基于深度对抗学习潜在表示分布的异常检测模型

席 亮,刘 涵,樊好义,张凤斌
(哈尔滨理工大学计算机科学与技术学院,黑龙江哈尔滨 150080)

摘 要: 针对已有的异常检测模型在高维、样本多样(类内多样)的数据背景下无法获得合理的潜在表示分布,不平衡数据较多(正常数据远大于异常数据)时特征提取准确性低,以及分类器超参数敏感等问题,本文提出一种基于深度对抗学习潜在表示分布的异常检测模型.基于正则化约束改进自编码器,将数据的原始特征空间映射到潜在特征空间形成低维的潜在表示,使其保持合理的空间分布;配以基于多判别器的生成对抗网络,在有效避免重构特征循环不一致和训练不稳定的基础上,来精确估计潜在表示的概率分布;以获得的潜在表示概率分布为单类分类器的输入,解决单类分类器超参数敏感问题,从而有效提高异常检测的整体性能.实验结果表明,相比于最新的基于机器学习和深度学习的异常检测模型,本文模型可在高维、样本多样、不平衡数据较多的应用背景下获得更合理的潜在表示空间分布并有效估计其概率分布,对单类分类器的超参数不敏感,并有效提高模型的检测性能.

关键词: 异常检测;深度学习;自编码器;生成对抗网络;潜在表示;特征融合
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112 (2021) 07-1257-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200970

Deep Adversarial Learning Latent Representation Distribution Model for Anomaly Detection

XI Liang, LIU Han, FAN Hao-yi, ZHANG Feng-bin
(*School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, Heilongjiang* 150080, *China*)

Abstract: To solve the problems of the existing anomaly detection models, such as incoherent latent representation distribution under in high-dimensional and diverse(within each class) data background, the low accuracy of feature extraction when unbalanced data(normal data far outweighs abnormal data) is large, and the sensitivity of classifier's hyperparameter, a deep adversarial learning latent representation distribution model for anomaly detection is proposed. Based on the regularization constraint, an improved autoencoder can map the original data feature space to a low-dimensional the latent feature space to get the reasonable latent representation distribution. On the premise of avoiding the problems of circulation inconsistent of reconstruction feature and unstable training, the multi-discriminator-based generative adversarial network can evaluate the latent representation probability distribution accurately, and to solve the hyperparameter sensitivity of one class classifier, so as to improve the overall performances of anomaly detection. Experimental results show that, compared with the up-to-date anomaly detection models based on machine learning and deep learning, the proposed model can obtain more coherent space distribution and ideal probability distribution of latent representation, is not sensitive to the hyperparameters of the single-class classifier, and effectively improve the detection performances under the application background with high-dimensional, diverse, unbalanced data.

Key words: anomaly detection; deep learning; autoencoder; generative adversarial network; latent representation; feature fusion

收稿日期:2020-09-03;修回日期:2021-02-20;责任编辑:覃怀银
基金项目:黑龙江省自然科学基金(No. F2018019)

1 引言

异常检测(anomaly detection)在各个应用领域都发挥着重要作用[1,2]. 然而,许多基于机器学习的异常检测模型,如支持向量数据域描述[3]、单类支持向量机(One Class Support Vector Machine, OC-SVM)[4]等,存在超参数敏感问题,而且寻优时间高[5]. 基于深度学习的异常检测的最大特点是能够学习数据的低维潜在表示,即表示学习. 这比原始特征更容易提取有用信息[6]. 然而已有的模型在高维、样本多样(类内多样)的数据集上无法获得理想的数据潜在表示分布,在不平衡数据较多(正常数据远大于异常数据)或缺少异常标记的数据集上检测性能不高. 这使得异常检测无法满足样本多样的数据应用背景.

自编码器(AutoEncoder, AE)[7]是一种优秀的表示学习方法,通过对正常数据进行训练获取重构特征,使其具有较小的重构误差(Reconstruction Error, RE),而异常数据的重构特征则会有较高的RE. 因此,RE通常作为异常的评分度量[8]. 生成对抗网络(Generative Adversarial Network, GAN)[9]是一种优秀的深度生成模型,通过生成器和判别器的对抗学习原始特征的概率分布,直接进行采样和推断,避免了传统的Markov模型的高复杂计算过程[10]. 同时,GAN是一种灵活的框架,可以与AE结合,充分利用潜在表示的优势,在具有高维、样本多样、不平衡特点的数据集上建模效果显著[11].

然而,经我们研究发现:(1)AE学习到潜在表示通常分布在空间的不同区域. 不合理的潜在表示空间分布会导致GAN学习速度慢和估计潜在表示概率分布不准确,无法避免分类器超参数敏感问题;(2)原始GAN使用随机采样方式会生成一些冗余、无用的低质量特征样本,会影响判别器的判别能力,学习能力会逐渐下降. 不稳定的训练过程也会影响潜在表示的概率分布估计;(3)在AE与GAN的结合过程中,原始特征的信息损失会造成重构特征的循环不一致性问题,影响GAN的稳定性,也影响潜在表示的概率分布估计. 因此,本文提出基于深度对抗学习潜在表示分布的异常检测模型,利用正则化约束的AE获得更合理的潜在表示空间分布,设计基于多判别器的GAN,综合考虑多种特征组合,在有效避免重构特征循环不一致和GAN训练不稳定的基础上,准确估计潜在表示的概率分布,并以此进行基于单类分类器的异常检测,解决其超参数敏感问题,在高维、样本多样、不平衡数据背景下完成高质量的异常检测任务.

