计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(10):206−210 [doi: 10.15888/j.cnki.csa.008752]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于层次密度聚类的去噪自适应混合采样①

姜新盈, 王舒梵, 严　涛

(上海工程技术大学 数理与统计学院, 上海 201620)
通信作者: 姜新盈, E-mail: jxynovelty@163.com

摘　要: 针对非平衡数据存在的类内不平衡、噪声、生成样本覆盖面小等问题, 提出了基于层次密度聚类的去噪自适应混合采样算法 (adaptive denoising hybrid sampling algorithm based on hierarchical density clustering, ADHSBHD). 首先引入 HDBSCAN 聚类算法, 将少数类和多数类分别聚类, 将全局离群点和局部离群点的交集视为噪声集, 在剔除噪声样本之后对原数据集进行处理, 其次, 根据少数类样本中每簇的平均距离, 采用覆盖面更广的采样方法自适应合成新样本, 最后删除一部分多数类样本集中的对分类贡献小的点, 使数据集均衡. ADHSBHD 算法在 7 个真实数据集上进行评估, 结果证明了其有效性.

关键词: 不平衡数据; 分类; 聚类; 混合采样

引用格式: 姜新盈,王舒梵,严涛.基于层次密度聚类的去噪自适应混合采样.计算机系统应用,2022,31(10):206–210. http://www.c-s-a.org.cn/1003-3254/8752.html

Denoising and Adaptive Hybrid Sampling Based on Hierarchical Density Clustering

JIANG Xin-Ying, WANG Shu-Fan, YAN Tao

(School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: As imbalanced data are exposed to problems such as intra-class imbalance, noise, and small coverage of generated samples, an adaptive denoising hybrid sampling algorithm based on hierarchical density clustering (ADHSBHD) is proposed. Firstly, the clustering algorithm HDBSCAN is introduced to perform clustering on minority classes and majority classes separately; the intersection of global and local outliers is regarded as the noise set, and the original data set is processed after noise samples are eliminated. Secondly, according to the average distance between clusters of samples in minority classes, the adaptive sampling method with broader coverage is used to synthesize new samples. Finally, some points that contribute little to the classification of majority classes are deleted to balance the dataset. The ADHSBHD algorithm is evaluated on six real data sets, and the results can prove its effectiveness.

Key words: imbalanced data; classification; cluster; hybrid sampling

现实中的很多领域存在的数据集都是不平衡的, 这类数据集有着不同类别数据样本不均、数量相差较大的特点, 其中大多数样本的类别称为多数类别, 其余样本的类别称为少数类别. 由于不平衡数据的广泛存在, 从不平衡数据中学习对于研究界及现实应用都至关重要, 例如疾病诊断[1] 和石油储层含油量识别[2] 等. 任何分类器的目标都是最大程度地提高总体准确性, 但是传统分类器往往更倾向于多数类样本, 这就导致少数类样本分类错误[3]. 实际应用中, 从不平衡数据中学习到的分类器需要同时在不同类别样本上均表现良好, 因此在保证多数类别分类精度的前提下, 如何处理不平衡数据以提高少数类别分类准确性.

1　引言

现有的一些广泛处理不平衡数据分类的方法主要分为数据预处理、算法改进及特征选择这 3 个层面, 当前

① 收稿时间: 2022-01-27; 修改时间: 2022-02-24; 采用时间: 2022-03-09; csa 在线出版时间: 2022-06-24

没有一种方法能够很好地解决所有不平衡数据集分类问题, 算法改进仅仅针对单一的分类器进行改进, 特征选择容易造成信息丢失, 但是数据层面的抽样方法显示了巨大的优越性, 该方法主要是改善数据集本身而不是分类器.

