计算机科学
COMPUTER SCIENCE

基于反向标签传播的多生成器主动学习算法及其在离群点检测中的应用研究

邢开颜, 陈文

引用本文

邢开颜, 陈文. 基于反向标签传播的多生成器主动学习算法及其在离群点检测中的应用研究[J]. 计算机科学, 2024, 51(4): 359-365.

XING Kaiyan, CHEN Wen. Multi-generator Active Learning Algorithm Based on Reverse Label Propagation and Its Application in Outlier Detection [J]. Computer Science, 2024, 51(4): 359-365.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500034

基于反向标签传播的多生成器主动学习算法及其在离群点检测中的应用研究

邢开颜　陈　文
四川大学网络空间安全学院　成都 610065
(xingkaiyan@stu.scu.edu.cn)

摘　要　当前正负类训练样本分布不均衡的问题已极大地限制了离群检测模型的性能。基于主动学习的离群点检测算法能够通过对样本分布的主动学习,自动合成离群点以平衡训练数据分布。然而,传统的基于主动学习的检测方法缺乏对合成离群点的质量评估和过滤筛选,导致通过主动学习过程合成的训练样本点中存在样本噪声,并降低了分类模型的性能。针对上述问题,提出了基于反向标签传播的多生成器主动学习算法(Multi-Generator Active Learning Algorithm Based on Reverse Label Propagation,MG-RLP),其包括多个神经网络生成器和一个用于离群点边界检测的鉴别器。MG-RLP通过多个子生成器生成多分布特征的样本数据,以防止单生成器合成的训练样本过于聚集而导致的模式崩塌问题。同时,MG-RLP利用反向标签传播过程对神经网络生成的样本点进行质量评估,以筛选出可信的合成样本。筛选后的样本被保留在训练样本中用于对鉴别器进行迭代训练,以提升对离群点的检测性能。基于5个公共数据集,对比验证了MG-RLP与6种典型的离群点检测算法的性能,结果表明,MG-RLP在AUC和检测精度指标上分别提高了15%和22%,结果验证了MG-RLP的有效性。

关键词:离群点检测;主动学习;生成对抗网络;标签传播
中图分类号　TP181

Multi-generator Active Learning Algorithm Based on Reverse Label Propagation and Its Application in Outlier Detection

XING Kaiyan and CHEN Wen
School of Cyber Science and Engineering,Sichuan University,Chengdu 610065,China

Abstract　The current problem of unbalanced distribution of positive and negative training samples has greatly limited the performance of outlier detection models. The outlier detection algorithm based on active learning can automatically synthesize outliers to balance the training data through active learning of sample distribution. However,the traditional detection method based on active learning lacks the quality assessment and filtering of synthetic outliers,which leads to the fact that the noise in the synthetic training samples degrades the performance of classification models. Aiming at the above problems,a multi-generator adversarial learning algorithm based on reverse label propagation(MG-RLP) is proposed,which consists of multiple neural network generators and a discriminator for outlier boundary detection. MG-RLP uses multiple sub-generators to generate sample data with multi-distribution features to prevent the mode collapse problem caused by the excessive aggregation of training samples synthesized by a single generator. At the same time,the proposed method utilizes the reverse label propagation to evaluate the quality of the sample points generated to screen out credible synthetic samples. The filtered samples are retained in the training samples to iteratively train the discriminator to improve the detection performance of outliers. The MG-RLP is compared with six typical outlier detection algorithms on five public datasets. The results show that the proposed algorithm improves AUC and detection precision by 15% and 22% respectively,which verifies its effectiveness.

Keywords　Outlier detection,Active learning,Generative adversarial networks,Label propagation

1　引言

离群点与大多数正常数据在分布形态上存在一定的差异。离群点检测的目的是在隐含的数据分布中检测到存在分布差异的样本,并从中发现潜在的异常信息。当前,离群点检测已经被广泛应用于医疗预测[1]、异常数据检测[2]、故障

到稿日期:2023-05-06　返修日期:2023-09-11
基金项目:国家重点研发计划(020YFB1805405,2019QY0800);国家自然科学基金(U19A2068,61872255)
This work was supported by the National Key Research and Development Program of China(020YFB1805405,2019QY0800) and National Natural Science Foundation of China(U19A2068,61872255).
通信作者:陈文(wenchen@scu.edu.cn)

检测[3]、金融数据分析[4]、图像处理等领域[5]。然而，在实际应用中，离群点分布大多具有不均衡性，因此离群点检测模型训练困难。检测边界往往受采集数量较多的正常样本影响而偏向正类样本，降低了检测准确率。此外，离群点还具有偶然性、多变性、不确定性等特点，因此离群点检测仍然面临着较大的挑战。

