

文章编号:1007-5321(2021)05-0028-07

DOI:10.13190/j.jbupt.2021-009

# 融合互信息估计和对抗自编码器的异常检测

霍纬纲<sup>1,2</sup>, 王星<sup>2</sup>, 梁锐<sup>2</sup>

(1. 中国民航大学 信息安全测评中心, 天津 30030; 2. 中国民航大学 计算机科学与技术学院, 天津 300300)

**摘要:** 无监督深度学习网络的训练目标从信息论的角度可解释为最大化训练样本及其表示之间的互信息. 对抗自编码器(AAE)通过生成对抗的方式学习训练样本集的分布,据此可以由AAE建立基于正常样本集的半监督异常检测模型,但是AAE无法显式最大化正常样本及其表示间的互信息. 为此,提出了一种互信息估计网络和AAE相融合(IAAE)的异常检测方法,该方法首先以重构误差最小化为目标,训练编码器和解码器;其次,在对抗正则化阶段将正常样本低维表示的聚集后验分布约束为先验分布,并最大化正常样本与其表示之间的互信息;最后由全连接神经网络估计正常样本与其表示之间的互信息. 由待测样本的重构误差及其表示在隐空间中的众数散度计算其异常得分值. 公开数据集上的实验结果表明,与已有典型相关的深度异常检测模型相比,IAAE模型在F1取值上具有更好的表现.

**关键词:** 对抗自编码器; 互信息估计; 异常检测; 深度生成模型; 半监督学习

**中图分类号:** TP18 **文献标志码:** A

## An Anomaly Detection Method Combining Mutual Information Estimation with Adversarial Autoencoder

HUO Wei-gang<sup>1,2</sup>, WANG Xing<sup>2</sup>, LIANG Rui<sup>2</sup>

(1. Information Security Evaluation Center of Civil Aviation, Civil Aviation University of China, Tianjin 300300, China;

2. School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

**Abstract:** According to the information theory, the training objective of the unsupervised deep learning networks can be interpreted as maximizing the mutual information between the training samples and their representations. Adversarial autoencoder (AAE) learns the distribution of the training samples by the generative adversarial method. So the semi-supervised anomaly detection model based on the normal sample sets can be established using AAE. However, AAE cannot maximize the mutual information between the normal samples and their representations explicitly. A semi-supervised anomaly detection method based on mutual information estimation network and AAE (IAAE) is proposed. Firstly, the encoder and decoder of the AAE are trained to minimize the reconstruction error. Then, in the adversarial regularization stage, the aggregated posterior of the normal sample's representations are matched to the arbitrary prior distribution, and the mutual information between normal samples and their representations is maximized. Finally, the mutual information between normal samples and their representations are estimated by fully connected neural network. The reconstruction error of the test sample and its mode divergence in the hidden space are used to calculate the abnormal score. The experimental results on public datasets show that the IAAE has better performance than the existing typical deep anomaly detection models in terms of

收稿日期: 2021-01-15

基金项目: 中央高校基本科研业务费专项项目(3122019190); 中国民航大学信息安全测评中心开放基金项目(ISECCA-202003)

作者简介: 霍纬纲(1978—), 男, 教授, 硕士生导师, E-mail: wghuo@cauc.edu.cn.

F1 values.

**Key words:** adversarial autoencoder; mutual information estimation; anomaly detection; deep generative mode; semi-supervised learning

异常检测是指在数据中发现与期望行为不符的数据模式.近年来,基于深度学习的异常检测方法引起了广泛的研究关注,并取得了比传统异常检测方法优越的检测效果<sup>[1]</sup>.在实际异常检测的应用场景中,异常样本出现的频次往往较低,正常样本较多且容易获得,因此学者们提出了很多无监督和基于正常样本的半监督深度异常检测方法.这些方法可大致分为基于自编码器(AE, autoencoder)的模型<sup>[2-3]</sup>、基于变分自编码器(VAE, variational autoencoder)与生成对抗网络(GAN, generative adversarial networks)的深度生成模型<sup>[4-10]</sup>.