2 相关工作

目前,基于深度学习的异常检测方法主要是以RE作为异常评分. 文献[12]提出一种改进的变分AE(Variational AE, VAE),由两个神经网络组成:全连接的VAE和跳过卷积VAE. 二者都是高效的生成网络. 文献[13]提出一种基于注意力机制LSTM的多尺度卷积循环编码器,利用特征残差签名矩阵对多元时间序列数据进行异常检测. 文献[14]提出一种深度自编码器高斯混合模型(Deep AE Gaussian Mixture Model, DAGMM)用于异常检测,利用AE为每个特征样本生成潜在表示,并将其嵌入到高斯混合模型(GMM),可以很好地平衡AE重构和潜在表示的概率分布估计.

而且,GAN也被广泛用于异常检测. 文献[15]提出一种基于GAN的无监督多元异常检测方法,考虑整个特征空间以捕获特征间的关系. 文献[16]使用一种双向GANs用于工业软件的异常检测. 文献[17]提出一种快速异常GAN(fast Anomaly GAN, f-AnoGAN),使用RE获得的潜在表示作为GAN的输入,使用判别器的特征残差和图片的AE组合快速准确地识别异常.

以上方法对高维、样本多样的数据进行了有效地特征学习和降维,利用GAN也可有效估计原始特征或潜在表示的概率分布,提高基于RE的异常检测的整体性能. 然而,这些方法忽略潜在表示的空间分布. 而且,AE和GAN结合的过程中还存在重构特征循环不一致性和GAN训练不稳定等问题. 这将影响模型在高维、样本多样、不平衡数据较多的数据集上的整体检测性能.

3 基于深度对抗学习潜在表示分布的异常检测

本文提出的基于深度对抗学习潜在表示分布(Deep Adversarial Learning latent Representation distribution, DALR)的异常检测模型利用正则化约束改进AE以获得更合理的潜在表示空间分布,并设计基于多判别器的GAN,以“潜在表示+原始特征+重构特征”为输入,结合生成器生成的生成特征,全面进行对抗博弈学习,有效避免重构特征循环不一致性和GAN训练不稳定,着重在高维、样本多样、不平衡数据较多的应用数据背景下获得更合理的潜在表示空间分布并有效估计其概率分布,以避免单类分类器的超参数敏感问题,从而提高模型的检测性能. 该模型主要由一个特征编码器、一个特征解码器、一个生成器、三个独立的判别器和一个单类分类器共七个模块组成,如图1所示,其中concat( )为特征级联融合函数.

3.1 基于正则化约束的自编码器

对于给定数据集的特征集 $X=\{\boldsymbol{x}^1,\boldsymbol{x}^2,...,\boldsymbol{x}^n\}$ 中的某个特征样本 $\boldsymbol{x}^f$,AE通过编码器获得潜在表示 $\boldsymbol{z}^f$. 但在不约束的情况下,$\boldsymbol{z}^f$ 会在潜在特征空间的不同区域并表

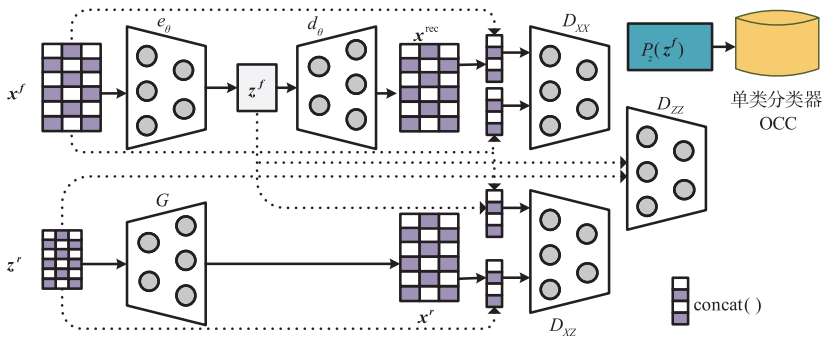


图1　基于深度对抗学习潜在表示分布的异常检测模型

现出不同的形状. 而且,在GAN中估计$\boldsymbol{x}^f$的似然需要很大的计算量,学习速度慢. 因此,DALR设计基于正则化约束的AE使潜在表示形成更合理的分布,从而获得准确的概率分布,提高GAN的学习速度,降低单类分类器的参数敏感性,以提高模型的整体检测性能.

定义1　特征编码器e_θ将输入的$\boldsymbol{x}^f \in X$映射到潜在特征空间得到$\boldsymbol{z}^f$:$\boldsymbol{z}^f = e_\theta(\boldsymbol{x}^f)$;特征解码器$d_\theta$将$\boldsymbol{z}^f$映射回原始特征空间,从而重构特征:$\boldsymbol{x}^{\text{rec}} = d_\theta(\boldsymbol{z}^f)$. e_θ和d_θ以非线性函数映射的多层感知机(Multi-Layer Perceptron, MLP)表示:

$$e_\theta(\boldsymbol{x}^f) = s_e(\boldsymbol{W}^e\boldsymbol{x}^f + \boldsymbol{b}^e) \tag{1}$$

$$d_\theta(\boldsymbol{z}^f) = s_d(\boldsymbol{W}^d\boldsymbol{z}^f + \boldsymbol{b}^d) \tag{2}$$

其中$\boldsymbol{W}^e$和$\boldsymbol{W}^d$分别表示e_θ和d_θ的权重矩阵,$\boldsymbol{b}^e$和$\boldsymbol{b}^d$分别表示e_θ和d_θ的偏移矢量矩阵. s_e和s_d分别表示其激活函数(本文选用tanh( )和lrelu( )).

