策边界。这些边界点位于分离区域上, 可以用封闭几何曲面近似拟合为类簇。与离群点不同, 这些边界点具有明确的标记, 并与类簇内部点相连[23]。因此, 检测到聚类边界点可以使分类器版本空间最小化, 更有利于模型预测。

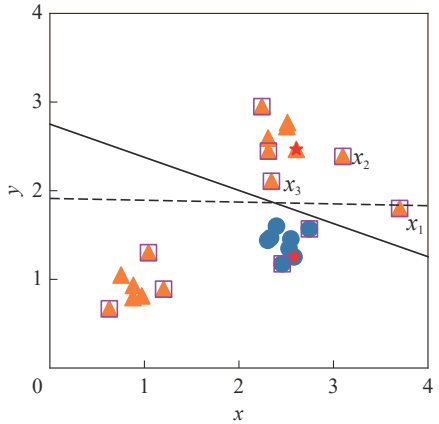
在 GAL 算法中, 一旦获取到聚类边界点, 便选取具有较大$\mathcal{M}_i$的样本进行标记。该算法与分类器假设无关, 性能较为稳定, 但存在几个不足之处: 1) GAL 算法在数据规模较大时收敛较慢; 2) 容易采集到离群点; 3) 没有考虑数据集的局部结构信息, 容易造成采样偏差。为提高算法收敛速度, 降低离群点的采集率, 本研究将信息熵与密度峰值聚类中的密度概念结合, 提出一种新的采集函数——密度熵 (density entropy-DE), 其公式如下

$$\text{DE}(x) = \text{Entropy}(x)^{1-\beta} \cdot \rho(x)^{\beta} \tag{6}$$

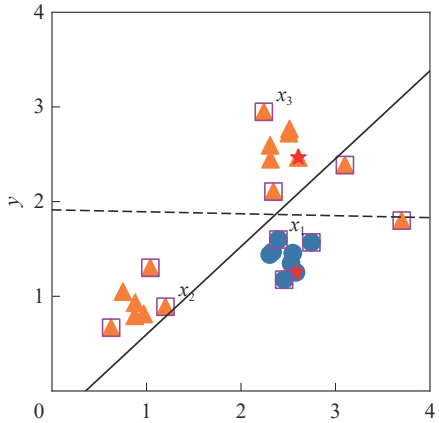
式中: $\text{Entropy}(x)$表示样本$x$的信息熵, 用于度量不确定性; $\rho(x)$表示密度, 采用式 (1) 进行计算, 反映一个点邻域内的点的数量, 可以刻画该点在局部区域中的代表性; $\beta$表示权衡参数, 取值范围在 0~1 间, 默认取 0.1。密度熵也可以看作是信息熵的推广, 特别的, 当$\beta$为 0 时, 密度熵就退化为信息熵。

这里给出一个示例对密度熵进行说明, 如图 1 所示。由图 1 可知, 圆形和三角形分别表示分类问题的 2 种类别; 正方型框选的表示算法选出的边界样本; 星型标注的表示初始标记样本; 英文

字母标注的表示主动学习选择标记的样本，按照标记次序依次记为$x_1$, $x_2$, $x_3$；虚线表示分类器经初始样本集训练后的决策边界；实线表示分类器经新的训练集（包含新标记样本）训练后的决策边界。显然，图 1(a) 中的$x_1$是一个离群点，不应该被优先选择。但是，根据信息熵的定义，其后验概率的信息熵较大，导致$x_1$被优先选择。根据式 (1) 计算其密度为 0，该点的密度熵也为 0，$x_1$将不会被优先标记。因此，在这种情况下，密度熵更符合数据的实际情况（如图 1(b) 所示，离群点没有被标记）。




(a) GAL




(b) BSDE

**图 1　GAL 和 BSDE 在合成数据集上的样本选择变化情况**

**Fig. 1　Variations in sample selection for GAL and BSDE on synthetic datasets**

为了避免局部区域采样率过高，造成采样偏差，首先，使用算法 1 对输入空间进行边界点采样；其次，使用密度熵对边界样本进行启发式搜索，找出信息量和代表性均较高的样本点。这里给出了一种基于聚类边界点采样与密度熵的主动学习 (BSDE) 算法。

**算法 2**　基于聚类边界点采样与密度熵的主动学习算法 (BSDE 算法)。

**输入**　数据集$\mathcal{D}$，边界离群点的密度偏离阈值$Z_{\text{th}}$，簇数$k$，边界样本占总样本的比例$\lambda$，权衡参数$\beta$，分类器$\theta$，初始标记样本数$p$，分类器的初始已标记样本集$D_l$，查询样本占未标记池的比例$q$，每轮的查询量 batch_size

**输出**　训练后的分类器$\theta$

1) 使用算法 1 获取$\mathcal{D}$的聚类边界样本集$S$。

2) count ← 0。

3) $N_q \leftarrow (|\mathcal{D}| - p)\cdot q$。

4) WHILE count < $N_q$。

5)　$\text{DE}(S) = \{\text{DE}(x), x \in S\}$。

6)　selected ← $\text{top}_{\text{batch\_size}}(\text{sort}_{\text{DE}}(S))$。

7)　selected ← query_label(selected)。

8)　$D_l \leftarrow D_l \cup \text{selected}$。

9)　$\theta \leftarrow \text{retrain}(\theta)$。

10)　$S \leftarrow S \setminus \text{selected}$。

11)　count ← count + |selected|。

12) END WHILE。

13) 输出$\theta$。

算法 2 描述了基于聚类边界点采样与密度熵的主动学习过程。1) 表示获取边界样本；2) −3) 表示获取当前标记量以及标记上限；5) −6) 表示计算边界样本的密度熵并选择最大的若干样本准备标记；7) 表示专家查询样本的标签；9) 表示分类器重训练。