根据使用的技术手段的不同，可以将离群点检测方法分为四大类：基于统计的检测算法[6-7]、基于近邻的检测算法[8-10]、基于聚类的检测算法[11-13]，以及基于分类的检测算法[14-19]。基于统计的检测算法需要事先了解被测对象的数据分布[6-7]，而在实际情况中，数据对象的分布往往是未知的。基于近邻的检测算法包括基于距离的检测和基于密度的检测[8-10]，其对参数的设置较为敏感，且难以处理高维数据。基于聚类的检测算法将数据对象划分为不同的簇，通过比较数据对象与簇之间的关系来确定离群点，如 DBSCAN[11]，DENCLU[12]，STING[13]等。基于聚类的检测算法易于实现，但聚类中心数量、聚类半径等参数的选择对算法的性能影响较大。此外聚类算法的时间复杂度较高，难以适用于样本规模较大的应用。基于分类的检测算法通过寻找决策边界来区分正常点与离群点。随着机器学习在各个领域的广泛应用，神经网络[18-19]、向量机模型[20]、决策树[15]等分类模型被广泛应用于离群点检测。然而在离群点检测的实际应用中往往面临着正负样本分布不均衡的问题，影响了检测模型的训练效果。

近年来，有学者提出通过机器学习合成离群点来弥补离群检测中训练样本点数量的不足，防止正负类样本不均衡造成的模型训练偏向性问题产生。Dèsir 等[15]提出了基于随机森林的离群点生成算法，在集成学习中使用随机化原则对特征数量以及训练集进行子采样以生成潜在离群点。Fan 等[16]提出在稀疏区域边缘生成更多的潜在离群点，扩大稀疏区域密度，从而学习到稀疏区域与其他区域的划分边界。Hempstalk 等[17]为目标数据建立参考密度，以此生成潜在异常。

受生成对抗网络(GAN)合成样本的过程启发，Dai 等利用生成器填充特征空间的低密度区域，即在特征空间中生成与真实数据互补的分布，使鉴别器学习低密度区域的分类边界，提高鉴别器分类精度[18]。Liu 等[19]提出了基于 GAN 的多生成器检测模型(MO-GAAL)，将单个生成器扩展为多个生成器，防止单个生成器陷入合成样本过度聚集导致的模式崩塌问题。

然而，上述方法都是基于 GAN，MO-GAAL 等神经网络生成器产生的合成训练样本，忽略了对新生成样本的质量评估，所合成的离群点可能偏离了真实的样本分布：其中可能存在一些噪声训练样本，进而影响了分类器训练，导致分类边界发生错误偏向。因此，需要对合成的离群点进行评估和筛选，以保证合成的离群点质量。

针对上述问题，本文提出了基于反向标签传播的多生成器主动学习算法(Multi-Generator Active Learning Algorithm Based on Reverse Label Propagation，MG-RLP)，并将其应用于离群点检测，本文的主要贡献有以下两个方面：

1)提出了一种基于主动学习的离群点检测模型，利用多个子生成器分布式地学习真实数据在不同特征空间区域的分布，避免了单生成器合成的离群点分布过于集中而导致模型崩塌，保证了样本分布的多样性，提高了鉴别器的分类精度。

2)将标签传播算法与多个子生成器相结合。在主动学习过程中，使用标签传播算法对生成的合成离群点对已知样本点进行反向标签传播。通过标签反向传播的评估结果，筛选出可信的合成离群点参与后续训练，减少了数据噪声，提升了模型的检测精度。

2 相关工作

2.1 生成对抗网络

在离群点检测中，已标记的样本点对于离群点识别训练有至关重要的作用，然而实际情况中通常面临样本不均衡，即正常类样本多、离群点样本少的问题。因此，近年来，研究人员将 GAN 网络引入离群点检测任务中[21-24]，利用其学习未知数据分布，凭借其合成新的样本数据的能力，增加离群点训练样本数量，提升模型的训练性能。

GAN 通常由生成器(G)和鉴别器(D)两部分组成[25]，通过两个神经网络相互对抗博弈进行学习。随机噪声 z 作为生成器的输入，G 学习真实数据的特征分布，尽可能生成与真实样本 x 相似的样本 $G(z)$。鉴别器 D 以真实样本 x 或者生成样本 $G(z)$ 作为输入，判断输入数据是否为真实样本，给出输入数据为真实样本的概率值 $D(x)$。训练过程中，生成器与鉴别器在迭代对抗过程中不断优化调整模型参数，直至达到对抗训练过程的纳什均衡[26]，即鉴别器无法判断输入数据的真实性时，停止训练。GAN 网络优化的目标函数如式(1)所示：

$$\min_G \max_D V(D,G)=\boldsymbol{E}_{x\sim p_{\text{data}}(x)}[\log D(\boldsymbol{x})]+ E_{z\sim P_z(z)}[\log(1-D(G(\boldsymbol{z})))] \tag{1}$$

其中，$\boldsymbol{z}$ 为输入 G 网络的噪声数据，$G(\boldsymbol{z})$ 为生成数据，$D(\boldsymbol{x})$ 为鉴别器判断 $\boldsymbol{x}$ 是否为真实样本的鉴别概率，$p_{\text{data}}(\boldsymbol{x})$ 为真实数据的分布，$P_z(\boldsymbol{z})$ 为噪声 $\boldsymbol{z}$ 的先验分布。