(1) e_θ:为由三层神经网络组成的MLP,每层都是全连接层(Fully Connected layers, FC):

$$\boldsymbol{z}^f = \text{MLP}(e_\theta(\boldsymbol{x}^f)) = \text{MLP}(s_e(\boldsymbol{W}^e\boldsymbol{x}^f + \boldsymbol{b}^e)) \tag{3}$$

(2) d_θ:结构与e_θ网络相似,但神经元个数是逐层递增的:

$$\boldsymbol{x}^{\text{rec}} = \text{MLP}(d_\theta(\boldsymbol{z}^f)) = \text{MLP}(s_d(\boldsymbol{W}^d\boldsymbol{z}^f + \boldsymbol{b}^d)) \tag{4}$$

(3) 损失函数:加入正则化约束项,约束$\boldsymbol{z}^f$集中在空间的特定区域,获得更合理的空间分布:

$$L_{\text{AE}}(\theta;\ \boldsymbol{x}^f, \boldsymbol{z}^f) = \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{x}^f, \boldsymbol{x}^{\text{rec}}) + \gamma\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{z}^f\right\|^2 \tag{5}$$

其中,θ为e_θ和d_θ需要学习的参数,基于$\boldsymbol{W}^e$,$\boldsymbol{W}^d$,$\boldsymbol{b}^e$和$\boldsymbol{b}^d$训练得出. 第一项为RE;第二项为正则化约束项,取平方后求均值可使$\boldsymbol{z}^f$随训练过程逐渐减小(由于归一化后的$\boldsymbol{x}^f \in [0,1]$),实现类内距离缩短,类间距离增大,更合理化潜在表示分布;γ是两项之间的权衡因子. 使用随机梯度下降法学习参数θ来最小化损失函数完成AE训练.

3.2　深度对抗学习潜在表示分布模型

GAN由生成器G和判别器D组成. G将随机生成的潜在表示$\boldsymbol{z}^r$(采样自Gaussian分布)映射到特征空间获得与$\boldsymbol{x}^f$更相似的生成特征,D尝试区分$\boldsymbol{x}^f$与生成特征$G(\boldsymbol{z}^r)$. 二者对抗博弈完成训练.

在AE与GAN的结合过程中,e_θ会使潜在表示中有利于分类判别的信息减少,引起重构特征信息损失,造成重构特征循环不一致问题,影响GAN的训练稳定性,造成潜在表示概率分布估计不准确. 因此,本文借鉴文献[18]提出由一个生成器和三个判别器组成的GAN,使用自适应矩估计梯度下降法不断降低三个判别器的损失,考虑各种有用的特征组合进行对抗学习,使生成特征的分布与原始特征分布差异逐渐变小,从而解决这些问题.

定义2　将$P_X(\boldsymbol{x}^f)$定义为$\boldsymbol{x}^f \in X$的概率分布,将$P_Z(\boldsymbol{z}^r)$定义为$\boldsymbol{z}^r \in \boldsymbol{Z}$的概率分布. 将$P_G(\boldsymbol{z}^r)$定义为$\boldsymbol{z}^r$通过$G$生成特征$\boldsymbol{x}^r$的概率分布,训练$G$和$D$解决鞍点问题,即$\min_G \max_D V(D,G)$:

$$\begin{aligned} V(D,G) = {} & E_{\boldsymbol{x}^f\sim p(\boldsymbol{x}^f)}[\log D(\boldsymbol{x}^f)] + \\ & E_{\boldsymbol{z}^r\sim p(\boldsymbol{z}^r)}[\log(1 - D(G(\boldsymbol{z}^r)))] \end{aligned} \tag{6}$$

对于固定的G,最优的D_G^*为

$$D_G^* = \frac{P_X(\boldsymbol{x}^f)}{P_X(\boldsymbol{x}^f) + P_G(\boldsymbol{z}^r)} \tag{7}$$

当且仅当$P_X(\boldsymbol{x}^f) = P_G(\boldsymbol{z}^r)$时,$G$和$D$同时达到全局最优.

(1) 生成器G:由三层神经网络组成,采用全连接方式. 第一层神经元个数为潜在特征空间维度,第三层神经元个数为原始特征空间维度. 对于这些生成特征,生成模型中的似然计算可表示为:

$$L = \prod_1^n P_G(\boldsymbol{x}^r;\ \lambda) \tag{8}$$

其目标是寻找最优参数λ^*使似然函数最大化,即$\lambda^* = \arg\max \prod_1^n P_G(x^r;\ \lambda)$.

(2) 判别器D_{XZ}:是最核心的判别器. 以concat$(\boldsymbol{x}^f, \boldsymbol{z}^f)$和concat$(\boldsymbol{x}^r, \boldsymbol{z}^r)$为输入进行对抗学习,得到带有判别信息的潜在表示,提高训练效率,快速达到

Nash均衡以完成对抗学习过程. 其中,生成特征与潜在表示的联合概率分布分别为:

$$P_G(\boldsymbol{x}^r, \boldsymbol{z}^r) = P_Z(\boldsymbol{z}^r)P_G(\boldsymbol{x}^r|\boldsymbol{z}^r) \tag{9}$$

$$P_{e_\theta}(\boldsymbol{x}^f, \boldsymbol{z}^f) = P_Z(\boldsymbol{z}^f)P_{e_\theta}(\boldsymbol{x}^f|\boldsymbol{z}^f) \tag{10}$$

这时,可以将式(6)改写为:

$$V(D_{XZ}, e_\theta, G) = E_{\boldsymbol{X}\sim P(\boldsymbol{x}^f)}[\log D_{XZ}(\boldsymbol{x}^f, e_\theta(\boldsymbol{x}^f))] + E_{\boldsymbol{X}\sim P(\boldsymbol{z}^r)}[1 - \log D_{XZ}(G(\boldsymbol{z}^r), \boldsymbol{z}^r)] \tag{11}$$