欠采样是对多数类样本进行处理, 选择一些多数类样本进行剔除以提高少数类样本的分类正确率. Yen 等[4] 提出 SBC 算法, 首先将整个数据集聚类, 再根据每簇采样数量进行欠采样, 但是容易丢失关键信息. 过采样是通过合成少数类样本以增加少数类样本的算法, Barua 等[5] 提出 MWMOTE 算法, 先选择少数类样本的适当子集, 根据不同类别样本间的距离对少数类样本分配权重, 再使用聚类方法并结合 SMOTE 算法合成新的少数类样本; Nekooeimehr 等[6] 提出自适应半无监督加权过采样, 先对少数类样本进行分层聚类, 并根据分类复杂度等自适应地对每个子集中靠近边界的少数类样本进行过采样, 避免生成与多数类重叠的合成少数类样本. 石洪波等[7] 研究表明使用单一的采样算法或导致过拟合或误删重要样本, 而文献 [8] 表明混合采样的分类性能比单个采样算法好, 在提高运行效率, 有效避免过拟合问题的情况下, 还不易丢失含有重要信息的多数类样本. 戴翔等[9] 提出 BCS-SK 算法, 采用 SMOTE 合成少数类样本, 然后采用 K-means 聚类算法对多数类样本进行欠采样; 史明华[10] 对整个数据集进行聚类, 根据每簇不平衡率的大小将数据集分为 4 类并采取不同的采样方法, 均衡了簇内的样本分布.

以上算法各具优势, 但大部分没有解决类内小分离及易合成低质量样本的问题, 也没有区分不同样本的重要性, 为解决以上问题, 本文提出了基于层次密度聚类的去噪自适应混合采样算法 (adaptive denoising hybrid sampling algorithm based on hierarchical density clustering, ADHSBHD) 来有效合成高质量样本.

2 相关理论

2.1 HDBSCAN 聚类

HDBSCAN 聚类算法[11] 是 McInnes 等人研究提出的一种基于层次聚类的最新算法, 是对 DBSCAN 密度聚类算法的优化, 能处理不同密度和任意形状的聚类[12], 一方面是不再将 *Eps* 值作为树状图的切割线, 而是通过查看分裂来压缩树状图, 使用该树选择最稳定的簇, 并以不同的高度来切割树, 这样可以根据簇的稳定性选择密度不同的簇; 另一方面是不再需要人工设置 *Eps* 参数, 只需要设置集群最小样本数量 *min_samples* 来确定样本是离群点还是分裂为两个新集群即可.

相比于 K-means、DBSCAN 等常用聚类具有对参数设置不敏感的优势, 通常设置两个参数: 最小集群数量 *min_cluster_size*, 集群最小样本数量 *min_samples*. 此外, HDBSCAN 具有噪声感知能力, −1 表示该样本点不在任何簇内, 并且聚类结果会给数据集中的每个样本都赋予一个 0.0–1.0 的概率值, 用于代表属于某个簇的可能性, 概率值为 0.0 则表示该样本点不在集群中 (所有的离群点都是这个值), 概率值为 1.0 则表示该样本点位于簇的核心.

2.2 Random-SMOTE 算法

由于 SMOTE 算法存在合成样本区域有限、易产生噪声等问题, 董燕杰[13] 提出 Random-SMOTE 算法, 可以提高算法效率并更符合原数据集样本空间. 该算法的核心思想是根据 3 个样本点构成三角区域, 在区域内生成新样本, 首先从少数类数据集中选择样本x作为根样本, 并且随机选择y_1、y_2这两个样本作为间接样本, 根据式 (1) 在y_1、y_2间线性插值, 根据设置的采样倍率N, 生成N个临时样本$p_j, j=1,2,\cdots,N$.

$$p_j = y_1 + rand(0,1)\times(y_2 - y_1), j=1,2,\cdots,N \quad (1)$$

其次根据式 (2) 在p_j和x之间线性插值得到新样本m_j.

$$m_j = x + rand(0,1)\times\left(p_j - x\right), j=1,2,\cdots,N \quad (2)$$

3 ADHSBHD 算法

文献 [6] 表明, 聚类是不错的解决类内不平衡及小分离问题的方法. 为了更深入研究解决噪声、类内不平衡和小分离问题, 提出了基于层次密度聚类的去噪自适应混合采样算法 (ADHSBHD).