2.2 标签传播算法

Zhu 等[27]提出的标签传播算法(Label Propagation Algorithm)是一种半监督学习方法。LPA 基于样本间的距离相似度，通过迭代的距离计算，将标签从已标记样本传播到未标记样本。根据 LPA 算法，用节点代表样本，节点与节点之间按照样本相似度进行传播，每一个节点标签都会受其相邻节点标签的影响，相邻节点与该节点的相似度越大，对该节点的影响权值越大，相邻节点的标签就越容易传播。每一轮传播结束后，检查初始已标记节点的标签，保持已标记节点的标签不变，进行下一轮传播。迭代多轮，直至标签矩阵收敛或迭代结束，完成标签传播过程。

具体算法如下：$\{c_1,c_2,\cdots,c_t\}$ 为所有标签类别，$\{x_1,x_2,\cdots,x_l\}$ 为已标记数据，其对应的标签分别为 $\{y_1,y_2,\cdots,y_l\}\in\{c_1,c_2,\cdots,c_t\}$，算法假设标签的类别数量 t 已知，且已标记数据的标签包含所有标签类别。$\{x_{l+1},x_{l+2},\cdots,x_{l+u}\}$ 为未标记数据，其对应标签 $\{y_{l+1},y_{l+2},\cdots,y_{l+u}\}$ 也是未知的，$\boldsymbol{X}=\{x_1,x_2,\cdots,x_{l+u}\}\in R^D$，通常情况下 $l\ll u$。首先创建一个含有全部节点的全连接图，任意节点 i 与节点 j 之间的边代表权重 w_{ij}，如式(2)所示：

$$w_{ij}=\exp\left(-\frac{d_{ij}^2}{\sigma^2}\right)=\exp\left(-\frac{\sum_{d=1}^{D}(x_i^d-x_j^d)^2}{\sigma^2}\right) \tag{2}$$

节点与节点之间的欧氏距离d_{ij}越近，则权重w_{ij}越大；σ为权重控制参数。

通过节点之间的边将节点标签传递给其他节点，边的权重越大，标签就越容易传播。文献[27]定义了一个$(l+u)\times(l+u)$的概率传播矩阵$\boldsymbol{T}$，如式(3)所示：

$$T_{ij}=P(j\rightarrow i)=\frac{w_{ij}}{\sum_{k=1}^{l+u}w_{kj}} \tag{3}$$

其中，T_{ij}为节点i到节点j的传播概率。同时定义了$(l+u)\times C$的标签矩阵$\boldsymbol{Y}$，矩阵第i行代表y_i结点的标签概率。

LPA算法传播的具体步骤如下：

1)标签矩阵$\boldsymbol{Y}$通过概率矩阵$\boldsymbol{T}$进行更新：$\boldsymbol{Y}\leftarrow\boldsymbol{TY}$；

2)保持已标记数据的概率分布，将$\boldsymbol{Y}$的前l行重置为真实标签值；

3)重复步骤1)和步骤2)，直至标签矩阵收敛或迭代结束。

LPA算法可以通过少量的标记样本对未标记样本进行预测，其不受样本分布的影响，且逻辑简单，容易实现，已经在多媒体分析[28]和网络社区挖掘[29-30]等领域得到了广泛应用。

3 MG-RLP算法

3.1 算法介绍

为了保证离群点检测的有效性，本文在多目标对抗学习中引入过滤机制，提出了MG-RLP算法。MG-RLP算法主要由k个生成器$Sub_G=\{G_1,G_2,\cdots,G_k\}$、$k$个反向传播区域和一个鉴别器$D$ 3部分组成，如图1所示。MG-RLP算法的基本思想是首先由k个生成器产生k组不同的合成离群点[19]，随后利用LPA算法对生成的每组合成离群点分别进行反向传播，筛选各组的高质量合成离群点，并加入训练数据集。在迭代训练过程中，鉴别器D不断通过新扩充的训练数据集学习正常数据与离群点的划分边界。