算法的时间复杂度分析：设样本点个数为$n$，每个簇的样本量都是$n^C$（$k$和$n^C$分别表示簇数和各个簇的样本量，$n^C$的值与簇号$C$有关）。1）表示基于密度峰值的边界采样过程，其时间复杂度是$O(k\cdot n^C\cdot\log(n^C))+O(n^2)$。2）−12）表示使用密度熵启发式搜索边界区域的过程，为分析方便，可以将聚类边界采样点的规模也视作$n$，这样问题就转为从$n$个样本中选取$N_q$个密度熵最大的样本，时间复杂度为$O(n\log(N_q))$。故算法时间复杂度为$O(n\cdot\log(N_q))+O(n^2)+O(k\cdot n^C\cdot\log(n^C))$，由于$N_q$小于$n$，$O(n\cdot\log(N_q))+O(n^2)=O(n^2)$。因此，算法 2 的时间复杂度为$O(k\cdot n^C\cdot\log(n^C))+O(n^2)$。

## 3　试验结果与分析

为了验证文中算法在边界采样和主动学习方面的有效性，下面将分别进行试验。在 3.1 中将在 flame 数据集上检验文中方法的有效性；在 3.2 中将在 Accuracy、F-score、ALC-Acc 等指标上与 5 种先进的主动学习算法进行对比。

### 3.1　基于密度峰值的聚类边界点采样的有效性

图 2 给出了二分类数据集 flame 上，基于密度

峰值的聚类采样算法在设置$k=2$，1，3，4时的边界采样效果, 图 2 中密度峰值点和聚类核心点被聚类边界点所包络。可以看到, $k=2$是最佳的聚类数, 此时边界采样区域能够完全覆盖可行的假设空间。此外, 当$k$偏离了真实最佳聚类数 2(即$k\in\{1,3,4\}$) 时, 算法的整体采样效果也较为稳定。

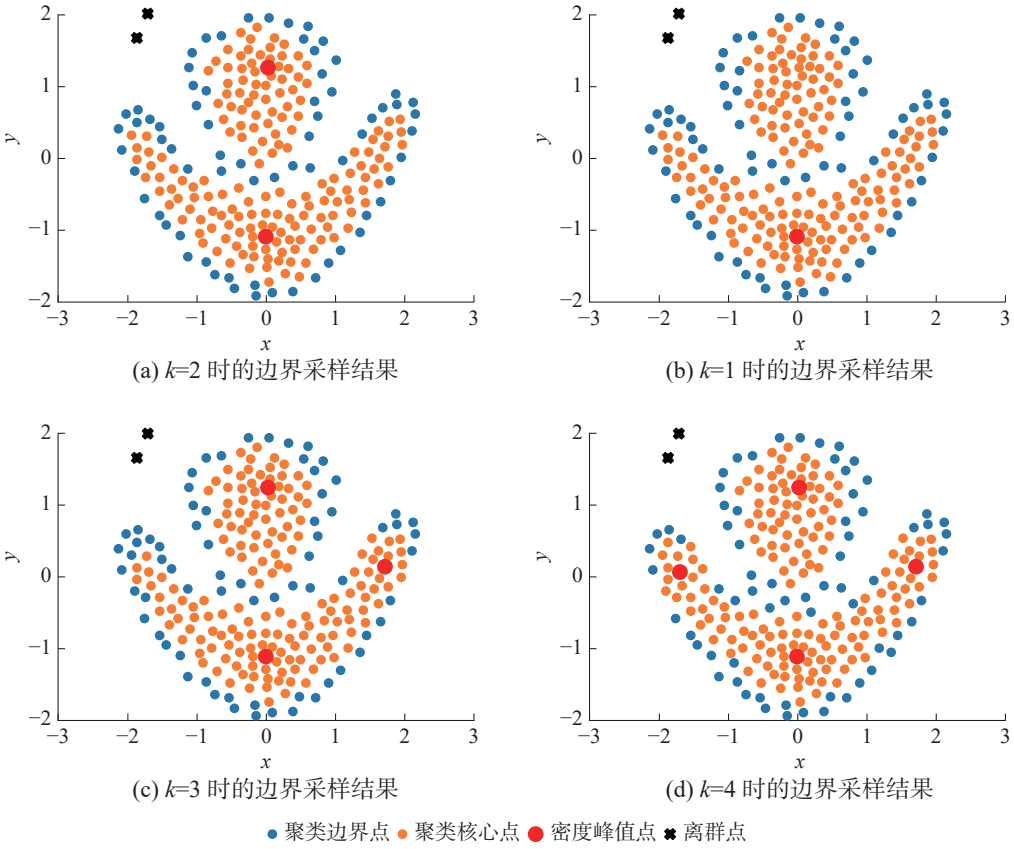



**图 2　边界采样算法在 flame 数据集上的结果**

**Fig. 2　Results of boundary sampling algorithm on the flame dataset**

### 3.2　BSDE 与几种先进主动学习算法的对比

#### 3.2.1　对比算法与数据集

这里对比了 5 种流行的主动学习算法, 分别如下。

1)LAL[24]: LAL 是一种期望错误最小化的方法, 其特点在于其将训练一个回归模型用于预测特定学习状态下候选样本的预期错误减少, 但只适用于二分类问题。在试验中, 使用笔者提供的预提取数据训练回归器。

