DOI:10.19651/j.cnki.emt.2417776

DenseNet 特征分组深度孤立森林图像异常检测*

周训会[1,2] 黄成泉[1,2] 肖洪湖[1,2] 董红来[1,2]
(1.贵州民族大学数据科学与信息工程学院 贵阳 550025;2.贵州民族大学工程技术人才实践训练中心 贵阳 550025)

摘 要:为了拓宽深度孤立森林(DIF)算法的应用领域。本文将深度学习预训练 DenseNet-121 模型和 DIF 算法相结合,提出了一种 DenseNet 深度孤立森林(DDIF)算法用于探索该方法在工业图像异常检测数据集 MVTec AD 上的应用效果。但是经 DenseNet-121 模型特征提取后特征向量维度相当高,在随机选择数据属性构建树时可能存在数据集中某些重要特征属性无法被选中的问题,因此本文又提出一种基于特征分组深度孤立森林(GDIF)算法并用在表格型数据集上。最后,在 DDIF 算法的基础上结合 GDIF 算法得到 DenseNet 特征分组深度孤立森林算法(DGDIF),解决了高维数据重要特征漏选问题。实验选取不同的数据集进行异常检测,发现 DDIF 方法在 15 个图像数据集中有 9 个优于其他基于深度学习的方法;GDIF 方法在 9 个表格数据集中较其他传统经典的异常检测算法表现出更优的 AUROC 值;DGDIF 方法在 15 个图像数据集中有 9 个优于不引用特征分组的 DDIF 方法。实验结果验证了所提出的 GDIF 算法,DDIF 算法和 DGDIF 算法的有效性。

关键词:异常检测;特征分组;DenseNet-121;深度孤立森林
中图分类号:TN014 **文献标识码**:A **国家标准学科分类代码**:510.50

DenseNet feature grouping deep isolated forest for image anomaly detection

Zhou Xunhui[1,2] Huang Chengquan[1,2] Xiao Honghu[1,2] Dong Honglai[1,2]
(1. School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China;
2. Engineering Training Center, Guizhou Minzu University, Guiyang 550025, China)

Abstract: In order to broaden the application field of Deep Isolation Forest (DIF) algorithm. we combine the deep learning pre-training DenseNet-121 model and DIF algorithm, and proposes a DenseNet Deep Isolation Forest (DDIF) algorithm for exploring the effectiveness of the method on the industrial image anomaly detection dataset MVTec AD. However, the dimension of feature vector after feature extraction by DenseNet-121 model is quite high, and there may be the problem that some important feature attributes in the dataset cannot be selected when randomly selecting data attributes to construct the tree, so we also propose a Group Deep Isolation Forest (GDIF) algorithm and applies it to tabular datasets. Finally, based on the DDIF algorithm and combined with the GDIF algorithm, the DenseNet Group Deep Isolation Forest (DGDIF) algorithm is obtained, which solves the problem of missing important features in high-dimensional data. Different datasets were selected for anomaly detection, and it was found that the DDIF method outperforms other deep learning-based methods in 9 out of 15 image datasets; the GDIF method showed better AUROC values than other traditional classical anomaly detection algorithms in the 9 tabular datasets; and the DGDIF method outperforms the DGDIF method in 15 image datasets by 9 outperforms the DDIF method without referencing feature grouping. The experimental results validate the effectiveness of the proposed GDIF algorithm, DDIF algorithm and DGDIF algorithm.

Keywords: anomaly detection; feature grouping; DenseNet-121; deep isolation Forest

0 引 言

异常检测是指识别与预期模式或行为显著不同的数据点的过程,用于发现不寻常的事件、行为或趋势。根据数据类型的不同(例如表格型、文本、时间序列和图像等),异常检测方法可以分为多种类别,分别应用于不同的数据类型,

收稿日期:2024-12-31
*基金项目:国家自然科学基金(62062024)、贵州省模式识别与智能系统重点实验室 2022 年度开放课题(GZMUKL[2022]KF03)、贵州省教育厅自然科学研究项目(黔教技[2022]015)资助