基于AE的异常检测模型问题在于,训练样本与其表示之间是确定的映射,这使得该模型对噪声比较敏感<sup>[3]</sup>;当处理复杂类型数据时,仅由重构误差无法正确地区分正常样本和异常样本<sup>[4]</sup>.Zong等<sup>[4]</sup>提出了一种深度自动编码高斯混合模型(DAGMM, deep autoencoding Gaussian mixture mode),该模型通过密度估计网络施加的正则化约束提高自编码器表示学习的质量.

An等<sup>[5]</sup>提出了基于VAE的异常检测方法,该方法由VAE的解码器网络生成重构正常样本分布的均值和方差,由此计算测试样本的重构概率,重构概率较小的样本判定为异常.Yao等<sup>[6]</sup>将VAE编码器网络生成的隐变量正态分布的均值作为输入样本的表示,由训练样本的表示向量集生成基于K-近邻、支持向量机、局部异常因子的异常检测模型.

Schlegl等<sup>[7]</sup>提出了用GAN学习正常视网膜图片的分布,通过定义残差损失和区分损失将待检测图片映射到生成器网络表示的隐空间中,如果该图片在隐空间中无法查询到与之匹配的表示向量,则将待测图片判定为异常.Zenati等<sup>[8]</sup>提出了一种基于双向生成对抗网络的异常检测方法,该方法利用了样本空间和隐空间循环一致特性<sup>[9]</sup>,采用生成器产生的图片与原始图片之间的重构误差检测异常,避免了文献[7]中需要将待测图片映射到生成器网络隐空间中的优化代价.Pidhorskyi等<sup>[10]</sup>提出了基于对抗自编码器(AAE, adversarial autoencoder)模型的图片异常检测模型.该模型通过对抗学习过程将编码器隐变量的聚集后验分布映射为先验分布,

由解码器将先验分布映射为训练数据分布,具有更强的数据生成能力.

从信息论的角度看,无监督深度学习网络的训练目标为最大化训练样本与其表示间的互信息<sup>[11]</sup>.Crescimanna等<sup>[12]</sup>提出了以最大化输入样本与其表示间的互信息为AE的目标函数,该方法要求样本在隐空间中的表示服从指数分布.Belghazi等<sup>[13]</sup>提出了一种基于神经网络的互信息估计器(MINE, mutual information neural estimator),该方法将互信息表达为2个随机变量的联合概率分布与边缘分布乘积的KL(Kullback-Leibler)散度,其优势在于,不需要知道随机变量的具体分布形式.Rezaabad等<sup>[14]</sup>提出了在VAE的训练目标函数中加入训练样本与其表示之间的互信息,克服VAE在最小化隐空间表示的聚集后验分布与先验分布的KL散度过程中,减少样本与其表示间互信息的问题.

笔者融合神经互信息估计器MINE和AAE,提出了一种互信息最大化对抗自编码器(IAAE, info-max adversarial autoencoder)半监督异常检测模型,通过显式最大化正常样本与其表示间的互信息,提高样本表示的学习质量,使得AAE模型更好地学习正常样本的数据分布.由测试样本在AAE模型上的重构误差及其在隐空间中的众数散度(MD, mode divergence)计算测试样本异常得分.公共数据集上的实验对比分析表明了所提模型的有效性.

## 1 神经互信息估计器

对于给定的2个随机变量 $X$ 和 $Z$ ,互信息可以表达两者之间的非线性统计相关性. $X$ 和 $Z$ 之间的互信息可以表示为KL散度形式:

$$I(X, Z) = D_{\text{KL}}(P_{XZ} \parallel P_X \otimes P_Z) \quad (1)$$

其中: $P_{XZ}$ 为2个随机变量的联合概率分布, $P_X$ 和 $P_Z$ 为边缘分布, $\otimes$ 为2个边缘分布的乘积, $D_{\text{KL}}(\cdot \parallel \cdot)$ 为计算2个分布间差异的非对称性度量,即KL散度.根据KL散度的对偶表示形式,随机变量间的互信息<sup>[13]</sup>可表达为

$$I_{\theta}(X, Z) = \sup_{\theta \in \Theta} E_{P_{XZ}}[T_{\theta}] - \log(E_{P_X \otimes P_Z}[e^{T_{\theta}}]) \quad (2)$$

其中:  $T_\theta$  为参数为  $\theta \in \Theta$  的深度神经网络表达的函数  $T_\theta: X \times Z \rightarrow \mathbf{R}$ ,  $E_{P_{XZ}}[T_\theta]$  为  $T_\theta$  在分布  $P_{XZ}$  下的期望,  $E_{P_{X \otimes P_Z}}[e^{T_\theta}]$  为  $e^{T_\theta}$  在分布  $P_X \otimes P_Z$  下的期望. Belghazi 等<sup>[13]</sup>以式(2)为优化目标,采用小批量梯度上升算法最大化  $X$  和  $Z$  之间的互信息. 具体的训练算法参见文献[13].