对于最佳的D_{XZ},当且仅当$P_G(\boldsymbol{x}^r, \boldsymbol{z}^r)$=$P_{e_\theta}(\boldsymbol{x}^f, \boldsymbol{z}^f)$时,$V(D_{XZ}, e_\theta, G)$全局达到最小值:

$$D_{XZ} = \frac{P_{e_\theta}(\boldsymbol{x}^f, \boldsymbol{z}^f)}{P_{e_\theta}(\boldsymbol{x}^f, \boldsymbol{z}^f) + P_G(\boldsymbol{x}^r, \boldsymbol{z}^r)} \tag{12}$$

(3) 判别器D_{XX}:以concat$(\boldsymbol{x}^f, \boldsymbol{x}^{\text{rec}})$和concat$(\boldsymbol{x}^f, \boldsymbol{x}^f)$为输入进行对抗学习,辅助AE降低重构特征信息损失. 借鉴文献[18],用式(13)使D_{XX}估计d_θ和e_θ的条件熵,使$d_\theta(e_\theta(\boldsymbol{x}^f)) = \boldsymbol{x}^{\text{rec}}$,解决重构特征循环不一致问题.

$$H^\beta(\boldsymbol{x}|\boldsymbol{z}) = -E_{\beta(\boldsymbol{x},\boldsymbol{z})}[\log\beta(\boldsymbol{x}|\boldsymbol{z})] \tag{13}$$

其中$\beta(\boldsymbol{x}, \boldsymbol{z})$是$\boldsymbol{x}$和$\boldsymbol{z}$的联合概率分布. 此时,GAN的鞍点问题就转化成:

$$\min_{d_\theta, e_\theta}\max_{D_{XX}} V(D_{XX}, e_\theta, d_\theta) \tag{14}$$

此时,式(6)改写为:

$$V(D_{XX}, e_\theta, d_\theta) = E_{X\sim P(\boldsymbol{x}^f)}[\log D_{XX}(\boldsymbol{x}^f, \boldsymbol{x}^f)] + E_{X\sim P(\boldsymbol{x}^f)}[1 - \log D_{XX}(\boldsymbol{x}^f, d_\theta(e_\theta(\boldsymbol{x}^f)))] \tag{15}$$

(4) 判别器D_{ZZ}:生成器随机采样方式会不可避免地生成一些冗余、无用的低质量特征样本,对判别器的判别能力产生负面影响,使学习能力逐渐下降,训练整体过程不稳定. D_{ZZ}以$\boldsymbol{z}^r$与$\boldsymbol{z}^f$(存在类内多样性)为输入进行对抗学习,使$\boldsymbol{z}^r$和$G(\boldsymbol{z}^r) = \boldsymbol{x}^r$更具多样性,提高GAN的学习能力及稳定性. 此时,式(6)改写为:

$$V(D_{ZZ}, e_\theta) = E_{\boldsymbol{Z}\sim P(\boldsymbol{z}^f)}[\log D_{ZZ}(\boldsymbol{z}^f)] + E_{\boldsymbol{Z}\sim P(\boldsymbol{z}^r)}[1 - \log D_{ZZ}(\boldsymbol{z}^r)] \tag{16}$$

(5) 损失函数:

$$\min_{G, e_\theta, d_\theta}\max_{D_{XZ}, D_{XX}, D_{ZZ}} V(D_{XZ}, D_{XX}, D_{ZZ}, e_\theta, G, d_\theta) \tag{17}$$

$$V(D_{XZ}, D_{XX}, D_{ZZ}, e_\theta, G, d_\theta) = V(D_{XZ}, e_\theta, G) + V(D_{XX}, e_\theta, d_\theta) + V(D_{ZZ}, e_\theta, G) \tag{18}$$

生成器、编码器、解码器的损失和三个判别器的损失如式(19)和式(20)所示:

$$L_{G, e_\theta, d_\theta} = \min V(D_{XZ}, D_{XX}, D_{ZZ}, e_\theta, G, d_\theta) \tag{19}$$

$$L_{D_{XZ}, D_{XX}, D_{XX}} = \max V(D_{XZ}, D_{XX}, D_{ZZ}, e_\theta, G, d_\theta) \tag{20}$$

基于以上论述,模型伪代码过程如下:

算法1 深度对抗学习潜在表示分布模型

输入:输入特征样本$X = \{\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^n\}$

输出:潜在表示$\boldsymbol{z}^f$

```
Input data preprocess  //数据预处理
Initialize W^e, W^d, b^e, b^d //初始化权重和偏置
For in range (num_step) //num_step为迭代次数
  For in range(batch_size)//batch_size为每个训练批次大小
    z^f = e_θ(x^f)   //编码器对原始特征编码
    Random sampling z^r//随机采样自Gaussian分布
    x^r = G(z^r)   //生成器生成特征
    x^rec = d_θ(z^f)   //解码器重构特征
    D_XZ ← concat(x^f, z^f), concat(x^r, z^r)
    D_XX ← concat(x^f, x^f), concat(x^f, x^rec)
    D_ZZ ← z^r, z^f
    loss = L_AE + L_{G,e_θ,d_θ} + L_{D_XZ,D_XX,D_ZZ}
  //根据式(5)、式(19)和式(20)计算损失
  END FOR
END FOR
```

3.3 单类分类器(OCC)

DALR选用2种经典的机器学习单类分类算法作为模型的分类器,以训练获得的$\boldsymbol{z}^f$及其概率分布进行异常检测,即OCC$\left(P_Z\left(\boldsymbol{z}^f\right)\right)$,验证DALR对单类分类器的参数调整不敏感.

(1) 单类支持向量机(OC-SVM)[4]:通过核函数将训练样本映射到特征空间,将原点视为异常点,并使原点尽可能远离超平面(支持向量). 决策函数在靠近原点的异常区域中返回负值.