自然语言处理法(natural language processing,NLP)[1]、情感分析法[2]等可以对文本数据进行异常检测;季节性分解和滑动窗口等方法可以对时间序列数据进行异常检测,杨晨龙等[3]提出的一种新的 STGAD 异常检测方法还可以解决现有多元时间序列存在的一些问题;常兴亚等[4]研究视频异常检测;卷积神经网络(convolutional neural networks,CNN)和生成对抗网络(generative adversarial network,GAN)架构常用来检测图像异常;基于距离的密度峰值聚类(density peak clustering,DPC)[5]、基于聚类的 K-均值聚类[6]和基于密度的聚类(density-based spatial clustering of applications with noise,DBSCAN)[7]、基于密度的局部离群因子(local outlier factor,LOF)[8]和一种新的数据流异常检测算法提取局部异常值因子(efficient local outlier factor,eLOF)[9]、基于树的孤立森林(isolation forest,IF)[10]等多种方法常用来检测表格型数据;其中,IF 是一种传统的经典无监督异常检测算法,最初由 Liu 等[10]提出并广泛应用,近年来,研究者们对该传统孤立森林进行了许多改进和应用,相继提出扩展孤立森林(extended isolation forest,EIF)[11]、深度孤立森林(deep isolation forest,DIF)[12]、分层孤立森林(layered isolation forest,LIF)[13]等改进算法,逐步弥补传统孤立森林算法的缺点,更加凸显出孤立森林在异常检测方面的优点,也将孤立森林与其他方法结合,提出许多具有创新性的方法和应用[14-18]。

尽管孤立森林在异常检测中表现出诸多优势,但其仍存在一些显著局限性。研究表明,无论是传统孤立森林算法还是其他一系列拓展方法,都依赖随机特征选取的机制,这使得部分重要特征可能在树的构建过程中被忽略,特别是在高维数据场景中,不同特征之间的强相关性往往得不到充分利用,从而导致模型性能下降。此外,孤立森林主要设计用于处理表格型数据,其在图像数据的异常检测方面的研究尚处于起步阶段。现有文献表明,DIF 算法虽已在时间序列和图数据中得到扩展[12],但并未探究该方法在图像异常检测领域的应用。然而,图像异常检测无疑是当前研究的热点之一,尤其在工业监控、医学影像分析、安全监控等领域具有重要意义。同时,图像异常检测面临多重挑战,例如异常样本通常极其稀少,使得模型难以在训练过程中捕捉足够的信息。此外,图像数据具有高维度和复杂的分布特点,传统基于统计的异常检测方法在应对这些问题时表现出局限性。为此,深度学习技术近年来逐渐成为研究热点,通过深度神经网络提取数据的高维嵌入表示,为图像异常检测提供了全新且有效的新思路。

基于上述背景,本文旨在进一步拓展 DIF 算法的应用领域,针对其在高维特征处理和图像数据异常检测中的不足提出改进。在现有研究基础上,本文将 DIF 算法与 CNN 结合,提出 DenseNet 深度孤立森林(DenseNet deep isolation forest,DDIF)算法,以适应图像数据的特征提取需求,使深度孤立森林适应图像数据;并针对特征提取后存在的特征漏选问题提出基于特征分组的深度孤立森林(group deep isolation forest,GDIF)优化算法,通过结合固定特征组与随机特征组的策略,增强模型对特征间关联的捕捉能力;最终,本文设计了结合特征分组的 DenseNet 特征分组深度孤立森林(DenseNet group deep isolation forest,DGDIF)算法,通过实验验证特征分组算法在应对高维数据重要特征遗漏问题上的有效性,同时探索其在图像异常检测领域的适用性及性能表现。本文的研究不仅填补了孤立森林算法在图像异常检测领域的研究空白,也为解决图像数据中的高维问题提供了新的思路和实践依据未来将拓展到更多异常图像数据集验证其适用性。