2)Entropy[20]: Entropy 是一种基于不确定性采样的经典方法, 使用信息熵度量样本的不确定性, 见式(3)。

3)TACS[11]: TACS 是基于聚类选择的主动学习算法。其将原始数据不断二分为块, 并结合三支决策理论对不同状态的块进行分别处理, 在块中查询标签时, 选择具有密度峰值的代表性实例或总距离最大的信息实例。

4)QUIRE[14]: QUIRE 基于主动学习的 min-max 框架, 实例的信息性通过已标记数据的预测不确定性刻画, 而其代表性通过未标记数据的预测不确定性衡量。

5)GAL[18]: GAL 是基于聚类边界点采样的主动学习算法, 其计算每个样本点距离$k$个近邻的概率转移长度, 通过排序筛选出聚类边界点作为待标记样本。

在 12 个数据集上进行了对比试验, 由于 LAL 只能用于二分类, 因此选择了 5 个二分类数据集, 其余为多分类数据集。其具体信息详见表 1。

#### 3.2.2　试验设置

具体试验设置如下: 1)许多数据集默认有序, 为满足数据的独立同分布假设, 对各数据集预先进行随机无放回采样(采样数等于样本总数); 2)数据集采用了标准化; 3)为保证对比试验稳定性, 采用随机分层 10 折交叉验证, 统计指标的均值和标准差。此外, 由于对比算法需要初始标记样本引导训练过程, 因此在训练集中, 从各类别中随机选择 1 个样本构成初始标记集; 4)使用逻辑回归作为基分类器; 5)使用测试集的 Ac-

curacy、F-score 指标衡量不同迭代轮次下分类器的性能。为评估整个主动学习过程的性能[25]，还计算了学习曲线下区域 (ALC) 这一性能指标。对于 Accuracy 和 F1-Weighed 曲线，分别可以得到 ALC-ACC 和 ALC-F1-Weighed；6）对于 LAL，笔者提供了 2 个版本，使用其在论文中推荐的表现更好的 LAL-iterative-2D；7）对于 TACS，使用了笔者提供的 Java 源码。为满足输入要求，在保证样本内容不变的前提下将数据集由 csv 格式转为 arff 格式；8）对于 QUIRE，使用了作者提供的源码，核函数采用文中建议的 RBF 核。在试验中发现 QUIRE 在规模较大的数据集上运行缓慢，由于试验条件限制，最终只在 9 个较小规模的数据集上对比了 QUIRE；9）对于 BSDE，设置参数 $p=1$，$q=0.2$、$\lambda=0.35$、$Z_{\text{th}}=2.5$、$\beta\in[0,1]$(通常取 0.1)；10）在真实环境下，对于带有聚类的主动学习，聚类过程可以在整个数据集上进行，但为保证试验的公平性，避免泄露测试集信息，把无标记样本池作为带有聚类过程算法 (TACS、BSDE) 的输入数据；11）标记预算为未标记池样本总量的 20%，每次迭代的查询量 (batch size) 为 1，即每轮查询一个样本。