## 2 基于 IAAE 的异常检测

### 2.1 IAAE 模型

AAE 模型与 VAE 模型本质上均为通过分布转换的方式学习训练样本的数据分布. 2 种模型的训练原理均为最小化训练样本  $\mathbf{x}$  的负对数似然上界<sup>[10]</sup>:

$$\begin{aligned} E_{\mathbf{x} \sim p_d(\mathbf{x})} [-\log p(\mathbf{x})] &< \\ E_{\mathbf{x}} [E_{q(\mathbf{z}|\mathbf{x})} [-\log (p(\mathbf{x}|\mathbf{z}))]] &+ \\ E_{\mathbf{x}} [D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))] &\quad (3) \end{aligned}$$

其中:  $p_d(\mathbf{x})$  为训练集的真实分布;  $p(\mathbf{x})$  为模型生成的分布;  $\mathbf{z}$  为编码器输出的隐向量,  $p(\mathbf{z})$  为  $\mathbf{z}$  的先验分布;  $q(\mathbf{z}|\mathbf{x})$  为编码器的输出分布;  $p(\mathbf{x}|\mathbf{z})$  为解码器的输出分布.

式(3)不等式右侧第2项与互信息  $I(\mathbf{x}, \mathbf{z})$  间的关系<sup>[14]</sup>为

$$\begin{aligned} E_{\mathbf{x}} [D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))] &= \\ \int q(\mathbf{x}, \mathbf{z}) \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} d\mathbf{x} d\mathbf{z} &\geq \\ \int q(\mathbf{x}, \mathbf{z}) \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} d\mathbf{x} d\mathbf{z} - D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z})) &= \\ \int q(\mathbf{x}, \mathbf{z}) \left[ \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} - \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right] d\mathbf{x} d\mathbf{z} &= \\ \int q(\mathbf{x}, \mathbf{z}) \log \frac{q(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})} d\mathbf{x} d\mathbf{z} = I(\mathbf{x}, \mathbf{z}) &\quad (4) \end{aligned}$$

其中:  $q(\mathbf{x}, \mathbf{z})$  为  $\mathbf{x}$  和  $\mathbf{z}$  的联合概率分布,  $q(\mathbf{z})$  为边缘分布. 式(3)中右侧第1项可解释为 AAE 模型中的重构误差, AAE 模型通过对抗训练的方式将训练样本  $\mathbf{x}$  的聚集后验分布  $q(\mathbf{z}|\mathbf{x})$  映射为先验分布  $p(\mathbf{z})$ , 即最小化式(3)不等式右侧的第2项. 但由式(4)可知, AAE 模型在此过程中减小了训练样本  $\mathbf{x}$  与其表示  $\mathbf{z}$  间的互信息. 从信息论的角度, 无监督深度模型的训练目标应为最大化训练样本与其表示间的互信息<sup>[11]</sup>. 因此, 在 AAE 模型中显式引入训练样本  $\mathbf{x}$  与其表示  $\mathbf{z}$  之间的互信息最大化估计网络, 以提高训练样本的表示学习质量.

所提 IAAE 模型结构如图1所示. IAAE 模型由

编码器、解码器、鉴别器和神经互信息估计器4部分组成. 图1中  $\mathbf{x}$  为正常数据组成的训练集中的任意一个样本,  $\hat{\mathbf{x}}$  为解码器对样本  $\mathbf{x}$  的重构输出,  $\mathbf{z}$  为编码器对  $\mathbf{x}$  产生的低维隐向量表示,  $\mathbf{z}'$  为从隐向量的任意先验分布  $p(\mathbf{z})$  中的采样样本,  $I(\mathbf{x}, \mathbf{z})$  为由神经互信息估计器计算得到的样本  $\mathbf{x}$  与其低维表示  $\mathbf{z}$  之间的互信息.