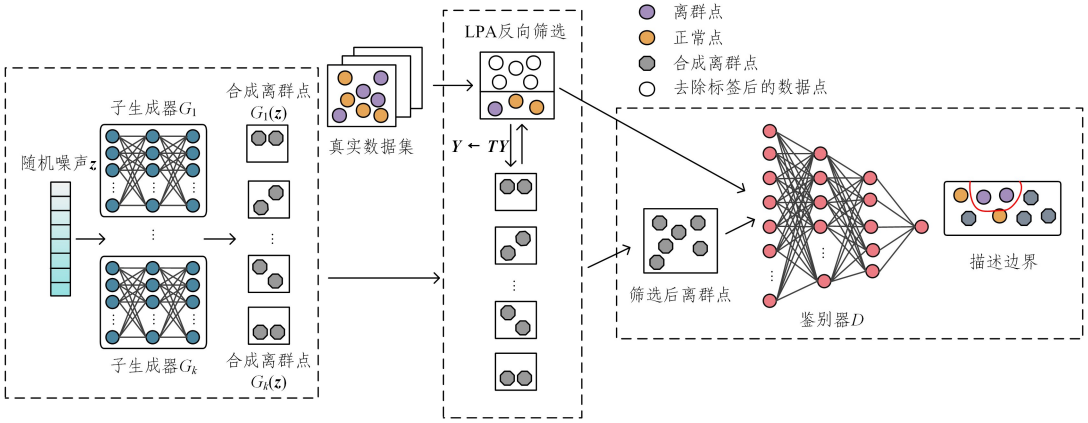

图1 MG-RLP框架图

Fig. 1 Framework of MG-RLP

3.1.1 多生成器

在MG-RLP中，随机噪声z作为各个子生成器G_i的输入从而生成合成离群点，k个子生成器生成不同的合成离群点$G_i(\boldsymbol{z}_i^{(j)};\theta_g)$，其中$\boldsymbol{z}_i^{(j)}$为均匀分布中随机采样得到的噪声变量，$\theta_g$为生成器的参数；鉴别器负责产生正类、负类的划分边界。首先，通过鉴别器的输出，将训练数据划分为k个子集。具体过程为，将鉴别器对真实样本的输出概率按数值大小进行升序排列，并划分为k等份，根据已知样本点的空间取值，选取在输出概率$\{0,1/k\cdots,(k-1)/k\}$位置附近的样本值$\{q_1,q_2,\cdots,q_k\}$作为子生成器的目标值，q_i代表第i个子区域的子生成器目标值。随后，子生成器通过生成与目标子集数据近似的合成离群点，学习目标子集数据的生成机制。

在训练的初始阶段，子生成器生成的合成离群点和目标子集数据相差较远，此时，鉴别器只能描述一个粗糙的边界来区分正常点与生成点。迭代多轮后，子生成器逐渐学习了目标子集数据的潜在分布，生成的合成离群点出现在目标子集数据的周围以及子集数据的内部。此时，经过训练的鉴别器可以描绘一个较准确的边界来区分正常点与生成点。子生成器通过生成距离目标子集数据相近的合成离群点，帮助鉴别器提升划分正常点与生成点的能力，实现了主动学习的过程。通过子生成器和鉴别器网络的对抗训练，子生成器学习到了目标子集数据的深层表示，鉴别器学习到了准确划分正常点和离群点的边界。

子生成器的损失函数如式(4)所示：

$$\min_{\theta_{g_i}}V_{G_i}=-\frac{1}{n_i}\sum_{j=1}^{n_i}[q_i\log(D(G_i(\boldsymbol{z}_i^{(j)})))+(1-q_i)\log(1-D(G_i(\boldsymbol{z}_i^{(j)})))] \tag{4}$$

其中，q_i为子生成器G_i的目标值，n_i为子生成器G_i生成的离群点数量。

鉴别器的损失函数如式(5)所示：

$$\max_{\theta_d}V_D=\frac{1}{n+a}[\sum_{j=1}^{n}\log(D(x^{(j)}))+\sum_{i=1}^{r}\sum_{j=1}^{n_{t_i}}\log(1-D(G_{t_i}(\boldsymbol{z}_i^{(j)})))] \tag{5}$$

其中，a为大于阈值组的合成离群点数量，$a=n_1+n_2+\cdots+n_r$，n_{t_i}为子生成器G_{t_i}生成的离群点数量，n为数据集中数据点的个数。

3.1.2 反向传播

传统的GAN网络中并没有对G生成的数据点进行筛选，生成的数据点中可能存在大量的噪声样本，对鉴别器的训练产生干扰，从而影响鉴别器正确划分边界。

因此，本文通过引入标签传播对数据点进行反向传播，以

此评价生成数据点的质量，从而减少噪声的干扰，传播过程如图 2 所示。

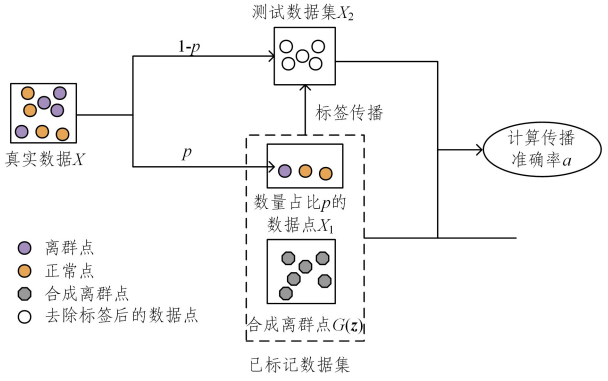

图 2 反向标签传播

Fig. 2 Back label propagation

在反向标签传播中，已标记数据包含两部分数据，一部分为生成器生成的合成离群点$G_i(\boldsymbol{z}_i^{(j)})$，其均被标记为负类；另一部分为从已知标签数据中随机选取数量占比为 p 的数据点X_1，加入已标记数据集。将剩余数量占比为 $1-p$ 的已标记数据去除标签后，作为测试数据集X_2，随后进行标签传播过程，利用 LPA 将标签从$G_i(\boldsymbol{z}_i^{(j)}) \cup X_1$传播到 X_2得到预测标签。传播完成后，计算每个子区域的标签传播准确率$\{a_1, a_2, \cdots, a_k\}$，其中$a_i = n_i / m_i$，其中$n_i$为第 i 个子区域中来自X_2的样本被正确传播的数量，m_i为第 i 个子区域中来自X_2的样本总数。a_i作为衡量合成离群点的质量的指标，用于对新标记样本集$G_i(\boldsymbol{z}_i^{(j)})$进行质量评估。对所有子生成器生成的 n 组合成离群点分别进行上述反向传播过程，并记录准确率。对 k 个区域的准确率进行统计，得到阈值 λ，小于 λ 的合成离群点组被丢弃，大于 λ 的 r 组合成离群点$\{G_{t_1}, G_{t_2}, \cdots, G_{t_r}\}$加入训练数据集，参与后续的训练，其中 $r \leqslant n$，$\{G_{t_1}, G_{t_2}, \cdots, G_{t_r}\} \in SG$。