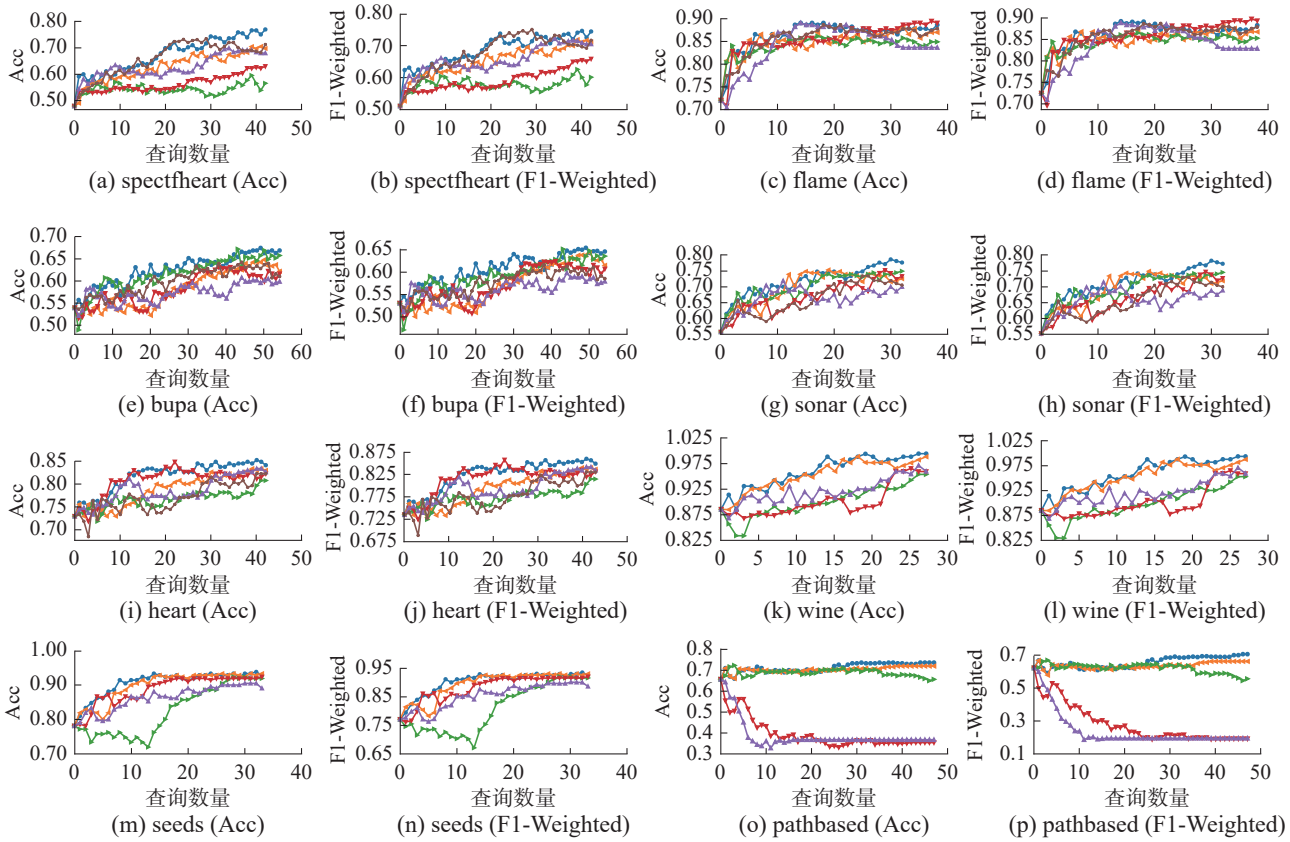
**表 1　试验数据集**

**Table 1　Experiment datasets**

| 数据集ID | 数据集名称 | 样本数 | 特征数 | 类别数 |
|---|---|---|---|---|
| 1 | spectfheart | 270 | 13 | 2 |
| 2 | flame | 240 | 2 | 2 |
| 3 | bupa | 345 | 6 | 2 |
| 4 | sonar | 208 | 60 | 2 |
| 5 | heart | 270 | 13 | 2 |
| 6 | wine | 178 | 13 | 3 |
| 7 | seeds | 210 | 7 | 3 |
| 8 | spiral | 321 | 2 | 3 |
| 9 | movement_libras | 360 | 90 | 15 |
| 10 | yeast | 1 484 | 8 | 10 |
| 11 | winequality-red | 1 599 | 11 | 6 |
| 12 | thyroid | 7 200 | 21 | 3 |

### 3.2.3　试验结果和分析

6 种主动学习方法的对比试验结果如图 3、表 2、表 3 所示。容易发现：本研究提出的 BSDE 算法在多数的数据集上取得领先 (如 spectheart、sonar、yeast)，在 12 个数据集上的 ALC-ACC 与 ALC-F1-Weighted 值的平均排名取得第一，在 6 种主动学习算法中取得了最好的表现。