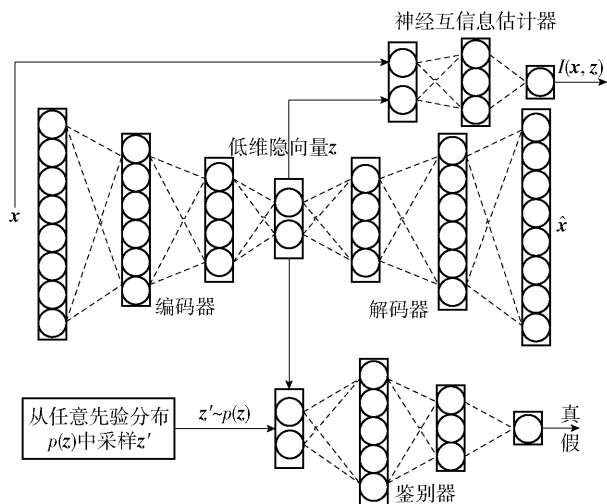


图1 IAAE 模型结构

设由梯度更新网络参数的小批量样本数目为  $m$ , 第  $i$  个样本记为  $\mathbf{x}_i$ , 编码器、解码器、鉴别器和神经互信息估计器对应的映射函数分别为  $Q_\phi(\cdot)$ ,  $G_\varphi(\cdot)$ ,  $D_w(\cdot)$  和  $T_\theta(\cdot)$ , 其中  $\phi, \varphi, w, \theta$  为相应的参数, 则对应的 IAAE 模型的训练过程分为以下3个阶段:

1) 最小化重构误差阶段: 固定鉴别器和神经互信息估计器的参数, 以最小化式(5)所示的均方误差为损失函数, 更新编码器和解码器的网络参数.

$$L_{Q,G} = \frac{1}{m} \sum_{i=1}^m (\mathbf{x}_i - G_\varphi(Q_\phi(\mathbf{x}_i)))^2 \quad (5)$$

2) 对抗正则化阶段: 固定解码器和神经互信息估计器的参数, 以对抗训练的方式更新编码器和鉴别器的网络参数. 首先固定编码器的参数, 循环更新  $k$  次鉴别器的参数, 然后更新一次编码器的参数. 从隐向量的任意先验分布  $p(\mathbf{z})$  中抽取的第  $i$  个样本记为  $\mathbf{z}'_i$ , 鉴别器由最大化下式(6)所示的损失函数更新网络参数. 编码器的训练目标为最大化下式(7)所示的损失函数. 其目标是把编码器生成的隐向量表示的分布映射为任意先验分布  $p(\mathbf{z})$ , 同时最大化样本  $\mathbf{x}_i$  与其由编码器生成的隐向量表示  $Q_\phi(\mathbf{x}_i)$  之间的互信息.

$$L_D = \frac{1}{m} \sum_{i=1}^m [\log(D_\omega(\mathbf{z}'_i)) + \log(1 - D_\omega(Q_\phi(\mathbf{x}_i)))] \quad (6)$$

$$L_Q = \frac{1}{m} \sum_{i=1}^m [\log(D_\omega(Q_\phi(\mathbf{x}_i))) + I(\mathbf{x}_i, Q_\phi(\mathbf{x}_i))] \quad (7)$$

3) 神经互信息估计阶段: 固定编码器、解码器和鉴别器网络参数, 以文献[13]中的算法训练优化神经互信息估计器。

IAAE 模型的具体训练过程如下所示。

初始化编码器、解码器、鉴别器和神经互信息估计器的网络参数;

**for** 每一个训练周期:

从训练集中随机采样  $m$  个样本, 记为  $\{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_m\}$ ;

阶段 1: 固定鉴别器和神经互信息估计器的网络参数, 由梯度下降法和式(5)更新编码器和解码器的网络参数;

$\phi, \varphi \leftarrow \nabla_{\phi, \varphi} L_{Q, G}$

阶段 2: 固定解码器和神经互信息估计器的网络参数, 更新编码器和鉴别器的参数

**for**  $k$  步:

从隐向量的先验分布  $p(\mathbf{z})$  中抽取  $m$  个样本  $\{\mathbf{z}'_1, \dots, \mathbf{z}'_i, \dots, \mathbf{z}'_m\}$ ;