1　相关理论及方法

本文的异常检测总体框架如图 1 所示,主要包括特征分组和特征提取,其中特征提取主要是通过 DenseNet 预训练模型。

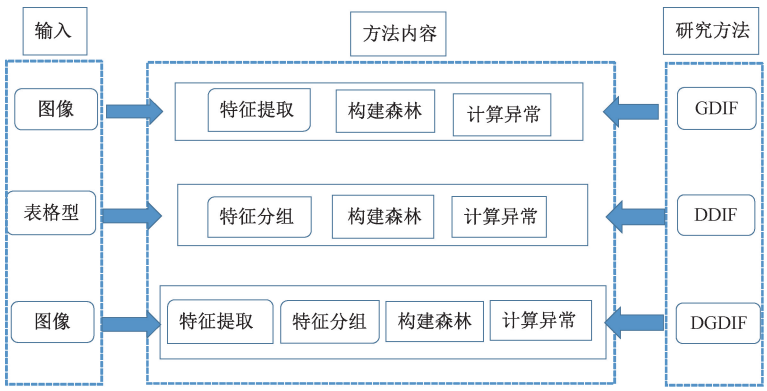

图 1　异常检测框架及模块

Fig. 1　Anomaly detection framework and modules

1.1　DenseNet 原理

DenseNet 是一种深度学习模型,是标准卷积神经网络架构的一种,由 Huang 等[19]在 2017 年提出,其设计目标在于增强信息流动与特征复用。DenseNet 的核心理念是

通过密集连接,将每一层与所有前序层相连,形成一个紧密的网络结构。DenseNet 引入由卷积和池化操作组成的过渡层,用于在不同的密集块之间调整特征图大小和数量。在每个密集块中,每一层接收所有前序层的输出作为输入,这种连接方式不仅使每层都能利用丰富的特征信息,还显著改善了梯度传播问题。这一设计在优化计算资源的同时,还大幅提升了模型的表达能力。

这里选取具有出色的特征学习能力的 DenseNet121 作为特征提取器,具体连接方式如图 2 所示,相比其他版本(如 DenseNet169 或 DenseNet201),DenseNet121 的层数较少,因此计算开销和内存消耗更低,适合在资源受限的环境中使用。在异常检测任务中,预训练的 DenseNet121 模型通过移除最后一层分类器,仅保留特征提取部分,专注于图像的高级特征提取。其输出为高维特征向量,能够有效捕捉复杂图像的高级特征,从而提高异常检测的准确性。

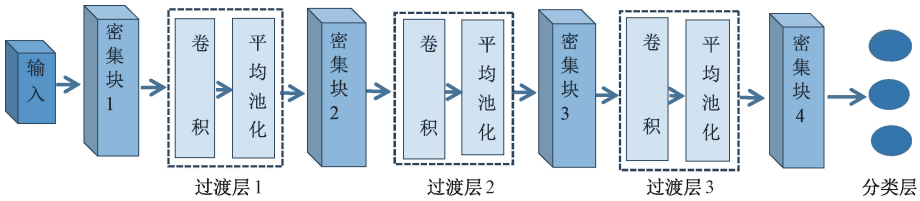

图 2 DenseNet-121 架构

Fig. 2 DenseNet-121 architecture

1.2 DIF 算法原理

DIF 是一种基于决策树集成思想的异常检测算法。在构造决策树之前,引入无优化自由的神经网络生成随机表示,这种随机表示范式中的随机性支撑了在原始数据空间中进行划分的高度自由性,随后用这些随机表示对数据进行分区后构造决策树,每棵决策树都是由训练集中的数据点构造而成。在树的每个节点上,从特征子集中随机选择一个属性特征,然后在该特征的最小值和最大值之间选择一个随机值在该节点上进行分割。总共构造了 T 棵决策树以组成森林来检测异常,数据对象 O 在森林 T 中所有树上的平均路径长度为 $E(|p(x_u|\tau_i)|)$,数据对象 O 的偏差增强隔离异常评分函数定义为:

$$S(O,T)=2^{-E(|p(x_u|\tau_i)|)/c(T)}\times E(g(x_u|\tau_i)) \tag{1}$$

$$g(x_u|\tau_i)=\frac{1}{|p(x_u|\tau_i)|}\sum_{k\in p(x_u|\tau_i)}|x_u^{j_k}-\eta_k| \tag{2}$$

$$C(T)=2H(T-1)-(2(T-1)/T) \tag{3}$$

$$H(T)=\ln(T)+\gamma \tag{4}$$

其中,$g(x_u|\tau_i)$ 表示 τ_i 中 x_u 的平均偏差度,$H(T)$ 表示谐波数,可使用式(4)计算。欧拉常数 γ 近似等于 0.577 215 665[9]。为了对所有树上的平均路径长度进行归一化,该算法使用归一化因子 $C(T)$。得分 S 在(0,1)之间,越接近 1,认为异常程度越高。因此,异常数据点在某些特征上具有极值,可以很容易地将其隔离,具有更短的路径长度。