3.1 基于 HDBSCAN 的去噪方法

在很多情形之下, 噪声样本会使分类器性能损失, 聚类作为潜在的噪声检测算法, 为识别噪声提供了另一研究思路. 为了有效识别噪声点, 提出了一种基于 HDBSCAN 的去噪方法, 首先对少数类样本集中的每个样本计算其 k 近邻, 若该样本 k 近邻全部是多数类样本, 则把该样本点视为全局离群点. 然后引入 HDBSCAN 聚类, 将少数类样本集进行聚类, 得到若干个簇, 将概率值为 0 的样本视为局部离群点, 取全部离群点集和局部离群点集的交集为噪声集, 这样可以较全面地识别出噪声点, 也避免直接删除处于小分离状态的样本, 为接下来提出的 ADHSBHD 算法在安全区域内生成

样本做准备.

3.2 对少数类样本的处理

若每个簇内都合成同样数量的样本, 那么容易造成类重叠, 也无益于类内不平衡的解决, 因此, 在对少数类样本聚类并去除噪声之后, 需要根据每簇的密集稀疏程度, 确定每个簇所需合成的样本数量, 分配给每个簇一个 0 到 1 的采样权重, 这样可以使稀疏区域的样本增添有益信息, 密集区域的样本尽可能地减少类重叠, 不会忽略对小规模集群的学习. 为了表示每簇的密集稀疏程度, 用平均距离来表示.

定义 1 (平均距离). 对于数据集C和少数类样本的任意簇$C^{(1k)},1\leqslant k\leqslant m$, 簇$C^{(1k)}$的平均距离记为 $Mean\text{-}dist(C^{(1k)})$, 即:

$$Meandist(C^{(1k)})=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}}{2n} \tag{3}$$

其中, d_{ij}为簇内各样本之间的欧式距离, n为簇内样本数量.

定义 2 (采样权重). 任意簇$C^{(1k)},1\leqslant k\leqslant m$的采样权重为某簇的平均距离与所有簇的平均距离之和的比值, 即:

$$W(C^{(1k)})=\frac{Meandist(C^{(1k)})}{\sum_{k=1}^{m}Meandist(C^{(1k)})} \tag{4}$$

当各簇的平均距离越小时, 说明簇内样本更密集, 反之, 则更稀疏. 相应的, 簇内样本越密集时, 需要合成的样本越少, 即采样权重就越小, 反之, 需要合成的样本就越多.

ADHSBHD 算法中新合成的少数类样本数量$\left|C_{\text{new}}^{(1)}\right|$定义如下:

$$\left|C_{\text{new}}^{(1)}\right|=\frac{|C|}{2}-\left|C^{(1)}\right| \tag{5}$$

其中, $|C|$表示原始数据集C的样本数量, $\left|C^{(1)}\right|$表示少数类样本的个数.

每簇$C^{(1k)},1\leqslant k\leqslant m$需要合成的样本数量$\left|C_{\text{new}}^{(1k)}\right|$定义如下:

$$\left|C_{\text{new}}^{(1k)}\right|=\left|C_{\text{new}}^{(1)}\right|\times W(C^{(1k)}) \tag{6}$$

3.3 对多数类样本的处理

根据 HDBSCAN 对多数类样本聚类, 得到若干个簇和离群点, 相应的得到每个簇的概率值矩阵, 当概率值较低时说明样本点越在簇的边缘, 过多的边缘样本会使分类器发生偏向而降低少数类的分类精度. 为此, 将所有多数类样本的概率值按照从小到大的顺序排列, 按照一定规则, 删减的多数类样本数量$\left|C_{\text{re}}^{(0)}\right|$定义如下:

$$\left|C_{\text{re}}^{(0)}\right|=\left|C^{(0)}\right|-\frac{|C|}{2} \tag{7}$$

3.4 ADHSBHD 算法流程

ADHSBHD 算法的具体步骤描述如算法 1.

算法 1. ADHSBHD 算法

输入: 原始数据集C, 近邻参数k, 最小集群数量 *min_cluster_size*, 集群最小样本数量 *min_samples*

输出: 均衡数据集C_{new}

Step 1. 将原始数据集划分为多数类样本集$C^{(0)}$和少数类样本集$C^{(1)}$且$C=C^{(0)}\cup C^{(1)}$且$C^{(0)}\cup C^{(1)}=\emptyset$, 并识别噪声.