固定编码器的网络参数, 由梯度上升法和式(6)更新鉴别器的参数:

$\omega \leftarrow \nabla_{\omega} L_D$

**end for**

固定鉴别器的网络参数, 由梯度上升法和式(7)更新编码器的参数:

$\phi \leftarrow \nabla_{\phi} L_Q$

阶段 3: 固定编码器、解码器、鉴别器的参数, 使用文献[13]算法更新神经互信息估计器的参数;

**end for**

## 2.2 异常得分计算方法

由正常样本集训练 IAAE 模型收敛后, 根据以下 2 种度量方式检测异常样本: ① 待检测样本在 IAAE 模型上的重构误差; ② 待测样本根据编码器生成的隐向量表示相对先验分布众数的散度 (MD, mode divergence). 将待检测的样本集记为  $\mathbf{X}_{\text{test}} = \{\mathbf{x}^1, \dots, \mathbf{x}^i, \dots, \mathbf{x}^m\}$ , 每个样本  $\mathbf{x}^i$  ( $1 \leq i \leq m$ ) 的重构样本记为  $\hat{\mathbf{x}}^i$ , 重构误差为  $E(\mathbf{x}^i, \hat{\mathbf{x}}^i) = \|\mathbf{x}^i - \hat{\mathbf{x}}^i\|^2$ , 将  $\mathbf{x}^i$  的重构误差归一化处理后作为其在重构误差度

量方式下的异常得分:

$$R(\mathbf{x}^i) = \frac{E(\mathbf{x}^i, \hat{\mathbf{x}}^i) - E_{\min}}{E_{\max} - E_{\min}} \quad (8)$$

其中  $E_{\min}$  和  $E_{\max}$  分别为测试集  $\mathbf{X}_{\text{test}}$  上的最小和最大重构误差值。

隐向量先验分布  $p(\mathbf{z})$  为由  $t$  个各向同性的多元高斯分布组成的高斯混合分布,  $t$  个高斯分布的众数表示为  $\{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^i, \dots, \boldsymbol{\mu}^t\}$ , 每个  $\boldsymbol{\mu}^i$  ( $1 \leq i \leq t$ ) 为第  $i$  个高斯分布的均值. 样本  $\mathbf{x}^i$  ( $1 \leq i \leq m$ ) 由编码器生成的隐向量表示记为  $\mathbf{z}^i$ ,  $\mathbf{x}^i$  在隐空间中的众数散度 MD 定义为:  $O(\mathbf{z}^i) = \min_t \|\mathbf{z}^i - \boldsymbol{\mu}^t\|^2$ . 将  $\mathbf{x}^i$  的众数散度归一化处理后作为其在隐空间散度量方式下的异常得分:

$$M(\mathbf{x}^i) = \frac{O(\mathbf{z}^i) - O_{\min}}{O_{\max} - O_{\min}} \quad (9)$$

其中  $O_{\min}$  和  $O_{\max}$  分别为测试集  $\mathbf{X}_{\text{test}}$  上的最小和最大众数散度值。

每个  $\mathbf{x}^i$  ( $1 \leq i \leq m$ ) 的异常得分值为

$$S(\mathbf{x}^i) = \alpha R(\mathbf{x}^i) + (1 - \alpha) M(\mathbf{x}^i) \quad (10)$$

其中  $\alpha$  为 2 种度量方式的调节因子.  $S(\mathbf{x}^i)$  取值越大, 样本  $\mathbf{x}^i$  越可能为异常。

## 3 实验分析

### 3.1 实验数据集及环境

实验中选取真实的金融欺诈检测数据集<sup>[15]</sup>、Arrhythmia 数据集和 Cardio 数据集 2 个疾病异常检测公共数据集<sup>[16]</sup>. 金融欺诈检测数据集中包含 2 种数值型属性和 6 种类别型属性, 使用 pandas 工具包将其中的类别型属性转化为独热编码, 编码后的特征维度为 618. Arrhythmia 数据集和 Cardio 数据集的特征属性分别由 274 维和 21 维数值型属性组成. 3 个数据集的样本总数、异常样本数及异常样本比例等信息如表 1 所示. 将每个数据集进行归一化处理后, 随机抽取正常样本的 80% 作为训练数据集, 其余的正常样本与异常样本构成测试集. 实验的主要软硬件环境为 Ubuntu18.04, PyTorch1.3; 硬件配置为 Inter i7-6700 CPU, NVIDIA GTX1080 GPU, 16 G 内存。