(2) 原点(CENtroid, CEN)[19]:使用单个Gaussian模型对训练数据进行建模,从CEN(原点)到被检测数据的距离(即半径)反映了其异常程度:值越大,异常的可能性越高. 该方法没有超参数.

4 实验与分析

4.1 实验环境与数据集

实验环境基于Windows操作系统,使用python语言,在tensorflow框架进行模型搭建. 主要的硬件环境为:GPU为GTX 1050 TI 4GB显存,CPU为Intel i5 8300H 4核,16GB内存,2TB硬盘.

实验选取权威数据集:(1) UNSW-NB15[20]:考察DALR在高维、样本多样、数据量大的数据集上的检测性能;(2)Shuttle[21]:考察DALR在数据量适中、维度较低、不平衡数据较多的数据集上的检测性能;(3)Arrhythmia[22]:考察DALR在高维但数据量少的数据集上的检测性能. 实验对数据集进行归一化处理,选取的数据样本统计详见表1,训练集均为正常样本. 除Shuttle(由于维度小)外,其他2个数据集采用one-hot编码方式.

表1　实验选取的数据集信息统计表

Dataset	Dimension	Training set	Test set	
			Normal	Abnormal
UNSW-NB15	196	56000	37000	45332
Shuttle	9	3410	11478	3022
Arrhythmia	259	189	48	37

4.2　度量标准

实验使用2个AUC(Area Under Curve)评价指标:基于潜在表示的AUC (Latent-Representation-based AUC, LR_AUC)和基于RE的AUC (RE_AUC). 模型越好,AUC越大. 使用这两个指标的目的为:

(1) 为了验证在合理潜在表示空间分布基础上使用潜在表示的性能要优于使用RE的性能;

(2) 对比各个模型的检测性能.

4.3　对比方法和参数设置

实验选取的对比方法为:(1)DALR使用的分类器:OC-SVM和CEN;(2)近几年5种代表性深度学习方法:GAN[9]和Efficient-GAN[23]、深度自编码器(Deep AE, DAE)[24]、收缩自编码器(Shrink AE, SAE)[8]、狄拉克变分自编码器(Dirac delta Variational AE, DVAE)[8].

本文通过参考相关文献和实验验证的方法设置AE、GAN和单类分类器的超参数并验证DALR对超参数不敏感.

(1) OC-SVM的权衡参数$v\in[0, 1]$,其含义为训练样本中被分类为异常的比例;

(2) CEN没有超参数;

(3) AE的隐含层神经元个数m采用经验法[25]:$m = 1 + \sqrt{n}$,其中n是输入的特征维度;

(4) GAN的隐含层神经元个数使用多次实验结果的最优参数.

不同数据集下相关参数设置均为大量实验所得出的最优设定,汇总如表2. 其中,batch_size为每个批次大小,num_step为迭代次数,learning_rate为学习率,γ为计算损失时的平衡因子. 由于UNSW-NB15数据量大,实验采用指数衰减学习率的方式使其随迭代次数的增加而减小.

表2　实验选取的相关参数信息统计表

Dataset	batch_size	num_step	Learning_rate	γ
UNSW-NB15	100	500	0.001	50
Shuttle	100	1000	0.001	50
Arrhythmia	100	1000	0.001	50

模型的网络参数根据数据集维度设置,如表3所示,其中FC(input, output,σ)表示输入input个神经元,输出output个神经元,σ表示激活函数,连接方式为全连接. 为保证公平,对比方法的参数设定与DALR保持一致,激活函数的选择、权重和偏置的初始化方式、梯度下降优化方法等选择其最优的网络优化方式.

表3　模型在不同数据集下的网络参数设置

模块	UNSW-NB15	Shuttle	Arrhythmia
e_θ	FC(196,136,tanh) FC(136,75,tanh) FC(75,15,tanh)	FC(9,7,tanh) FC(7,6,tanh) FC(6,4,tanh)	FC(259,178,tanh) FC(178,98,tanh) FC(98,17,tanh)
d_θ	FC(15,75,tanh) FC(75,136,tanh) FC(136,196,tanh)	FC(4,6,tanh) FC(6,7,tanh) FC(7,9,tanh)	FC(17,98,tanh) FC(98,178,tanh) FC(178,259,tanh)
G	FC(15,32,relu) FC(32,64,relu) FC(64,196,relu)	FC(4,32,relu) FC(32,64,relu) FC(64,9,relu)	FC(17,32,relu) FC(32,64,relu) FC(64,259,relu)
D_{XZ}	FC(256,128,lrelu) FC(128,1,lrelu)	FC(256,128,lrelu) FC(128,1,lrelu)	FC(256,128,lrelu) FC(128,1,lrelu)
D_{XX}	FC(392,128,lrelu) FC(128,1,lrelu)	FC(18,128,lrelu) FC(128,1,lrelu)	FC(518,128,lrelu) FC(128,1,lrelu)
D_{ZZ}	FC(30,128,lrelu) FC(128,1,lrelu)	FC(8,128,lrelu) FC(128,1,lrelu)	FC(34,128,lrelu) FC(128,1,lrelu)

4.4　实验结果分析

4.4.1　异常检测实验

本实验在3个数据集上分别进行. OC-SVM的超参数v=0.1(参数敏感性实验证实v=0.1时效果最优). OC-SVM和CEN没有RE,无法计算RE_AUC. GAN和Efficient-GAN不考虑潜在表示,无法计算LR_AUC. 实验结果如表4所示.