1.3 DDIF 算法

由于深度孤立森林算法有诸多优势,且已在表格型、时间序列和图结构数据 3 种数据上验证了有效性,因此,本文提出了 DDIF 算法,结合 DenseNet121 和随机神经网络,探索深度孤立森林算法在图像领域的异常检测效果,该算法将深度孤立森林与深度神经网络提取的高维图像特征相结合,在特征空间中生成非线性切割,使其能够用于图像异常检测领域,从而增强了孤立森林对高维图像数据的异常检测能力。

$$\boldsymbol{x}_i=f_{DenseNet121}(y_i) \tag{5}$$

$$p(x_u|\tau_i)=\phi(\boldsymbol{x}_i) \tag{6}$$

具体而言,首先对原始图像数据集 $\{y_i\}_{i=1}^N$ 进行预处理,使用预训练的 DenseNet121 网络提取图像特征 $\{x_i\}_{i=1}^N$,其中 $\boldsymbol{x}_i$ 表示通过深度卷积神经网络映射得到的特征向量,然后通过随机初始化的无优化神经网络 $\phi(\cdot)$ 对特征 x_i 进行非线性映射,经过式(6)生成新的特征表示,多样化的表示能够更好地捕捉图像数据中的非线性模式,从而提高对复杂异常的检测能力。其中 $\phi(\cdot)$ 的权重矩阵初始化为随机分布,用于产生不同的随机投影。在特征空间 $\{p(x_u|\tau_i)\}$ 上,利用简单的轴平行分割来构建孤立树,并利用式(1)计算异常分数。

1.4 GDIF 算法

将 DIF 算法拓宽到图像异常检测领域时,先通过预训练 DenseNet121 特征提取,然而,经过特征提取后图像的特征向量 $\boldsymbol{x}_i$ 的维度相当高,在这种高维特征中,部分特征之间可能存在强相关性,随机选择策略可能忽视这些特征间的依赖关系或者数据集中某些重要特征属性无法被选中,导致性能下降。为解决此问题,本文又提出了 GDIF 优化算法,通过固定特征组和随机特征组的组合,提升模型对特征间关联的捕捉能力,作为 DIF 算法的补充。

1)特征分组策略

首先将高维特征划分为多个组,每组包含相邻的特征。组内特征可能具有较强相关性,有助于模型学习局部模式。

假设数据集 $X\in R^{n\times d}$,其中 n 为样本数,d 为特征数。将特征划分为 G 组,每组包含的特征数量为 $g=[\frac{d}{G}]$。按照索引顺序将特征划分为 G 组:

$$G=\{\{x_1,x_2,\cdots,x_g\},\{x_{g+1},x_{g+2},\cdots,x_{2g}\},\cdots,\{x_{(G-1)g+1},x_{(G-1)g+2},\cdots,x_d\}\} \tag{7}$$

这种策略确保每个特征组的特征数量尽可能均匀，且分组间特征不重叠，使得模型在面对高维数据时，能够更好地捕捉数据的内在结构，可以降低深度孤立森林在高维数据中随机选择特征导致的性能不稳定问题，提升了对复杂数据分布的适应性和检测精度。

2)特征选择与模型训练

在每次训练过程中，为了既保留特定特征组的信息，又引入随机性，本文采用“固定＋随机”策略，确保模型能够学习该组特征的特定信息，同时，从剩余特征中随机选择若干特征，提供额外的随机性，增强模型的泛化能力。选择当前的特征组 $F_{fixed}\in G$，从剩余特征中随机选择特征 $F_{random}\in d$，再将 F_{fixed} 与 F_{random} 组合作为当前训练的输入特征 F_G。对于每个特征组合 F_G，训练一个深度孤立森林模型 M_i，计算决策分数。这种组合方式不仅保留了固定特征的稳定性，还引入了随机特征的多样性，提升了模型的泛化能力。