表 1 3 个数据集的参数

数据集	样本总数	异常样本数	异常样本比例/%
金融欺诈检测	533 009	100	0.019
Arrhythmia	452	66	14.6
Cardio	1 831	176	9.6



3.2 IAAE 异常检测模型参数设置

IAAE 异常检测模型中的 4 个全连接网络在 Cardio 数据集上的层数设置为 3,在其余 2 个数据集上层数设置为 4,在 3 个数据集的具体维数设置如表 2 所示. 解码器和鉴别器的第 4 层均采用 sigmoid 激活函数,其余网络各层均采用 ReLU 激活函数. 使用 Adam 优化器对 IAAE 模型中各网络的参数进行优化,参数  $\beta_1$  和  $\beta_2$  分别设置为 0.5 和 0.999;编码器和解码器的学习率设为 0.001,鉴别器和神经

互信息估计器的学习率设为 0.000 1,式(6)和式(7)中对数函数的底数为 e. 小批量梯度优化算法的小批量样本数设置为:金融欺诈检测数据集为 256,Arrhythmia 数据集为 64,Cardio 数据集为 128. 在 3 个数据集上,IAAE 模型的训练周期数均为 200. 异常得分计算过程中的调节因子  $\alpha$  在各数据集上均设置为 0.5. 异常检测阈值设置为:在金融欺诈检测数据集设为 0.95,在 Arrhythmia 数据集和 Cardio 数据集均设为 0.9.

表 2 IAAE 模型在 3 个数据集的各网络层数的维度设置

数据集	编码器	解码器	鉴别器	神经互信息估计器
金融欺诈检测	618,256,64,32,16	16,32,64,256,618	16,32,8,4,1	634,128,32,16,1
Arrhythmia	274,128,64,32,8	8,32,64,128,274	8,16,32,8,1	282,64,16,8,1
Cardio	21,16,8,4	4,8,16,21	4,8,4,1	25,16,8,1

3.3 IAAE 模型异常检测性能分析

图 2 所示为 IAAE 模型中隐空间先验混合高斯分布的子分布个数  $t$  对异常检测性能的影响结果. 不难看出,金融欺诈检测数据集,Arrhythmia 数据集和 Cardio 数据集中  $t$  值分别为 12,8 和 4 时,F1 值最大. 子分布个数  $t$  设置过大或过小,都会使得 IAAE 模型异常检测性能有明显的下降. 这是因为每个训练数据集在其隐空间的真实分布是确定的,子分布个数  $t$  设置不当,使得先验分布和真实分布差距变大,IAAE 模型的对抗训练过程很难将隐空间分布映射为先验分布. 子分布的个数  $t$  也反映了样本集数据分布的复杂性,实验结果反映了金融欺诈检测数据集、Arrhythmia 数据集的分布复杂性高于 Cardio 数据集.

验分析. 实验过程中,AAE-AD 模型和 IAAE 模型的参数  $Q_\phi, G_\varphi, D_w$  完全相同. 为了消除隐空间先验混合高斯分布中子分布个数对实验结果的影响,2 种模型均采用标准多元正态分布作为低维隐向量的先验分布,以待检测数据的归一化重构误差为异常得分,其他参数均按 3.2 节进行设置. 表 3 所示为 2 种模型在 3 个数据集上异常检测结果的 F1 值. 由表 3 可知,IAAE 模型的 F1 值在金融欺诈检测数据集和 Cardio 数据集上高于 AAE-AD 模型,而在 Arrhythmia 数据集上与 AAE-AD 模型相当. 这是因为 Arrhythmia 数据集的异常比例较高,AAE-AD 模型的学习能力足以区分正常样本和异常样本,使得 IAAE 模型的异常检测性能提升空间较小. 因此,IAAE 模型更适用于异常样本比例较低的场景. 表 3 的实验表明,IAAE 模型能在 AAE-AD 模型的基础上提升训练样本隐空间表示的质量,更好地学习正常样本的分布形态,从而获得较好的异常检测性能.