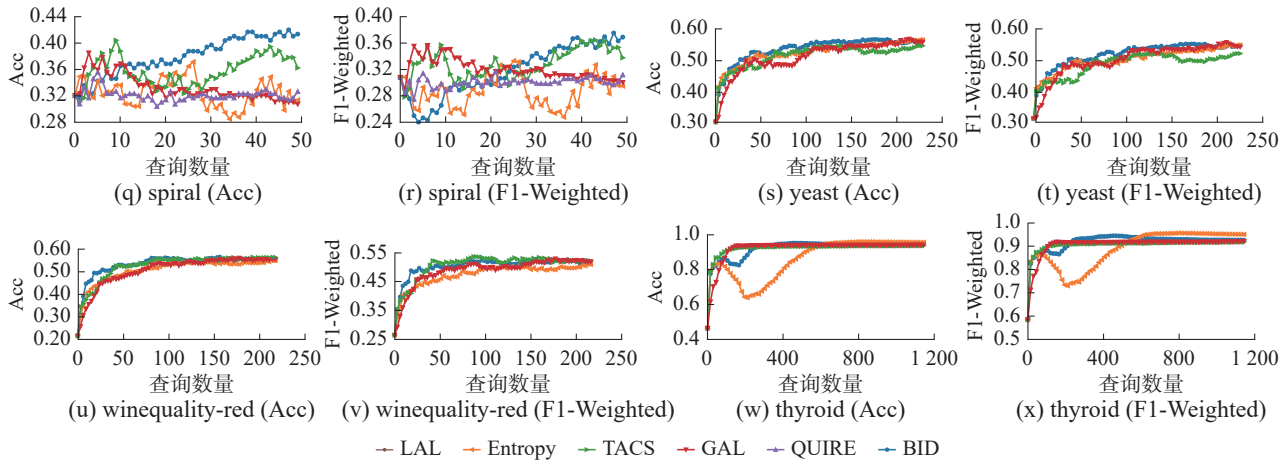



**图 3　12 个数据集上 Accuracy 和 F1-score 的变化曲线**

**Fig. 3　Accuracy and F1-score variation curves on 12 datasets**

**表 2　主动学习算法在不同标注比例下的 F1-Weighted**

**Table 2　F1-Weighted of active learning algorithms at different labeling ratios**

| 数据集ID | 数据集名称 | 标注比例 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1% | 3% | 5% | 7% | 10% | 13% | 15% | 20% |
| 1 | LAL | 0.536±0.102 | 0.597±0.068 | 0.637±0.059 | **0.681±0.067** | 0.704±0.065 | **0.739±0.049** | 0.725±0.066 | 0.716±0.060 |
| | Entropy | 0.525±0.077 | 0.604±0.073 | 0.614±0.081 | 0.636±0.106 | 0.660±0.072 | 0.690±0.071 | 0.686±0.060 | 0.711±0.077 |
| | TACS | 0.554±0.100 | 0.599±0.083 | 0.603±0.098 | 0.593±0.101 | 0.564±0.103 | 0.582±0.108 | 0.554±0.079 | 0.600±0.109 |
| | GAL | 0.552±0.128 | 0.555±0.107 | 0.570±0.113 | 0.570±0.096 | 0.569±0.091 | 0.603±0.098 | 0.616±0.089 | 0.657±0.089 |
| | QUIRE | 0.570±0.102 | **0.632±0.105** | **0.660±0.106** | 0.632±0.088 | 0.632±0.088 | 0.641±0.075 | 0.704±0.071 | 0.704±0.088 |
| | BSDE | **0.622±0.069** | 0.601±0.106 | 0.631±0.097 | 0.664±0.093 | **0.718±0.102** | 0.720±0.105 | **0.727±0.084** | **0.744±0.055** |
| 2 | LAL | 0.724±0.148 | 0.818±0.091 | 0.844±0.072 | **0.870±0.076** | 0.885±0.061 | 0.873±0.069 | 0.861±0.070 | 0.878±0.070 |
| | Entropy | 0.724±0.148 | 0.809±0.101 | 0.841±0.096 | 0.850±0.098 | 0.854±0.075 | 0.866±0.072 | 0.866±0.080 | 0.867±0.067 |
| | TACS | 0.724±0.148 | 0.813±0.106 | 0.831±0.083 | 0.855±0.092 | 0.860±0.076 | 0.860±0.080 | 0.848±0.083 | 0.852±0.081 |
| | GAL | 0.724±0.148 | **0.847±0.085** | **0.851±0.070** | 0.848±0.067 | 0.852±0.068 | **0.873±0.060** | **0.877±0.055** | **0.892±0.049** |
| | QUIRE | 0.724±0.148 | 0.782±0.093 | 0.812±0.114 | 0.858±0.078 | 0.878±0.076 | 0.870±0.073 | 0.854±0.087 | 0.828±0.058 |
| | BSDE | 0.724±0.148 | 0.827±0.083 | 0.849±0.063 | 0.856±0.082 | **0.890±0.072** | 0.870±0.072 | 0.868±0.073 | 0.882±0.076 |
| 3 | LAL | **0.521±0.110** | 0.554±0.121 | 0.556±0.108 | 0.536±0.099 | 0.581±0.097 | 0.595±0.078 | 0.583±0.088 | 0.585±0.054 |
| | Entropy | 0.517±0.071 | 0.510±0.054 | 0.521±0.063 | 0.529±0.090 | 0.574±0.071 | 0.601±0.084 | 0.617±0.073 | 0.615±0.103 |
| | TACS | 0.470±0.085 | **0.575±0.093** | **0.572±0.073** | **0.570±0.063** | **0.597±0.061** | 0.600±0.072 | **0.624±0.075** | **0.635±0.061** |
| | GAL | 0.496±0.100 | 0.553±0.115 | 0.546±0.109 | 0.511±0.087 | 0.552±0.094 | **0.617±0.114** | 0.621±0.099 | 0.607±0.092 |
| | QUIRE | 0.510±0.101 | 0.561±0.085 | 0.527±0.084 | 0.542±0.075 | 0.558±0.077 | 0.577±0.075 | 0.554±0.097 | 0.579±0.087 |
| | BSDE | 0.545±0.090 | 0.561±0.052 | 0.545±0.063 | 0.592±0.069 | 0.603±0.068 | 0.632±0.060 | 0.617±0.084 | 0.646±0.086 |
| 4 | LAL | 0.552±0.081 | 0.656±0.139 | 0.598±0.113 | 0.622±0.094 | 0.628±0.120 | 0.708±0.148 | 0.719±0.128 | 0.702±0.085 |
| | Entropy | 0.552±0.081 | 0.634±0.182 | 0.607±0.125 | 0.662±0.153 | **0.747±0.109** | **0.743±0.096** | 0.739±0.090 | 0.722±0.131 |
| | TACS | 0.552±0.081 | **0.677±0.138** | 0.676±0.127 | 0.673±0.124 | 0.680±0.122 | 0.728±0.087 | **0.740±0.081** | 0.749±0.059 |
| | GAL | 0.552±0.081 | 0.605±0.143 | 0.643±0.144 | 0.614±0.144 | 0.666±0.173 | 0.683±0.134 | 0.692±0.116 | 0.730±0.141 |
| | QUIRE | 0.552±0.081 | 0.618±0.136 | 0.667±0.165 | 0.682±0.132 | 0.684±0.114 | 0.669±0.118 | 0.640±0.124 | 0.691±0.152 |
| | BSDE | 0.552±0.081 | 0.654±0.129 | **0.687±0.129** | **0.696±0.115** | 0.732±0.104 | 0.742±0.101 | 0.730±0.111 | **0.777±0.108** |