Step 1.1. 计算$C^{(1)}$中各个样本x, 计算其k近邻, 若k近邻均为多数类样本, 则将该添加到全局离群点集N_t中.

Step 1.2. 用 HDBSCAN 对$C^{(1)}$聚类, 得到m个相互独立且不同规模的簇$C^{(11)},C^{(12)},C^{(13)},\cdots,C^{(1m)}$和若干个离群点, 并且得到每个簇的样本概率值矩阵$p_{ij},0<i\leqslant m,0<j\leqslant\left|C^{(1i)}\right|$, 而离群点的概率值为 0, 将其添加到局部离群点集N_l中, 则噪声集群$N=N_t\cap N_l$. 从原始数据集C中删去噪声集群N, 得到C', 其样本量$\left|C'\right|=|C|-|N|$.

Step 2. for i=1 to $\left|C^{(1i)}\right|$, 计算每个簇需要合成的新的样本数量并自适应合成.

Step 2.1. 计算每个簇中样本之间的欧氏距离, 并得到各簇的平均距离$Meandist(C^{(1k)})$.

Step 2.2. 计算每簇的采样权重$W(C^{(1k)})$.

Step 2.3. 计算每簇新的样本数量$\left|C_{\text{new}}^{(1k)}\right|$, 并选中簇内样本$x$做目标样本, 从它的$k$近邻中随机选择两个样本$y_1$、$y_2$, 先通过$y_1$、$y_2$随机生成一个辅助样本$a$, 再在目标样本$x$和辅助样本$a$之间线性插值随机生成新样本, 即采用 Random-SMOTE 在三角形区域内随机合成样本, 将合成样本添加到$C_{\text{new}}^{(1k)}$中, 直到新样本数量为$\left|C_{\text{new}}^{(1k)}\right|$.

Step 2.4. 合并每簇的$C^{(1k)}$与$C_{\text{new}}^{(1k)}$, 各簇原始的少数类样本集与合成样本组成新的数据集$C_{\text{new}}^{(1)}$中.

Step 3. 对多数类样本$C^{(0)}$进行处理.

Step 3.1. 对$C^{(0)}$用 HDBSCAN 聚类, 得到n个相互独立且不同规模的簇$C^{(01)},C^{(02)},C^{(03)},\cdots,C^{(0n)}$和$t$个离群点, 并且得到每个簇的样本概率值矩阵$p_{ij},0<i\leqslant n,0<j\leqslant\left|C^{(0i)}\right|$.

Step 3.2. 当$t<\left|C^{(0)}\right|-\frac{|C|}{2}$时, 将$C^{(0)}$中的样本按照概率值从小到大排序, 删去离群点和部分概率值小的点, 共$\left|C^{(0)}\right|-\frac{|C|}{2}$个样本; 当$t\geqslant\left|C^{(0)}\right|-\frac{|C|}{2}$时, 对$C^{(0)}$删除$\left|C^{(0)}\right|-\frac{|C|}{2}$个离群点, 剩余的多数类样本添加到新的数据集$C_{\text{new}}^{(0)}$中.

Step 4. $C_{\text{new}}=C_{\text{new}}^{(0)}\cup C_{\text{new}}^{(1)}$并对分类器进行训练.

4 实验结果与分析

4.1 评价指标

准确率作为分类器评价指标对于非平衡数据有失公平, 为了客观评价分类算法性能好坏, 研究学者常常

根据混淆矩阵引入的概念来评估算法性能. 根据表 1 的混淆矩阵, 引入查全率 (*Recall*)、查准率 (*Precision*)、*F-value* 值、*G-mean* 值、*AUC* 等定义.

表 1 混淆矩阵

类别	预测为正类	预测为负类
实际为正类	*TP*	*FN*
实际为负类	*FP*	*TN*

各定义的公式如下:

$$R=\frac{TP}{TP+FN} \tag{8}$$

$$P=\frac{TP}{TP+FP} \tag{9}$$

$$F\text{-}value=\frac{2PR}{R+P} \tag{10}$$

$$G\text{-}mean=\sqrt{\frac{TP}{TP+FN}}\times\sqrt{\frac{TN}{FP+TN}} \tag{11}$$

此外, *AUC* 作为可视化指标, 是根据引入的 *ROC* 曲线下的面积来表示, 值越大, 说明分类器性能越好. *ROC* 曲线则是以真正例率为纵轴, 以假正例率为横轴绘制的.