3)DGDIF 算法

最后，本文在 DDIF 算法的基础上结合 GDIF 算法引出 DGDIF 算法，在图像异常检测领域验证其解决高维数据重要特征漏选方面的有效性，先对图像进行特征提取，再通过特征分组后输入到 DIF 算法检测异常，整个过程中不仅充分利用了神经网络的表示学习能力探索 DIF 算法在图像异常检测领域的适应性，还利用特征分组方法解决了高维特征选择中的有效性，希望提高图像异常检测中的检测效果。

2　实验数据及评价指标

2.1　数据集特征

为了验证所提方法的有效性，本文采用公开可用的真实世界数据集进行评估，包括来自公开可用 UCI 机器学习存储库的 9 个表格数据集，详细信息如表 1 所示。

表 1　真实世界基准数据集属性

Table 1　Properties of the real-world benchmark dataset

数据集	样本量	特征数	异常比/%
Creditcard	284 807	30	0.17
Pageblocks	5 393	10	9.46
Shuttle	1 013	9	1.28
Pima	768	8	35.00
Breastw	569	30	37.26
Ionosphere	351	33	36.00
Arrhythmia	452	268	45.80
Thyroid	3 772	6	7.53
Diabetes	768	8	34.80

MVTec AD 数据集里面包含多种工业场景下的图像数据，如瓶子、胶囊、金属零件等 15 种类型(5 类纹理图像和 10 类物体图像)，共计 5 354 张图像，含盖了正常样本和多种不同类型的异常样本。本实验将数据集分为两类，即正常样本和异常样本，具体信息如表 2 和图 3(a)、(b)所示。

表 2　MVTec AD 数据集特征

Table 2　Characteristics of MVTec AD dataset

数据集	样本量	异常数	数据集	样本量	异常数
bottle	292	63	zipper	391	119
cable	374	92	tile	347	84
capsule	351	109	grid	342	57
toothbrush	102	30	leather	369	92
transistor	313	40	wood	326	60
hazelnut	501	70	carpet	397	89
metal_nut	335	93	—	—	—
pill	434	141	—	—	—
screw	480	119	—	—	—

(a) 榛子正常与部分异常样例

(a) Normal and partially abnormal samples of hazelnut

(b) 皮革正常与部分异常样例

(b) Normal and partially abnormal samples of leather

图 3　MVTec AD 正常与异常样例

Fig. 3　MVTec AD normal and abnormal samples

2.2　评价指标

本文采用异常检测的主流评价方案 ROC 曲线下面积(AUROC)和 PR 曲线下面积(AUCPR)两个互补指标来评价检测精度。ROC 曲线由假正率(FPR)和真正率(TPR)绘制，AUC 为 ROC 曲线与坐标轴围成的面积。而 PR 曲线是由精确率(Precision)与召回率(Recall)绘制的关系图，召回率通常与真正率 TPR 相同。对应计算公式分别为：

$$F_{TPR}=\frac{TP}{TP+FN} \tag{8}$$

$$F_{FPR}=\frac{FP}{TN+FP} \tag{9}$$

$$Precision = \frac{TP}{TP + FN} \tag{10}$$

$$AUC = \int_0^1 TPR(FPR)\mathrm{d}(FPR) \tag{11}$$

本研究使用 ROC 曲线与 x 轴围成的面积和 PR 曲线与 x 轴围成的面积作为评价指标,分别记为 AUROC、AUCPR,越接近 1 表示模型性能越好。在实验中,使用 AUROC 和 AUCPR 来评估表格型数据的异常检测性能,使用 I-AUROC 评估图像级异常性能。

具体而言,为了全面评估模型的鲁棒性,通过迭代 100 次获得 AUROC、AUCPR 等指标值,排序后取前 10%性能指标进行平均计算,强调了高置信度样本在模型性能评估中的权重,降低了极端低分样本的干扰风险,进一步验证算法的有效性和稳定性,相比于全局平均值的比较,这种方法更具针对性,更准确地反映模型在关键场景下的表现。

3 实验结果与分析

本文进行了 3 个实验,第 1 个实验是在表格型数据集上验证 GDIF 算法的效果,第 2 个实验是验证 DDIF 算法在图像异常检测领域的可拓展性,第 3 个实验是验证 DGDIF 算法在图像异常检测领域的有效性。所有实验的运行环境为 Intel Xeon Gold 6226R CPU @ 2.90 GHz,使用 Python 3.8.20 环境下 PyTorch 1.13 与 CUDA 11.8,规格为 3 核心 15 G 内存 1 个加速器,操作系统为 64 位的 Windows 11。