续表 2

| 数据集ID | 数据集名称 | 标注比例 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1% | 3% | 5% | 7% | 10% | 13% | 15% | 20% |
| 5 | LAL | 0.733±0.088 | 0.721±0.135 | 0.746±0.147 | 0.766±0.113 | 0.747±0.074 | 0.774±0.085 | 0.806±0.052 | 0.821±0.042 |
| | Entropy | 0.753±0.061 | 0.722±0.076 | 0.727±0.080 | 0.765±0.072 | 0.791±0.056 | 0.801±0.048 | 0.820±0.039 | 0.828±0.036 |
| | TACS | 0.736±0.068 | 0.718±0.077 | 0.773±0.051 | 0.751±0.062 | 0.767±0.095 | 0.772±0.080 | 0.773±0.054 | 0.806±0.076 |
| | GAL | 0.743±0.074 | 0.741±0.078 | **0.803±0.079** | 0.822±0.055 | **0.837±0.058** | **0.830±0.078** | 0.822±0.065 | 0.822±0.074 |
| | QUIRE | 0.743±0.054 | 0.755±0.076 | 0.772±0.066 | 0.799±0.067 | 0.770±0.068 | 0.781±0.087 | 0.818±0.056 | 0.825±0.097 |
| | BSDE | **0.758±0.084** | **0.762±0.150** | 0.798±0.060 | **0.824±0.056** | 0.831±0.041 | 0.825±0.058 | **0.848±0.059** | **0.840±0.047** |
| 6 | Entropy | 0.885±0.062 | 0.914±0.107 | **0.930±0.073** | 0.925±0.055 | 0.961±0.053 | 0.977±0.041 | 0.977±0.042 | 0.988±0.025 |
| | TACS | 0.885±0.062 | 0.829±0.104 | 0.877±0.090 | 0.878±0.084 | 0.895±0.076 | 0.920±0.064 | 0.926±0.068 | 0.954±0.063 |
| | GAL | 0.885±0.062 | 0.868±0.082 | 0.875±0.065 | 0.879±0.076 | 0.895±0.112 | 0.878±0.089 | **0.889±0.089** | 0.960±0.056 |
| | QUIRE | 0.885±0.062 | 0.908±0.108 | 0.913±0.091 | 0.902±0.087 | 0.926±0.039 | 0.914±0.031 | 0.925±0.040 | 0.960±0.039 |
| | BSDE | 0.885±0.062 | **0.930±0.068** | 0.919±0.058 | **0.936±0.065** | **0.971±0.041** | **0.983±0.028** | 0.988±0.024 | **0.994±0.018** |
| 7 | Entropy | 0.770±0.176 | 0.804±0.112 | 0.872±0.056 | 0.899±0.058 | **0.924±0.033** | **0.928±0.041** | 0.928±0.041 | 0.933±0.034 |
| | TACS | 0.770±0.176 | 0.737±0.174 | 0.710±0.181 | 0.701±0.130 | 0.783±0.132 | 0.864±0.090 | 0.885±0.058 | 0.918±0.065 |
| | GAL | 0.770±0.176 | **0.862±0.059** | 0.822±0.112 | 0.855±0.086 | 0.908±0.062 | 0.917±0.047 | 0.913±0.051 | 0.923±0.053 |
| | QUIRE | 0.770±0.176 | 0.771±0.174 | 0.807±0.171 | 0.853±0.133 | 0.866±0.064 | 0.880±0.073 | 0.894±0.059 | 0.888±0.052 |
| | BSDE | 0.770±0.176 | 0.855±0.084 | **0.883±0.056** | **0.913±0.060** | 0.922±0.054 | 0.918±0.040 | **0.933±0.040** | **0.928±0.051** |
| 8 | Entropy | **0.309±0.081** | 0.274±0.065 | 0.254±0.056 | 0.290±0.082 | **0.325±0.106** | 0.265±0.090 | 0.247±0.096 | 0.295±0.113 |
| | TACS | 0.278±0.111 | 0.356±0.121 | 0.314±0.086 | 0.310±0.097 | 0.296±0.119 | 0.320±0.100 | 0.343±0.091 | 0.339±0.100 |
| | GAL | 0.290±0.105 | **0.358±0.067** | **0.352±0.094** | **0.333±0.081** | 0.324±0.077 | 0.321±0.083 | 0.312±0.083 | 0.301±0.089 |
| | QUIRE | 0.284±0.085 | 0.305±0.078 | 0.290±0.064 | 0.309±0.076 | 0.301±0.074 | 0.299±0.068 | 0.307±0.084 | 0.312±0.084 |
| | BSDE | 0.291±0.099 | 0.243±0.088 | 0.293±0.111 | 0.292±0.109 | 0.318±0.084 | **0.339±0.078** | **0.354±0.085** | **0.371±0.082** |
| 9 | Entropy | 0.360±0.099 | **0.396±0.084** | 0.373±0.088 | 0.402±0.093 | 0.423±0.125 | 0.426±0.087 | 0.430±0.071 | 0.456±0.060 |
| | TACS | 0.351±0.097 | 0.380±0.088 | 0.393±0.065 | 0.421±0.093 | 0.433±0.103 | **0.469±0.101** | 0.478±0.094 | 0.460±0.090 |
| | GAL | **0.370±0.110** | 0.376±0.076 | 0.393±0.098 | 0.397±0.089 | 0.434±0.085 | 0.413±0.112 | 0.424±0.096 | 0.431±0.094 |
| | QUIRE | 0.341±0.101 | 0.389±0.084 | 0.407±0.074 | 0.399±0.075 | 0.408±0.093 | 0.413±0.088 | 0.427±0.095 | 0.439±0.087 |
| | BSDE | 0.358±0.102 | 0.372±0.087 | **0.398±0.097** | **0.426±0.106** | **0.452±0.083** | 0.469±0.073 | **0.485±0.062** | **0.496±0.096** |
| 10 | Entropy | 0.441±0.062 | 0.476±0.054 | 0.484±0.035 | 0.492±0.029 | 0.518±0.041 | 0.522±0.044 | 0.532±0.051 | **0.551±0.056** |
| | TACS | 0.419±0.064 | 0.427±0.072 | 0.470±0.043 | 0.497±0.030 | 0.520±0.033 | 0.515±0.022 | 0.498±0.036 | 0.527±0.033 |
| | GAL | 0.350±0.055 | 0.486±0.054 | 0.479±0.045 | 0.483±0.053 | 0.533±0.041 | **0.544±0.037** | 0.538±0.047 | 0.543±0.054 |
| | BSDE | **0.431±0.099** | **0.490±0.047** | **0.499±0.060** | **0.528±0.059** | **0.539±0.067** | **0.544±0.056** | **0.556±0.037** | 0.545±0.045 |
| 11 | Entropy | 0.402±0.062 | 0.445±0.057 | 0.464±0.046 | 0.482±0.047 | 0.495±0.056 | 0.497±0.046 | 0.493±0.041 | 0.506±0.034 |
| | TACS | 0.393±0.073 | 0.471±0.053 | **0.506±0.031** | **0.515±0.029** | **0.528±0.045** | **0.520±0.044** | **0.533±0.033** | **0.518±0.020** |
| | GAL | 0.359±0.080 | 0.465±0.052 | 0.481±0.038 | 0.495±0.037 | 0.496±0.033 | 0.499±0.031 | 0.519±0.044 | 0.516±0.037 |
| | BSDE | **0.433±0.052** | **0.482±0.046** | 0.501±0.039 | 0.505±0.032 | 0.513±0.036 | 0.513±0.033 | 0.513±0.042 | 0.515±0.033 |
| 12 | Entropy | 0.866±0.084 | 0.780±0.109 | 0.760±0.090 | 0.851±0.060 | 0.937±0.022 | **0.955±0.009** | **0.956±0.006** | 0.951±0.007 |
| | TACS | **0.873±0.012** | 0.908±0.006 | 0.911±0.004 | 0.913±0.004 | 0.914±0.004 | 0.917±0.005 | 0.918±0.006 | 0.919±0.007 |
| | GAL | 0.823±0.060 | **0.916±0.007** | 0.916±0.004 | 0.918±0.002 | 0.918±0.003 | 0.919±0.002 | 0.918±0.003 | 0.921±0.004 |
| | BSDE | 0.866±0.033 | 0.871±0.016 | **0.934±0.009** | **0.943±0.009** | **0.940±0.007** | 0.931±0.006 | 0.930±0.006 | **0.927±0.005** |