3.2 算法流程

本文提出的 MG-RLP 算法流程如算法 1 所示。

算法 1 MG-RLP 算法

输入：训练集 X；测试集 T；子生成器个数 k；最小批次 b；标签传播比例 p；最大迭代次数 h

输出：测试集分类结果 R

```
1. 初始化子生成器{G1,G2,…,Gk}和鉴别器 D
2. repeat
3.     for m←1 to b do
4.         从数据集中获取批次数据集xm
5.         从均匀分布中随机采样获得噪声zj
6.         for i←1 to k do
7.             Gi生成合成离群点Gi(z_i^(j))
8.         end for
9.         X1=P*X
10.        X2=(1-P)*X
11.        for i←1 to k do
12.            #反向标签传播 LPA
13.            lpa(Gi(z_i^(j))+X1→X2)
14.            记录准确率ai
15.    end for
16.    统计{a1,a2,…,ak}得到阈值 λ
17.    A={a_t1,a_t2,…,a_tr}&{x∈A | x>λ},{G_t1,G_t2,…,G_tq}加入训练数据集
18.    基于损失函数(5)对鉴别器 D 进行训练
19.    D 进行预测 D(xm)，得到{q1,q2,…,qk}
20.    基于损失函数(4)对子生成器Gi进行训练
21.    end for
22. until 达到最大迭代次数 h
23. 鉴别器输出测试集 T 的分类结果 R
```

MG-RLP 算法中首先初始化子生成器$\{G_1, G_2, \cdots, G_k\}$和鉴别器 D(第 1 行)；k 个子生成器生成不同的合成离群点(第 4－7 行)；根据 3.1.2 小节中提出的 LPA 反向传播对每个区域生成的合成离群点进行置信度评估，若准确率大于阈值 λ，则加入训练集(第 11－16 行)；根据式(5)对鉴别器进行训练(第 18 行)；根据式(4)对子生成器进行训练(第 20 行)；达到迭代次数后停止训练(第 23 行)；输出测试集 T 的分类结果(第 24 行)。

4 实验

4.1 数据集与实验设置

为了验证 MG-RLP 算法的有效性，本文从 DAMI 数据存储库[31]中选取数据集进行实验，此数据库中收集了大量常用于离群点检测研究的相关数据集，同时为了满足离群点检测任务的需求，数据集已被转换为适用于离群点评估的数据集，被广泛应用于各类离群点检测模型的性能验证实验中[32-34]。因此，本文从中选取了在数据分布、特征维度和离群点数量占比上均有差异的 5 个离群点检测数据集进行实验评估。5 个数据集的详细信息如表 1 所列。

表 1 数据集介绍

Table 1 Dataset introduction

数据集	样本数			维度
	总量	正常点	离群点	
Pima	768	500	268	8
Stamps	340	309	31	9
Waveform	3443	3343	100	21
WDBC	367	357	10	30
Ionosphere	351	225	126	32

实验在 Windows11，Python3.7 环境下进行，实验中采用的生成器与鉴别器均采用学习能力相对较弱的三层神经网络结构[19]，以验证基于反向标签传播对合成样本进行质量评估对模型性能持续提升的有效性。子生成器的节点设置为 $l * l * l$，采用 ReLU 激活函数，子生成器的数量设置为 10。鉴别器采用含有一个隐藏层的神经网络结构，节点设置为 $l * \sqrt{n} * 1$，采用 Sigmoid 激活函数，l 为维度大小，n 为训练数据量大小。训练过程中采用 SGD 优化算法，生成器的学习率设置为 0.0001，鉴别器的学习率设置为 0.01。为减少训练过程中出现过拟合现象，设置了生成器训练停止条件：当生成器损失值的下降趋势减缓时，停止训练生成器[19]。实验中训练集与测试集按 8∶2 的比例进行划分，从训练数据中随机抽取 15%数据作为已标记样本，剩余的样本作为无标记样本，参与反向传播过程。

本文选取了 6 个具有代表性的离群点检测算法进行对比，分别为基于距离的检测算法 KNN[36]、基于密度的检测算法 LOF[37]、单分类的检测算法 OC-SVM[20] 和 DeepSVDD[38]、基于对抗网络的检测算法 MO-GAAL[19]，以及基于图神经网络的检测算法 Lunar[39]。本文使用开源工具 Pyod[40] 对 DeepSVDD 和 Lunar 进行评估，MO-GAAL 使用其作者发布的实验系统进行评估，其余均在离群点检测框架 ELKI 下进行评估。