表4　不同数据集的异常检测结果

模型	UNSW-NB15		Shuttle		Arrhythmia	
	LR_AUC	RE_AUC	LR_AUC	RE_AUC	LR_AUC	RE_AUC
OC-SVM[4]	79.2%	----	76.0%	----	80.7%	----
CEN[19]	73.8%	----	88.1%	----	81.6%	----
GAN[9]	----	42.7%	----	31.8%	----	35.5%
Efficient-GAN[23]	----	58.1%	----	49.3%	----	44.2%
DAE[24]+OC-SVM	53.6%	87.3%	76.2%	82.1%	66.8%	82.4%
DAE[24]+CEN	74.3%		93.1%		73.8%	
SAE[8]+OC-SVM	89.3%	80.0%	78.1%	88.4%	74.0%	78.7%
SAE+CEN	88.6%		80.0%		75.4%	
DVAE[8]+OC-SVM	87.2%	80.3%	80.4%	88.0%	78.0%	78.5%
DVAE+CEN	87.9%		84.9%		77.7%	
DALR+OC-SVM	**90.9%**	**89.3%**	**93.3%**	**92.7%**	**82.8%**	73.3%
DALR+CEN	**90.8%**		**93.2%**		**82.9%**	

(1) 在所有方法中,DALR的效果最好. 这就说明在充分考虑潜在表示空间分布合理性的前提下,DALR可以更有效发挥深度学习模型的优势,适用于多种类

型、不同特点的数据集.

(2) 在深度学习方法中,GAN的效果最差,Efficient-GAN的效果提升也不理想,DAE好于GANs,SAE和DVAE效果也很好,仅次于DALR. 这就说明,GAN和Efficient-GAN不考虑潜在表示的空间分布,无法准确估计原始特征的概率分布,这也说明基于潜在表示的方式要优于基于原始特征的GANs;在考虑了潜在表示分布后,SAE和DVAE的效果获得明显提升(二者对潜在表示进行了一定的约束但不充分).

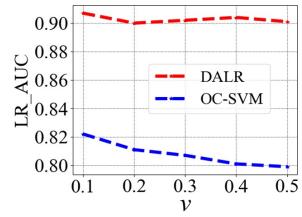
(3) 从除GAN、Efficient-GAN外的其他深度学习方法和机器学习方法(OC-SVM和CEN)的对比来看:在高维但数据量少的Arrhythmia数据集上,机器学习方法比DAE、SAE、DVAE的LR_AUC效果更好;在数据量适中、维度较低、但不平衡数据较多的Shuttle数据集上,机器学习方法与DAE、SAE、DVAE效果相当;在高维、样本多样、数据量大的UNSW-NB15数据集上,机器学习方法仅比DAE的LR_AUC优秀,与SAE和DVAE相差很大. 这就说明,机器学习方法受数据规模和维度、不平衡数据的影响较大;而深度学习方法只有在获得理想的潜在表示空间分布的前提下才能发挥其强大的表示学习能力,避免分类器超参数敏感问题,以适用于各种特点的数据背景.

(4) 对比AE相关方法(DAE、SAE、DVAE和DALR)的结果还可以发现:在UNSW-NB15数据集上,除了DAE,其他方法的LR_AUC都优于RE_AUC;在Shuttle和Arrhythmia数据集上,除DALR外的其他方法的LR_AUC都相当于或弱于RE_AUC. 这就说明,在对潜在表示进行合理分布约束后的模型,基于LR进行异常检测好于基于传统的RE方式,可有效克服数据背景的各种不利条件,不因维度和数据量限制而影响单类分类器的分类效果,更适用于与各种特点的数据背景.

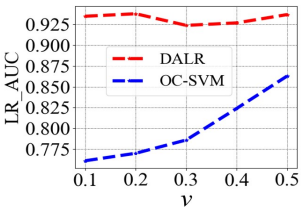
综合以上实验可以看出,DALR可以在不同特点的数据集下,通过对潜在表示空间分布的约束,在有效避免重构特征循环不一致性和GAN训练不稳定的基础上准确估计潜在表示的概率分布,提高模型在高维、样本多样、不平衡数据较多等特点的应用背景下的检测性能.

4.4.2 参数敏感性实验

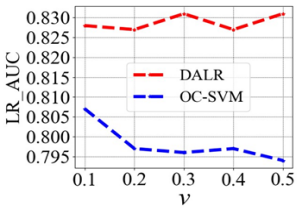
本节实验在3个数据集上分别进行,并与单独的OC-SVM进行对比,验证DALR对单类分类器的超参数不敏感. OC-SVM的超参数$v\in[0.1,0.5]$,其余过程同上一实验,最终结果如图2所示. 从中可以看出,在不同特点的数据集下,DALR在检测性能和稳定性等方面明显优于OC-SVM,超参数v的取值对DALR的性能影响很小. 这就说明DALR对单类分类器超参数不敏感.



(a) UNSW-NB15



(b) Shuttle

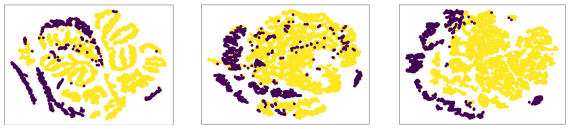


(c) Arrhythmia

图2 参数敏感性实验结果

4.4.3 数据可视化实验

实验使用t-SNE工具包对DALR、SAE、DAE训练后学习到的潜在表示进行可视化,验证DALR在构建数据潜在表示分布方面的性能,结果如图3所示,其中深紫色代表异常样本,黄色代表正常样本. 在DAE和SAE的结果中,只有部分正常和异常样本集中在一起,其他样本散落在空间的不同区域. 这是由于未对潜在表示进行约束或约束不充分. 相反,在DALR的结果中,正常和异常样本被有效分离,潜在表示空间分布十分理想. 这就体现了DALR可较好地约束潜在表示. 这些结果也对应了异常检测的实验结果,凸显了DALR相对于其他方法在构造潜在表示空间分布方面的优势.