注: 黑体表示最好结果, 下同。

**表 3　主动学习算法的 ALC-F1-Weighted**

**Table 3　ALC-F1-Weighted of active learning algorithms**

| 数据集ID | LAL | Entropy | TACS | GAL | QUIRE | BSDE |
|---|---|---|---|---|---|---|
| 1 | 29.011±1.467 | 27.939±2.499 | 24.821±2.889 | 25.305±3.939 | 28.119±2.78 | **29.392±2.705** |
| 2 | 33.315±2.225 | 33.105±2.516 | 32.871±2.782 | 33.363±2.023 | 32.667±2.432 | **33.548±2.287** |
| 3 | 31.656±3.011 | 31.388±3.102 | 32.482±2.907 | 31.443±4.638 | 30.864±3.605 | **33.321±2.898** |
| 4 | 21.712±3.128 | 22.95±3.025 | 22.953±2.365 | 21.929±3.758 | 21.743±3.388 | **23.539±3.156** |
| 5 | 33.053±2.973 | 33.719±1.533 | 32.873±2.041 | 34.644±1.921 | 33.789±2.601 | **35.037±1.976** |
| 6 | — | 26.653±0.949 | 25.164±1.69 | 25.222±1.782 | 25.773±1.239 | **26.843±0.795** |
| 7 | — | 30.413±1.338 | 27.478±2.78 | 29.955±1.198 | 29.016±2.449 | **30.708±1.639** |
| 8 | — | 14.408±2.606 | **16.489±3.745** | 16.041±3.258 | 15.06±3.265 | 15.943±3.565 |
| 9 | — | 20.722±3.783 | 21.637±4.086 | 20.587±4.045 | 20.447±3.231 | **21.881±3.752** |
| 10 | — | 116.602±7.14 | 112.643±7.37 | 116.409±9.55 | — | **120.335±9.949** |
| 11 | — | 102.347±8.972 | **109.021±5.667** | 104.985±6.097 | — | 108.586±6.006 |
| 12 | — | 1 017.49±28.683 | 1 037.681±5.134 | 1 035.1±6.399 | — | **1 050.47±5.864** |
| 平均值 | 29.749 | 123.145 | 124.676 | 124.582 | 26.386 | **127.467** |
| 平均排名 | 3.8 | 3.5 | 3.417 | 3.25 | 4.333 | **1.25** |

分析其原因, 可以发现: 1) TACS 在分块过程中进行标注, 如果分块过程出现问题, 容易导致整个块产生错误标记, 如 seeds 数据集; 2) Entropy 容易陷入空间中难以学习的局部区域, 造成分类器性能的意外下降[26], 如 thyroid 数据集, 在迭代前期分类器的性能产生骤降, 导致整体的学习曲线不佳; 3) QUIRE 在大部分数据集上的排名不高, 主要由于数据集局部区域的样本性质存在差异, 违背算法的基本假设; 4) LAL 从简单的二维合成数据集训练随机森林回归器, 并预测预期误差减少, 每轮主动学习迭代开始前, LAL 都需要重新训练回归器, 时间效率较低, 且由于该算法只能用于二分类, 进一步限制了其应用范围; 5) GAL 在缩减到稳定的版本空间前需要查询足够的聚类边界样本点, 导致分类器收敛速度较慢。从图 3 中可以看到, GAL 在大部分数据集上的学习曲线都较为平缓 (如 spectheart 和 winequality-red)。

## 4　结束语

许多基于池的主动学习方法根据当前分类假设和标记样本挑选样本标记, 在标记样本量较少的情况下, 学习曲线未能显著提升。受到分类器性能由聚类边界点决定这一观点的启发, 本研究首先提出一种基于密度峰值的聚类边界点采样方法; 在此基础之上, 为解决聚类边界采样主动学习收敛缓慢的问题, 定义了一种名为密度熵的采集函数, 其具有不确定性采样适用性强、收敛快的特点, 同时避免了其容易陷入局部区域查询的问题; 最后, 提出了一种新的主动学习算法——BSDE。试验结果表明, 本研究算法能够有效发掘聚类边界点、启发式搜索有价值的样本进行标记, 能在少量标记条件下有效提高分类器性能, 在对比试验中取得了较好结果。但是, 本研究提出的算法使用了经典的密度峰值算法版本, 存在时间复杂度较高、高维性能退化的问题, 这些因素可能会限制 BSDE 的性能。可以考虑采用更适用的距离定义, 如 MMD 或 Wasserstein 距离, 或者基于度量学习的方法来获得更准确的样本空间表征, 这是未来的主要工作。

## 参考文献:


[1] SHAHRAKI A, ABBASI M, TAHERKORDI A, et al. Active learning for network traffic classification: a technical study[J]. IEEE transactions on cognitive communications and networking, 2021, 8(1): 422–439.