4.2 实验结果分析

离群点检测通常面临着数据不平衡的问题，因此本文采用 5 个常用的评价指标来对不同的检测算法进行综合评估，分别是 AUC(ROC 曲线下面积)、精确率 precision、平均精度 AP(average precision)、召回率 recall 和 F1 分数(F1 Score)。AUC 可以评估分类器对于正类和负类的分类性能，对样本数据不平衡具有很好的鲁棒性。在离群点检测中，我们更关注对少量离群样本的检测，而精确率指标可以通过设定分类阈值，得到预测结果 top-n 中真正为离群点的数量。在某些情况下，我们并不知道离群点的数量，因此，本文还引入了平均精度指标 AP 来评估算法的性能。除此之外，实验过程中还采用了召回率 recall 和 F1 分数指标，以综合评估 MG-RLP 模型的性能。

实验结果如表 2－表 6 所列。

表 2 AUC 实验对比结果

Table 2 AUC experiment comparison results

数据集	KNN	LOF	DeepSVDD	OC-SVM	MO-GAAL	Lunar	MG-RLP
Pima	0.7220	0.6639	0.5235	0.5342	0.7580	0.7147	**0.7757**
Stamps	0.9011	0.7453	0.5907	0.5813	0.7319	0.6561	**0.9623**
Ionosphere	0.9273	0.8866	0.5752	0.9289	0.8697	0.9272	**0.9395**
WDBC	0.9249	0.9039	0.6560	0.8960	0.9817	0.9059	**0.9859**
Waveform	0.7495	0.7224	0.5908	0.5894	0.8474	0.7214	**0.9062**

表 3 Precision 实验对比结果

Table 3 Precision experiment comparison results

数据集	KNN	LOF	DeepSVDD	OC-SVM	MO-GAAL	Lunar	MG-RLP
Pima	0.5597	0.5112	0.3769	0.3808	0.5576	0.5560	**0.5818**
Stamps	0.2581	0.1290	0.0645	0.1935	0.4062	0.2581	**0.7142**
Ionosphere	0.8492	0.7540	0.3333	0.8571	0.7716	0.8571	**0.9230**
WDBC	0.6000	0.6000	0.1000	0.4000	0.6363	0.5000	**0.6666**
Waveform	0.2000	0.1400	0.0800	0.1000	0.1584	0.1600	**0.2857**

表 4 AP 实验对比结果

Table 4 AP experiment comparison results

数据集	KNN	LOF	DeepSVDD	OC-SVM	MO-GAAL	Lunar	MG-RLP
Pima	0.5235	0.4637	0.3706	0.3779	0.5731	0.5237	**0.6179**
Stamps	0.3355	0.2187	0.0997	0.1391	0.3531	0.2414	**0.6273**
Ionosphere	0.9299	0.8232	0.4335	0.9257	0.8342	0.9225	**0.9853**
WDBC	0.5953	0.6437	0.1137	0.3700	0.5887	0.4536	**0.9999**
Waveform	0.1257	0.0784	0.0528	0.0479	0.1113	0.1173	**0.3632**

表 5 Recall 实验对比结果

Table 5 Recall experiment comparison results

数据集	KNN	LOF	DeepSVDD	OC-SVM	MO-GAAL	Lunar	MG-RLP
Pima	0.1455	0.1343	0.3731	0.1492	0.6380	0.1716	**0.6481**
Stamps	0.2903	0.1612	0.1290	0.3870	0.3225	0.2580	**0.8333**
Ionosphere	0.2777	0.2619	0.4365	0.2777	0.6984	0.2778	**0.9200**
WDBC	**0.8000**	**0.8000**	0.1000	**0.8000**	**0.8000**	0.7000	0.5000
Waveform	0.3600	0.3100	0.0700	0.1500	0.1400	**0.3700**	0.1500

表 6 F1 实验对比结果

Table 6 F1 experiment comparison results

数据集	KNN	LOF	DeepSVDD	OC-SVM	MO-GAAL	Lunar	MG-RLP
Pima	0.2260	0.2086	0.3731	0.2318	0.6368	0.2666	**0.6422**
Stamps	0.2769	0.1538	0.1290	0.3692	0.3174	0.2461	**0.7692**
Ionosphere	0.4347	0.4099	0.4365	0.4347	0.6956	0.4348	**0.9019**
WDBC	0.3404	0.3404	0.1000	0.3404	**0.7619**	0.2979	0.4000
Waveform	0.1618	0.1393	0.0700	0.0674	0.1393	**0.1663**	0.1463

表 2－表 6 列出了 6 种典型的离群点检测算法和本文提出的 MG-RLP 算法在不同数据集下的实验结果对比，最好的结果用粗体标出，次好结果用下划线标出。实验结果表明，针对 5 个数据集，本文算法除了在 WDBC 和 Waveform 数据集上的 Recall 和 F1 指标上有所下降外，在其余指标上均表现最优。在 AUC 指标上的对比结果中，本文提出的算法在 Pima