(a) DAE (b) SAE (c) DALR

图3 数据可视化结果

4.4.4 消融实验

实验使用UNSW-NB15数据集,验证每个模块对性能的影响,参数设置同上,结果如表5所示.

(1) GAN:其潜在表示是随机采样自Gaussian分布,不能采用单类分类器进行检测,只能考察RE_AUC,

表5　UNSW-NB15数据集消融实验结果

模型	LR_AUC	RE_AUC
GAN	----	42.7%
AE+OC-SVM	81.8%	81.1%
AE+CEN	79.6%	
正则化AE+OC-SVM	84.5%	82.2%
正则化AE+CEN	84.7%	
AE+GAN+OC-SVM	73.4%	79.8%
AE+GAN+CEN	73.7%	
DALR+OC-SVM	**90.9%**	**89.3%**
DALR+CEN	**90.8%**	

结果也最差. 这说明由原始GAN随机生成的特征质量低,降低异常检测性能;

(2) AE相关方法:正则化AE比AE有2.7%以上的性能提升. 这说明正则化可有效约束潜在表示的空间分布,从而提高单类分类器的分类性能;

(3) AE+GAN:整体的检测结果不如AE相关方法. 这表明未正则化的AE会获得不合理的潜在表示空间分布,导致GAN无法准确估计潜在表示的概率分布,影响单类分类器核函数的决策,导致单类分类器识别异常的能力大打折扣;

(4) DALR(正则化AE+GAN):整体效果优于其他缺少某些模块的方法. 这说明正则化AE的必要性,使GAN可以准确估计潜在表示的概率分布,达到全局最优,令单类分类器准确识别出异常;

(5) 使用正则化AE的模型中,LR_AUC要优于RE_AUC. 这也说明AE正则化以获得合理潜在表示空间分布的必要性.

4.4.5　训练效率实验

本节实验使用三个数据集验证DALR的训练效率(单位:ms),参数设置同上,实验结果如表6所示. 总体来看,Efficient-GAN和CEN的效率最高,DALR居中,DVAE效率最低. 这是由于DVAE使用KL散度计算损失函数,计算效率远低于使用交叉熵计算损失的神经网络. DALR涉及较多的对抗学习过程,效率虽然不是最优,但在牺牲有限的效率的同时获得了最优的异常检测性能,这在对时间要求不高的训练中是可以接受的.

综上,异常检测实验体现出DALR在异常检测过程中具有良好的检测性能,参数敏感性实验体现出DALR的鲁棒性,数据可视化实验体现出DALR在构建潜在表示分布时的优良效果,消融实验验证了DALR正则化AE和多判别器对抗学习的必要性,训练效率实验证实DALR在可以接受的时间代价基础上获得很好的检测效果是可行的.

表6　DALR算法效率

Dataset / 模型	UNSW-NB15	Shuttle	Arrhythmia
OC-SVM[4]	38.96	0.41	0.08
CEN[19]	0.56	**0.02**	**0.01**
GAN[9]	1.72	0.10	0.03
Efficient-GAN[23]	**0.14**	**0.02**	0.04
DAE[24]	0.58	0.08	0.03
SAE[8]	0.64	0.08	0.03
DVAE[8]	152.33	26.73	8.05
DALR	0.92	0.18	0.10

5　总结与展望

基于深度学习的异常检测已有成果主要专注于以RE进行异常检测,而很少考虑潜在表示的分布约束. 这会导致在高维、样本多样(类内多样)的数据集上无法获得合理的潜在表示空间分布,在不平衡数据较多的数据集上特征提取的准确性低. 而且,传统机器学习方法对参数敏感性较高也影响了以此为分类器的相关模型的实际检测性能. 因此,本文提出了深度对抗学习潜在表示分布模型,在AE上加入正则化以实现对潜在表示分布约束,使用三个判别器在有效避免重构特征循环不一致性和GAN训练不稳定的基础上准确估计潜在表示的概率分布,最后基于此使用单类分类器进行异常检测,解决参数敏感性问题,提高模型在高维、样本多样、不平衡数据较多等应用背景下的检测性能. 实验也证实了以上结论. 未来我们还会继续深入挖掘数据的潜在表示形式,并搭建一个简单、有效、可以适用于多种应用背景下的深度异常检测模型.

参考文献

[1] 梅御东, 陈旭, 孙毓忠, 等. 一种基于日志信息和CNN-text的软件系统异常检测方法[J]. 计算机学报, 2020, 43(02): 366 – 380.
MEI Yu-dong, CHEN Xu, SUN Yu-zhong, et al. A method for software system anomaly detection based on log information and CNN-Text[J]. Chinese Journal of Computers, 2020, 43(02): 366 – 380. (in Chinese)

[2] 闻佳, 王宏君, 邓佳, 刘鹏飞. 基于深度学习的异常事件检测[J]. 电子学报, 2020, 48(2): 308 – 313.
WEN Jia, WANG Hong-jun, DENG Jia, LIU Peng-fei. Abnormal event detection based on deep learning[J]. Acta Electronica Sinica, 2020, 48(2): 308 – 313. (in Chinese)

[3] Pauwels E J, Ambekar O. One class classification for

anomaly detection: support vector data description revisited[A]. Proceedings of the 11th Industrial Conference on Data Mining: Applications and Theoretical Aspects [C]. New York, USA: Springer, 2011. 25 – 39.

[4] Zhang M, Xu B, Gong J. An anomaly detection model based on one-class SVM to detect network intrusions[A]. Proceedings of the 11th International Conference on Mobile Ad-hoc and Sensor Networks[C]. Shenzhen, China: IEEE, 2015. 102 – 107.