3.1 特征分组对表格数据的有效性分析

为了验证 GDIF 算法的有效性,将其应用于表格型数据,并选取经典无监督检测算法:基于密度的 LOF[8]、基于树的标准 IF[10]和 DIF[12],与模型 GDIF 在 9 个数据集上做对比,实验结果如表 3 所示。

由表 3 可知,在 9 个表格型数据集中,相较于 LOF 算法、传统的标准 IF 算法和 DIF 算法,GDIF 算法减少了假阴性,使得 GDIF 算法在平均检测精度上显著提高,获得了更好的 AUCPR 和 AUROC 性能。在平均 AUCPR 方面,GDIF 明显优于 LOF47.2%、优于 IF20.4%、优于 DIF17.8%;在平均 AUC-ROC 方面,GDIF 优于 LOF23.2%、优于 IF3%、优于 DIF5.8%。

表 3 GDIF 有效性分析实验结果

Table 3 Experimental results of GDIF effectiveness analysis

数据集	AUROC				AUCPR			
	LOF	IF	DIF	GDIF	LOF	IF	DIF	GDIF
Breastw	0.528	0.817	0.626	**0.824**	0.406	**0.698**	0.443	0.662
arrhythmia	0.756	0.775	0.783	**0.784**	0.729	0.776	**0.779**	0.767
ionosphere	0.887	0.865	0.904	**0.907**	0.836	0.823	**0.893**	0.882
pageblocks	0.727	0.911	0.912	**0.912**	0.366	0.507	0.582	**0.595**
pima	0.597	0.683	0.663	**0.687**	0.420	0.480	0.458	**0.492**
shuttle	0.985	0.908	0.964	**0.991**	0.311	0.115	0.215	**0.451**
Thyroid	0.677	0.866	0.764	**0.914**	0.154	0.383	0.300	**0.576**
creditcard	0.486	0.950	0.953	**0.955**	0.002	0.155	0.387	**0.407**
Diabetes	0.598	0.692	0.689	**0.708**	0.420	0.520	0.494	**0.532**
Average	0.693	0.830	0.807	**0.854**	0.405	0.495	0.506	**0.596**

3.2 DDIF 和 DGDIF 对图像数据的有效性分析

为了验证 DDIF 方法的有效性,将 DDIF 模型与 AE-SSIM[20]、U-Net[21]、DAGAN[22]3 种模型进行对比实验,在 MVTec AD 数据集上的实验结果如表 4 所示。

从表 4 中可以看出,基于预训练 DenseNet121 的特征提取和 DIF 算法的组合在图像异常检测任务上表现出了良好的效果,这不仅保留了深度孤立森林方法的优势,同时也拓宽了他的应用领域,通过与其他几个异常检测方法对比,发现 15 个数据集中有 9 个数据集表现出最优 I-AUROC,验证了其在工业场景中的应用潜力。

采用 DGDIF 方法在 MVTec AD 数据集上验证其有效性,并与没有结合特征分组的 DDIF 算法做对比,实验结果如表 5 所示。

实验结果表明,相较于 DDIF,结合特征分组的 DGDIF 方法在工业图像数据集 MVTec AD 的 15 个数据集中有 9 个数据集的 I-AUROC 值得到改善,可见特征分组在解决高维特征数据随机选择问题方面具有一定的有效性。

3.3 参数分析

该异常检测任务中,特征分组的选择对模型性能具有重要影响,通过控制特征子集的大小 G 来研究其对检测性能的影响,绘制了不同特征组大小对性能 AUROC 和对应运行时间的时长,如图 4 所示。

从图 4 可以看出,随着特征分组的增加,AUROC 和运行时间大致呈先上升后下降的趋势,说明特征分组的优化不仅影响准确性,也影响计算效率。综合考虑这两者的重要性后选择特征组大小 $G = 4$,当特征组大小取 4 时,

AUROC 和运行时间达到一个较为平衡的性能。因此，合理调整特征分组大小对异常检测性能具有重要影响。未来工作将进一步研究更自适应的特征选择策略，以进一步提升模型表现。