数据集上相比其他算法提升了1%～25%；在Stamps数据集上提升了6%～38%；在Ionosphere数据集上提升了1%～36%；在WDBC数据集上提升了0.42%～32%；在Waveform数据集上提升了5.88%～31%。结果显示在小型数据集上，MG-RLP算法的效果提升更明显。这是因为在反向LPA过程中，小型数据集可以获得更好的评估和筛选，从而挑选出更符合现实数据集分布的合成离群点，因此提升效果更加显著。同时，MG-RLP在precision和AP上也表现得更好。由表5—表6可以看出，在recall和F1指标上，MG-RLP算法对比同类算法在WDBC和Waveform数据集上性能有所下降，但在其余3个数据集上仍是最优。这是因为其在反向LPA过程中，为获取高质量合成离群点，过滤了合成离群点中的样本噪声，但同时部分分布接近于噪声的有效样本也被过滤了，导致模型偏向于高质量样本(典型离群点)，从而出现了部分样本分布过拟合的现象。针对该问题，可根据训练过程动态调整反向标签传播样本的筛选阈值λ，以提高模型的泛化能力。

本文选取了不同比例(15%，10%，5%)的已标记样本进行实验，并统计AUC结果，统计结果如图3所示。结果表明，尽管使用不同比例的已标记数据进行反向传播，MG-RLP的AUC仍然高于其他算法。对实验结果进行分析可以发现，相较于其他模型，由于MG-RLP使用多个子生成器，合成了更多接近真实样本的离群点，且通过反向的LPA筛选过程过滤了数据噪声，保证了样本质量，提升了模型的检测精度。实验结果证实了MG-RLP对离群点检测的有效性。

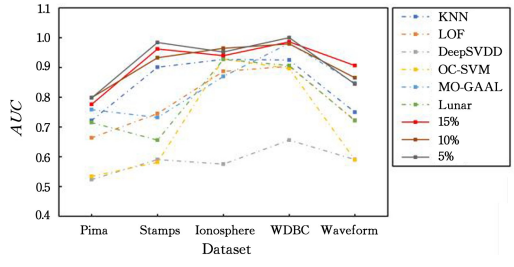

图3 不同数据集下各算法性能对比

Fig. 3 Performance comparison of each algorithm on different datasets

结束语 本文对主流的离群点检测算法进行了总结与分类，提出了一种基于主动学习的离群点检测模型MG-RLP，该算法利用多个子生成器生成多样化的合成离群点，防止生成器产生模式崩溃问题。同时，本文利用反向标签传播对离群点进行评估和筛选，避免训练数据中含有噪声数据，从而提高分类精度。在公开数据集上的多组对比实验结果表明，MG-RLP算法的整体表现优于同类算法，通过多子生成器分布式产生离群点以及基于反向标签传播的样本筛选能够有效提升离群点检测的准确性。同时，我们也注意到基于反向标签传播的样本筛选在保证合成样本质量的同时，也可能会出现模型对部分样本过拟合的情况。在后继的研究中，可以通过动态调整筛选阈值、随机筛选保留等机制进一步提升模型的学习能力。

参 考 文 献

[1] YANG Y, FAN C J, CHEN L, et al. IPMOD: An efficient outlier detection model for high-dimensional medical data streams[J]. Expert Systems with Applications, 2022, 191: 116212.

[2] BEULAH J R, PUNITHAVATHANI D S. An efficient mixed attribute outlier detection method for identifying network intrusions[J]. International Journal of Information Security and Privacy(IJISP), 2020, 14(3): 115-133.

[3] SU Y, ZHAO Y, SUN M, et al. Detecting outlier machine instances through gaussian mixture variational autoencoder with one dimensional cnn[J]. IEEE Transactions on Computers, 2021, 71(4): 892-905.

[4] HILAL W, GADSDEN S A, YAWNEY J. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances [J]. Expert System with Application, 2022, 193: 116429.

[5] BERGMANN P, BATZNER K, FAUSER M, et al. The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection[J]. International Journal of Computer Vision, 2021, 129(4): 1038-1059.

[6] VINUE G, EPIFANIO I. Robust archetypoids for anomaly detection in big functional data[J]. Advances in Data Analysis and Classification, 2021, 15: 437-462.

[7] WILLIAMS J, HILL R R, PIGNATIELLO JR J J, et al. Wavelet analysis of variance box plot[J]. Journal of Applied Statistics, 2022, 49(14): 3536-3563.

[8] YANG J, CHEN Y, RAHARDJA S. Neighborhood representative for improving outlier detectors[J]. Information Sciences, 2023, 625: 192-205.

[9] LI K, GAO X, FU S, et al. Robust outlier detection based on the changing rate of directed density ratio[J]. Expert Systems with Applications, 2022, 207: 117988.

[10] MUHR D, AFFENZELLER M. Little data is often enough for distance-based outlier detection[J]. Procedia Computer Science, 2022, 200: 984-992.

[11] PEŁKA M. Outlier Identification for Symbolic Data with the Application of the DBSCAN Algorithm[C]// Modern Classification and Data Analysis: Methodology and Applications to Micro- and Macroeconomic Problems. Cham: Springer International Publishing, 2022: 53-62.

[12] HINNEBURG A, KEIM D A. An efficient approach to clustering in large multimedia databases with noise[M]. Bibliothek der Universität Konstanz, 1998.

[13] WANG W, YANG J, MUNTZ R. STING: A statistical information grid approach to spatial data mining[C]// VLDB. 1997: 186-195.