[5] Erfani, Sarah M, Baktashmotlagh, Mahsa, Rajasegarar, Sutharshan. R1SVM: a randomised nonlinear approach to large-scale anomaly detection[A]. Proceedings of the 29th AAAI Conference on Artificial Intelligence[C]. Austin, Texas, USA: AAAI, 2015. 1 – 7.

[6] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798 – 1828.

[7] Sakurada M, Yairi T. Anomaly detection using autoencoders with nonlinear dimensionality reduction[A]. Proceedings of the 2nd Workshop on Machine Learning for Sensory Data Analysis [C]. Gold Coast, Australia: ACM, 2014. 4 – 11.

[8] Nicolau M, McDermott J. Learning neural representations for network anomaly detection[J]. IEEE Transactions on Cybernetics, 2018, 49(8): 3074 – 3087.

[9] Akcay S, Atapour-Abarghouei A, Breckon T P. GANomaly: Semi-supervised anomaly detection via adversarial training[A]. Proceedings of the 14th Asian Conference on Computer Vision[C]. Perth, Australia: Springer, 2018. 622 – 637.

[10] Ngo P C, Winarto A A, Kou C, et al. Fence GAN: towards better anomaly detection[A]. Proceedings of the IEEE 31st International Conference on Tools with Artificial Intelligence[C]. Portland, OR, USA: IEEE, 2019. 141 – 148.

[11] Jiang W, Hong Y, Zhou B, et al. A GAN-based anomaly detection approach for imbalanced industrial time series [J]. IEEE Access, 2019, 7: 143608 – 143619.

[12] Malaiya R K, Kwon D, Kim J, et al. An empirical evaluation of deep learning for network anomaly detection [A]. Proceedings of the 2018 International Conference on Computing, Networking and Communications [C]. Maui, HI, USA: IEEE, 2018. 893 – 898.

[13] Zhang C, Song D, Chen Y, et al. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data[A]. Proceedings of the 34th AAAI Conference on Artificial Intelligence [C]. New York, USA: AAAI, 2019. 1409 – 1416.

[14] Zong B, Song Q, Min M R, et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection[A]. Proceedings of the 6th International Conference on Learning Representations[C]. Vancouver, BC, Canada: OpenReview, 2018. 1 – 19.

[15] Li D, Chen D, Jin B, et al. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks[A]. Proceedings of the 28th International Conference on Artificial Neural Networks [C]. Munich, Germany: Springer, 2019. 703 – 716.

[16] Kumarage T, Ranathunga S, Kuruppu C, et al. Generative adversarial networks (GAN) based anomaly detection in industrial software systems[A]. Proceedings of the 2019 Moratuwa Engineering Research Conference [C]. Moratuwa, Sri Lanka: IEEE, 2019. 43 – 48.

[17] Schlegl T, Seeböck P, Waldstein S M, et al. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks[J]. Medical image analysis, 2019, 54: 30 – 44.

[18] Li C, Liu H, Chen C, et al. Alice: Towards understanding adversarial learning for joint distribution matching [A]. Proceedings of the 31st Annual Conference on Neural Information Processing Systems[C]. Long Beach, USA: ACM, 2017. 5495 – 5503.

[19] Sarmiento A, Fondón Irene, Durán-Díaz Iván, et al. Centroid-based clustering with αβ-divergences[J]. Entropy, 2019, 21(2): ArticleID: 196.

[20] Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)[A]. Proceedings of the 2015 Military Communications and Information Systems Conference [C]. Canberra, ACT, Australia: IEEE, 2015. 1 – 6.

[21] Amarnath B, Balamurugan S A A. Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset [J]. Journal of Engineering Science and Technology, 2016, 11(11):1639 – 1646.

[22] DeCamilla J, Xia X, Wang M, et al. The multiple arrhythmia dataset evaluation database (MADAE) [J]. Journal of Electrocardiology, 2018, 51(6): S106 – S112.

[23] Fujioka T, Kubota K, Mori M, et al. Efficient anomaly detection with generative adversarial network for breast ultrasound imaging[J]. Diagnostics, 2020, 10(7): Arti-

cleID: 456.

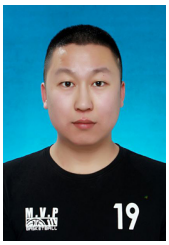
[24] Zhao H, Liu H, Hu W, et al. Anomaly detection and fault anAlysis of Wind tuRbine components based on deep learning network [J]. Renewable Energy, 2018, 127: 825 – 834.

[25] Seyed-Allaei H. Phase diagram of spiking neural networks [J]. Frontiers in Computational Neuroscience, 2015, 9: ArticleID: 19.

作者简介

席　亮　男, 1983年5月出生于河北省邢台市. 现为哈尔滨理工大学副教授、硕士生导师. 主要研究方向为人工智能及其应用、网络与信息安全、深度学习等.

E-mail:xiliang@hrbust.edu.cn



刘　涵　男, 1996年10月出生于黑龙江省哈尔滨市. 硕士研究生. 主要研究方向为人工智能及其应用, 深度学习.

E-mail:liuhanharbin@163.com

樊好义　男, 1994年6月出生于河南省新乡市. 博士研究生, 主要研究方向为网络嵌入、异常检测、时间序列信号分析、深度学习等.

E-mail:isfanhy@gmail.com

张凤斌　男, 1965年7月出生于黑龙江省哈尔滨市. 现为哈尔滨理工大学教授, 博士生导师, CCF高级会员. 主要研究方向为网络与信息安全, 人工智能与应用.

E-mail:zhangfb@hrbust.edu.cn