4.2 数据集描述

为了验证 ADHSBHD 的有效性, 本文从国际机器学习标准库 UCI 中选取了 Ionosphere、Glass、Abalone、Haberman、Vehicle、Ecoli、Yeast 这 7 组不平衡数据集, 其中, Glass 的少数类特征为“1”; Abalone 的少数类特征为“F”; Vehicle 数据集中, 将第一类视为少数类; Ecoli 的少数类特征为 “om”“omL”和“pp”; Yeast 数据集中将“MIT”视为少数类. 7 组数据集的具体信息如表 2 所示.

表 2 数据集信息

数据集	样本总量	少数类数量	多数类数量	不平衡度
Ionosphere	351	126	225	1.79
Glass	214	70	144	2.06
Abalone	4 177	1 307	2 870	2.20
Haberman	306	81	225	2.74
Vehicle	846	199	647	3.25
Ecoli	336	77	259	3.36
Yeast	1 484	244	1 240	5.08

4.3 实验分析

为了验证本节所提出的 ADHSBHD 算法表现良好, ADHSBHD 算法分别与 SMOTE 算法、ADASYN 算法、Random-SMOTE 算法 (以下简称 R-SMOTE)、SVMSOMTE 算法在这 7 组数据集上做重采样的对比实验, 用 *F-value*、*G-mean*、*AUC* 作为评价指标, 实验环境均在 Jupyter Notebook 中运行, 所使用的对比算法除 R-SMOTE 外均调用 imbalanced-learn 程序包实现.

此外, SVM 中核函数为高斯核函数, 其他超参数均为 imbalanced-learn 中的默认值, 为了直观验证上述各对比采样算法的有效性, 设置 k=5, 相应的算法中的其他超参数也均为默认值, 而 ADHSBHD 算法中使用到的 HDBSCAN 算法设置最小集群数量 *min_cluster_size*=2, 集群最小样本数量 *min_samples*=4.

表 3、表 4 和表 5 展示了 7 组数据集 6 种不同采样算法下的 *F-value* 值、*G-mean* 值和 *AUC* 值以此来衡量算法分类结果, 黑体加粗的数值表示同一数据集的最优算法对应指标值, avg 表示不同数据集在不同组合形式的算法下的平均值. 从各表中可以看出, 虽然对 Haberman 数据集使用 Random-SMOTE 算法的结果优于本文算法结果, 这是由于本文在引入聚类算法后, 在如何选择最佳样本参与合成时存在不足, 而 Random-SMOTE 算法的合成方法更符合 Haberman 数据集. 但是从整体性能上看, ADHSBHD 算法应用到 SVM 分类器后提升了整体分类精度, 在 *F-value*、*G-mean*、*AUC* 这几种性能指标方面均优于文中所提其他 4 种对比算法.

表 3 不同数据集在支持向量机下的 *F-value* 性能对比

数据集	ADHSBHD	SMOTE	ADASYN	R-SMOTE	SVMSOMTE
Ionosphere	**0.980**	0.978	0.974	0.801	0.978
Glass	**0.861**	0.847	0.794	0.857	0.850
Abalone	0.679	**0.686**	0.663	0.551	0.677
Haberman	0.737	0.637	0.695	**0.929**	0.770
Vehicle	**0.973**	0.966	0.963	0.831	0.966
Ecoli	**0.990**	0.941	0.928	0.952	0.955
Yeast	**0.894**	0.797	0.782	0.629	0.862
avg	0.873	0.836	0.828	0.793	0.865