[2] NATH V, YANG Dong, LANDMAN B A, et al. Diminishing uncertainty within the training pool: active learning for medical image segmentation[J]. IEEE transactions on medical imaging, 2021, 40(10): 2534–2547.

[3] 陈立伟, 房赫, 朱海峰. 多视图主动学习的多样性样本选择方法研究 [J]. 智能系统学报, 2021, 16(6): 1007–1014.
CHEN Liwei, FANG He, ZHU Haifeng. Diversity sample selection method of multiview active learning classification[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1007–1014.

[4] CARCILLO F, LE BORGNE Y A, CAELEN O, et al. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization[J]. International journal of data science and analytics, 2018,

5(4): 285–300.

[5] OWOYELE O, PAL P, VIDAL TORREIRA A. An automated machine learning-genetic algorithm framework with active learning for design optimization[J]. Journal of energy resources technology, 2021, 143(8): 082305.

[6] AGGARWAL C C, KONG X, GU Q, et al. Active learning: A survey [M]. [S. l. ]: Algorithms and Applications, 2014: 571−605.

[7] LEWIS D D, GALE W A. A sequential algorithm for training text classifiers[C]//Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.[S.l.]: SIGIR 1994, 29: 3−12.

[8] KEE S, DEL CASTILLO E, RUNGER G. Query-by-committee improvement with diversity and density in batch active learning[J]. Information sciences, 2018, 454/455: 401–418.

[9] SHAO Hao. Query by diverse committee in transfer active learning[J]. Frontiers of computer science, 2019, 13(2): 280–291.

[10] SETTLES B, CRAVEN M. An analysis of active learning strategies for sequence labeling tasks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. New York: ACM, 2008: 1070−1079.

[11] MIN Fan, ZHANG Shiming, CIUCCI D, et al. Three-way active learning through clustering selection[J]. International journal of machine learning and cybernetics, 2020, 11(5): 1033–1046.

[12] YAO Yiyu. Three-way decisions with probabilistic rough sets[J]. Information sciences, 2010, 180(3): 341–353.

[13] HOI S C H, JIN Rong, ZHU Jianke, et al. Semi-supervised SVM batch mode active learning for image retrieval[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008: 1−7.

[14] HUANG Shengjun, JIN Rong, ZHOU Zhihua. Active learning by querying informative and representative examples[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(10): 1936–1949.

[15] DONG Shi. Multi class SVM algorithm with active learning for network traffic classification[J]. Expert systems with applications, 2021, 176: 114885.

[16] SIDDIQUI Y, VALENTIN J, NIESSNER M. ViewAL: active learning with viewpoint entropy for semantic segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 9430−9440.

[17] CAO Xiaofeng. A structured perspective of volumes on active learning[J]. Neurocomputing, 2020, 377: 200–212.

[18] CAO Xiaofeng. A divide-and-conquer approach to geometric sampling for active learning[J]. Expert systems with applications, 2020, 140: 112907.

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

[20] SETTIES B. Curious machines: active learning with structured instances[J]. Journal of chemical information and modeling, 2013, 53(9): 1689–1699.

[21] XIA Chenyi, HSU W, LEE M L, et al. BORDER: efficient computation of boundary points[J]. IEEE transactions on knowledge and data engineering, 2006, 18(3): 289–303.

[22] QIU Baozhi, CAO Xiaofeng. Clustering boundary detection for high dimensional space based on space inversion and Hopkins statistics[J]. Knowledge-based systems, 2016, 98: 216–225.

[23] AGGARWAL C C. An introduction to outlier analysis[M]//Outlier Analysis. Cham: Springer International Publishing, 2016: 1−34.

[24] KONYUSHKOVA K, SZNITMAN R, FUA P. Learning active learning from data[J]. Conference and workshop on neural information processing systems, 2017, 31(12): 4226–4236.

[25] HE Deniu, YU Hong, WANG Guoyin, et al. A two-stage clustering-based cold-start method for active learning[J]. Intelligent data analysis, 2021, 25(5): 1169–1185.

[26] KARAMCHETI S, KRISHNA R, LI Feifei, et al. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 7265−7281.

## 作者简介：

胡峰，教授，博士，主要研究方向为粗糙集、粒计算、数据挖掘。主持和参与国家自然科学基金项目 4 项，参与科技部重点研发计划项目 3 项，作为参与者获吴文俊人工智能科学技术奖、重庆市自然科学奖各 1 项，发表学术论文 40 余篇。E-mail：hufeng@cqupt.edu.cn。

李路正，硕士研究生，主要研究方向为数据挖掘、主动学习。E-mail：isluzheng.li@foxmail.com。

代劲，教授，博士，重庆邮电大学软件学院副院长。主要研究方向为大数据知识工程、智能信息处理。先后承担和完成省部级科研项目 4 项，出版专著 1 部，发表学术论文 20 余篇。E-mail：331545392@qq.com。