表 4　DDIF 可拓展性分析结果

Table 4　Results of DDIF scalability analysis

数据集	I-AUROC			
	AE-SSIM	U-Net	DAGAN	DDIF
carpet	0.874	0.774	0.903	**0.994**
grid	**0.942**	0.857	0.867	**0.664**
leather	0.784	0.870	0.944	**0.998**
wood	0.734	0.958	0.979	**0.990**
tile	0.592	0.964	0.961	**0.973**
zipper	0.881	0.750	0.781	**0.901**
cable	0.764	0.636	0.665	**0.770**
toothbrush	0.923	0.811	**0.950**	0.851
transistor	0.890	0.674	0.794	**0.898**
pill	**0.912**	0.781	0.768	0.744
metal_nut	0.800	0.676	0.815	**0.801**
bottle	0.931	0.863	0.963	**0.968**
hazelnut	0.972	0.996	**1**	0.875
capsule	**0.943**	0.673	0.687	0.719
screw	0.964	**1**	**1**	0.634

表 5　DGDIF 有效性分析结果

Table 5　Results of DGDIF effectiveness analysis

数据集	I-AUROC	
	DDIF	DGDIF
carpet	0.994	0.968
grid	0.664	0.600
leather	0.998	**0.999**
wood	0.99	0.975
tile	0.973	**0.986**
zipper	0.901	**0.935**
cable	0.770	**0.801**
toothbrush	0.851	**0.897**
transistor	0.898	0.769
pill	0.744	0.733
Metal_nut	0.801	**0.836**
bottle	0.968	**0.983**
hazelnut	0.875	0.860
capsule	0.719	**0.811**
screw	0.634	**0.647**

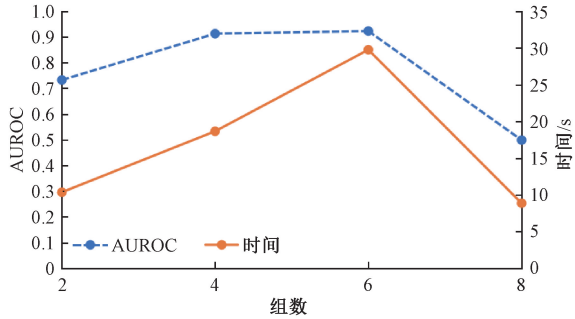

图 4　Thyroid 数据集上特征分组大小对 AUROC 和运行时间的影响

Fig. 4　Effect of feature group size on AUROC and runtime on Thyroid dataset

4　结　　论

本文探索深度孤立森林算法在图像异常检测上的性能优势，将深度学习的预训练特征提取器和无监督的深度孤立森林算法相结合，提出了一种新的图像异常检测算法 DDIF，扩充了深度孤立森林算法的应用领域，MVTec AD 数据集上的实验证明了该方法的实用性。虽然 DDIF 算法给图像异常检测带来了良好的检测效果，但是特征提取过程中，带来了高维重要特征漏选问题，因此又提出了 GDIF 异常检测算法，将数据特征进行分组，每次固定一组特征并随机选取未分组特征一起组合后输入到深度孤立森林模型中，确保数据的所有特征至少被选中一次参与模型的构建，有效提高了异常检测的性能，经 9 个表格数据集的实验验证，该算法优于经典的基于密度的 LOF 算法、传统孤立森林算法和深度孤立森林算法。最后将 DDIF 算法结合 GDIF 算法引出 DGDIF 算法应用于 MVTec AD 图像异常检测数据集，再次验证特征分组的有效性，改善了随机选取特征构建孤立树带来的问题。

参考文献

[1] KANG Y, CAI Z, TAN C W, et al. Natural language processing (NLP) in management research: A literature review [J]. Journal of Management Analytics, 2020, 7(2): 139-172.

[2] 董昱灿，赵奎. 基于注意力机制多特征融合与文本情感分析的日志异常检测方法[J]. 四川大学学报(自然科学版)，2024，61(2)：76-86.
DONG Y C, ZHAO K. A log anomaly detection method based on attention mechanism multi-feature fusion and text sentiment analysis [J]. Journal of Sichuan University(Natural Sciences Edition), 2024, 61(2): 76-86.