表 4 不同数据集在支持向量机下的 *G-mean* 性能对比

数据集	ADHSBHD	SMOTE	ADASYN	R-SMOTE	SVMSOMTE
Ionosphere	**0.980**	0.977	0.969	0.781	0.977
Glass	**0.869**	0.836	0.765	0.849	0.833
Abalone	0.668	**0.676**	0.650	0.529	0.669
Haberman	0.732	0.638	0.688	**0.928**	0.756
Vehicle	**0.973**	0.963	0.963	0.820	0.965
Ecoli	**0.990**	0.934	0.928	0.951	0.955
Yeast	**0.894**	0.794	0.782	0.628	0.862
avg	0.872	0.831	0.821	0.784	0.860

表 5 不同数据集在支持向量机下的 *AUC* 性能对比

数据集	ADHSBHD	SMOTE	ADASYN	R-SMOTE	SVMSOMTE
Ionosphere	**0.980**	0.977	0.969	0.795	0.977
Glass	**0.863**	0.846	0.793	0.853	0.839
Abalone	0.684	**0.691**	0.671	0.549	0.686
Haberman	0.735	0.644	0.697	**0.929**	0.770
Vehicle	**0.973**	0.964	0.963	0.834	0.965
Ecoli	**0.990**	0.935	0.928	0.952	0.956
Yeast	**0.894**	0.797	0.782	0.629	0.862
avg	0.874	0.836	0.829	0.791	0.865

5 结论

本文提出了基于层次密度聚类的去噪自适应混合采样算法 (ADHSBHD), 以层次密度聚类为基础, 考虑到不同类别样本的样本空间分布, 为了验证 ADHSBHD 的有效性及稳定性, 将该算法与 SVM 分类器结合在一起, 并与 4 种采样算法进行实验对比, 实验结果表明, 该算法与不同的分类器的组合有较好的泛化性, *F-value*、*G-mean*、*AUC* 这 3 个评价指标在上述大部分数据集上都有所提升. 由于该算法是基于聚类算法展开, 后续可以研究其他的聚类技术能否为该算法带来更高性能.

参考文献

1 Bhattacharya S, Rajan V, Shrivastava H. ICU mortality prediction: A classification algorithm for imbalanced datasets. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017. 1288–1294.

2 李诒靖, 郭海湘, 李亚楠, 等. 一种基于 Boosting 的集成学习算法在不均衡数据中的分类. 系统工程理论与实践, 2016, 36(1): 189–199. [doi: 10.12011/1000-6788(2016)01-0189-11]

3 赵楠, 张小芳, 张利军. 不平衡数据分类研究综述. 计算机科学, 2018, 45(6A): 22–27, 57.

4 Yen SJ, Lee YS. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 2009, 36(3): 5718–5727. [doi: 10.1016/j.eswa.2008.06.108]

5 Barua S, Islam MM, Yao X, *et al*. MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 405–425. [doi: 10.1109/TKDE.2012.232]

6 Nekooeimehr I, Lai-Yuen SK. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. Expert Systems with Applications, 2016, 46: 405–416. [doi: 10.1016/j.eswa.2015.10.031]

7 石洪波, 陈雨文, 陈鑫. SMOTE 过采样及其改进算法研究综述. 智能系统学报, 2019, 14(6): 1073–1083.

8 Gazzah S, Hechkel A, Amara NEB. A hybrid sampling method for imbalanced data. Proceedings of the IEEE 12th International Multi-conference on Systems, Signals & Devices (SSD15). Mahdia: IEEE, 2015. 1–6.

9 戴翔, 毛宇光. 基于集成混合采样的软件缺陷预测研究. 计算机工程与科学, 2015, 37(5): 930–936. [doi: 10.3969/j.issn.1007-130X.2015.05.012]

10 史明华. 不平衡分类问题中的去噪混合采样算法研究 [硕士学位论文]. 广州: 华南理工大学, 2020.

11 McInnes L, Healy J. Accelerated hierarchical density based clustering. 2017 IEEE International Conference on Data Mining Workshops (ICDMW). New Orleans: IEEE Press, 2017. 33–42.

12 董宏成, 赵学华, 赵成, 等. 基于 HDBACAN 聚类的自适应过采样技术. 计算机工程与设计, 2020, 41(5): 1295–1300.

13 董燕杰. 不平衡数据集分类的 Random-SMOTE 方法研究 [硕士学位论文]. 大连: 大连理工大学, 2009.

(校对责编: 孙君艳)