[14] ALAVERDYAN Z, JUNG J, BOUET R, et al. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening[J]. Medical Image Analysis, 2020, 60: 101618.

[15] DÉSIR C, BERNARD S, PETITJEAN C, et al. One class random forests[J]. Pattern Recognition, 2013, 46(12): 3490-3506.

[16] FAN W, MILLER M, STOLFO S, et al. Using artificial anomalies to detect unknown and known network intrusions[J]. Knowledge and Information Systems, 2004, 6: 507-527.

[17] HEMPSTALK K,FRANK E,WITTEN I H. One-class classification by combining density and class probability estimation [C]// Machine Learning and Knowledge Discovery in Databases:European Conference,ECML PKDD 2008,Antwerp,Belgium. Berlin Heidelberg:Springer,2008:505-519.

[18] DAI Z,YANG Z,YANG F,et al. Good semi-supervised learning that requires a bad GAN[J]. arXiv:1705.09783,2017.

[19] LIU Y,LI Z,ZHOU C,et al. Generative adversarial active learning for unsupervised outlier detection[J]. IEEE Transactions on Knowledge and Data Engineering,2019,32(8):1517-1528.

[20] SCHÖLKOPF B,PLATT J C,SHAWE-TAYLOR J,et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation,2001,13(7):1443-1471.

[21] SCHLEGL T,SEEBÖCK P,WALDSTEIN S M,et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery[C]// Information Processing in Medical Imaging:25th International Conference,IPMI 2017,Boone,NC,USA. Cham:Springer International Publishing,2017:146-157.

[22] ZENATI H,FOO C S,LECOUAT B,et al. Efficient gan-based anomaly detection[J]. arXiv:1802.06222,2018.

[23] AKCAY S,ATAPOUR-ABARGHOUEI A,BRECKON T P. Ganomaly: Semi-supervised anomaly detection via adversarial training[C]// Computer Vision—ACCV 2018:14th Asian Conference on Computer Vision,Perth,Australia,Revised Selected Papers,Part III 14. Springer International Publishing,2019:622-637.

[24] DEECKE L,VANDERMEULEN R,RUFF L,et al. Image anomaly detection with generative adversarial networks[C]// Machine Learning and Knowledge Discovery in Databases:European Conference,ECML PKDD 2018,Dublin,Ireland. Springer International Publishing,2019:3-17.

[25] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial networks[J]. Communications of the ACM,2020,63(11):139-144.

[26] HEUSEL M,RAMSAUER H,UNTERTHINER T,et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. arXiv:1706.08500,2017.

[27] ZHU X,GHAHRAMANI Z. Learning from labeled and unlabeled data with label propagation [J]. Tech Report,2002,3175(2004):237-244.

[28] BADRINARAYANAN V,GALASSO F,CIPOLLA R. Label propagation in video sequences[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE,2010:3265-3272.

[29] XIE J,SZYMANSKI B K. Community detection using a neighborhood strength driven label propagation algorithm[C]// 2011 IEEE Network Science Workshop. IEEE,2011:188-195.

[30] WU Z H,LIN Y F,GREGORY S,et al. Balanced multi-label propagation for overlapping community detection in social networks[J]. Journal of Computer Science and Technology,2012,27(3):468-479.

[31] CAMPOS G O,ZIMEK A,SANDER J,et al. On the evaluation of unsupervised outlier detection:measures,datasets,and an empirical study[J]. Data Mining and Knowledge Discovery,2016,30:891-927.

[32] HAN S,HU X,HUANG H,et al. Adbench:Anomaly detection benchmark[J]. Advances in Neural Information Processing Systems,2022,35:32142-32159.

[33] LI Z,ZHAO Y,BOTTA N,et al. COPOD:copula-based outlier detection[C]// 2020 IEEE International Conference on Data Mining(ICDM). IEEE,2020:1118-1123.

[34] XU H,PANG G,WANG Y,et al. Deep isolation forest for anomaly detection[J]. IEEE Transactions on Knowledge and Data Engineering,2023,35(12):12591-12604.

[35] RAMASWAMY S,RASTOGI R,SHIM K. Efficient algorithms for mining outliers from large data sets[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. 2000:427-438.

[36] BREUNIG M M,KRIEGEL H P,NG R T,et al. LOF:identifying density-based local outliers[C]// Proceedings of the 2000 ACM Sigmod International Conference on Management of Data. 2000:93-104.

[37] RUFF L,VANDERMEULEN R,GOERNITZ N,et al. Deep one-class classification[C]// International Conference on Machine Learning. PMLR,2018:4393-4402.

[38] GOODGE A,HOOI B,NG S K,et al. Lunar:Unifying local outlier detection methods via graph neural networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022:6737-6745.

[39] ZHAO Y,NASRULLAH Z,LI Z. Pyod:A python toolbox for scalable outlier detection[J]. arXiv:1901.01588,2019.

XING Kaiyan, born in 1999, postgraduate. Her main research interests include machine learning and data mining.

CHEN Wen, born in 1983, Ph.D, associate professor, Ph.D supervisor. His main research interests include network security and data mining.

(责任编辑:何杨)