[3] 杨晨龙，孙晔，刘晓悦. 基于 GAT-AGRU 的多元时序数据异常检测[J]. 电子测量技术，2024，47(17)：

38-46.
YANG CH L, SUN Y, LIU X Y. Multiple timing data anomaly detection based on GAT-AGRU[J]. Electronic Measurement Technology, 2024, 47(17): 38-46.

[4] 常兴亚, 武云鹤, 陈东岳, 等. 基于多任务学习的视频异常检测方法[J]. 仪器仪表学报, 2023, 44(8):21-29.
CHANG X Y, WU Y H, CHEN D Y, et al. Video anomaly detection method based on multi-task learning[J]. Journal of Instrumentation, 2023, 44(8): 21-29.

[5] CHEN Y, HU X, FAN W, et al. Fast density peak clustering for large scale data based on kNN[J]. Knowledge-Based Systems, 2020, 187: 104824.

[6] AHMED M, SERAJ R, ISLAM S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.

[7] DENG D. DBSCAN clustering algorithm based on density[C]. 2020 7th International Forum on Electrical Engineering and Automation (IFEEA). IEEE, 2020: 949-953.

[8] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: Identifying density-based local outliers[C]. 2000 ACM SIGMOD International Conference on Management of Data, 2000: 93-104.

[9] YANG Y, CHEN L, FAN C J. ELOF: Fast and memory-efficient anomaly detection algorithm in data streams[J]. Soft Computing, 2021, 25(6): 4283-4294.

[10] LIU F T, TING K M, ZHOU Z H. Isolation-based anomaly detection[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2012, 6(1): 1-39.

[11] HARIRI S, KIND M C, BRUNNER R J. Extended isolation forest[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(4): 1479-1489.

[12] 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.

[13] LIU T, ZHOU Z, YANG L. Layered isolation forest: A multi-level subspace algorithm for improving isolation forest[J]. Neurocomputing, 2024, 581: 127525.

[14] KARCZMAREK P, KIERSZTYN A, PEDRYCZ W, et al. K-means-based isolation forest[J]. Knowledge-Based Systems, 2020, 195: 105659.

[15] KARCZMAREK P, KIERSZTYN A, PEDRYCZ W, et al. Fuzzy c-means-based isolation forest[J]. Applied Soft Computing, 2021, 106: 107354.

[16] CHATER M, BORGI A, SLAMA M T, et al. Fuzzy isolation forest for anomaly detection[J]. Procedia Computer Science, 2022, 207: 916-925.

[17] 沈萍,陈俊丽.基于孤立森林评分扩展的流量异常检测方法[J].电子测量技术, 2024, 47(8): 157-163.
SHEN P, CHEN J L. A traffic anomaly detection method based on isolated forest score extension[J]. Electronic Measurement Technology, 2024, 47(8): 157-163.

[18] 夏志祥,李准,徐伟.大气电场测量数据的异常检测及校正方法研究[J].电子测量技术, 2023, 46(1): 90-96.
XIA ZH X, LI ZH, XU W. Research on anomaly detection and correction methods of atmospheric electric field measurement data[J]. Electronic Measurement Technology, 2023, 46(1): 90-96.

[19] HUANG G, LIU S, VAN DER MAATEN L, et al. Condensenet: An efficient densenet using learned group convolutions[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2752-2761.

[20] BERGMANN P, FAUSER M, SATTLEGGER D, et al. MVTec AD-A comprehensive real-world dataset for unsupervised anomaly detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 9592-9600.

[21] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, 2015: 234-241.

[22] TANG T W, KUO W H, LAN J H, et al. Anomaly detection neural network with dual auto-encoders GAN and its industrial inspection applications[J]. Sensors, 2020, 20(12): 3336.

作者简介

周训会,硕士研究生,主要研究方向为异常检测、机器学习与模式识别。
E-mail:1838972714@qq.com

黄成泉(通信作者),博士,教授,主要研究方向为机器学习、模式识别与图像处理。
E-mail:hcq863@163.com

肖洪湖,硕士研究生,主要研究方向为深度学习。
E-mail:2143821719@qq.com

董红来,硕士研究生,主要研究方向为深度学习、虚拟试穿。
E-mail:1957468512@qq.com