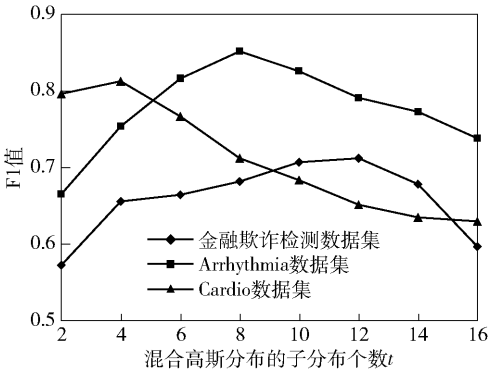


图 2 子分布个数  $t$  对 IAAE 模型的 F1 值的影响

为了验证 IAAE 模型中神经互信息估计网络的作用,将 IAAE 模型和基于 AAE 模型的异常检测(AAE-AD,AAE anomaly detection)模型进行消融实

表 3 IAAE 模型与 AAE-AD 模型在 3 个数据集上的 F1 值对比

数据集	AAE-AD	IAAE
金融欺诈检测	0.669	0.682
Arrhythmia	0.801	0.805
Cardio	0.742	0.754

3.4 IAAE 模型与相关模型对比

将 IAAE 模型与单类支持向量机(OC-SVM,one class support vector machine)<sup>[17]</sup>, DAGMM<sup>[4]</sup>, 基于

AE<sup>[2-3]</sup>, VAE<sup>[5-6]</sup>, GAN<sup>[7]</sup> 的异常检测模型(分别记为 AE-AD, VAE-AD, GAN-AD)做了实验对比. 其中 OC-SVM 采用径向基核函数, 核函数系数为各数据集的特征维数的倒数, 表示支持向量比例下界的参数为各数据集异常样本占总样本数的比例. AE-AD 和 VAE-AD 的模型结构与 IAAE 模型中的编码器、解码器网络结构相同. GAN-AD 的生成器、鉴别器与 IAAE 模型中的编码器、鉴别器网络结构相同. DAGMM 的压缩网络与 IAAE 模型中的编码器、解码器网络结构相同. AE-AD 以检测样本归一化处理后的重构误差值为异常得分. DAGMM 以检测样本归一化处理能量值<sup>[4]</sup>为异常得分. VAE-AD 以检测样本的重建概率值<sup>[5]</sup>为异常得分, GAN-AD 以待测样本在鉴别器中的输出值为异常得分. 以上各模型在 3 个数据集上的异常检测阈值与 3.2 节中 IAAE 模型阈值设置相同.

表 4 至表 6 所示为上述模型在 3 个数据集上的性能对比. 由表可知, OC-SVM 在维度较低的 Cardio 数据集上的 F1 值较高, 但在金融欺诈检测和 Arrhythmia 这 2 个高维数据集上, F1 明显低于其他模型, 这表明, 传统的 OC-SVM 模型无法适用于高维数据上的异常检测. AE-AD 在 2 个高维度数据上的 F1 值低于其他模型, 这是因为 AE-AD 中的重构误差无法较好地地区分正常样本和异常样本. DAGMM 在所有数据上的 F1 值均高于 VAE-AD 和 GAN-AD. 原因是 VAE-AD 将正常样本的表示约束为高斯分布; GAN-AD 通过对抗训练的方式学习正常样本的分布, 该模型训练不稳定, 不能很好地表达正常样本的分布; 而 DAGMM 将正常样本表示、正常样本与其重构样本的差异约束为混合高斯分布, 有较强的区分异常样本的能力. IAAE 模型的 F1 值在 3 个数据集上均高于绝大多数的对比模型. 这是因为 IAAE 模型与 DAGMM 原理类似, 但 IAAE 模型使用对抗训练的方式将正常样本的低维隐空间约束为混合高

表 4 6 种模型在金融欺诈检测数据集上的性能对比

模型	精确率	召回率	F1 值
OC-SVM	0.614	0.597	0.605
AE-AD	0.628	0.609	0.618
VAE-AD	0.646	0.661	0.653
GAN-AD	0.595	0.708	0.647
DAGMM	0.670	0.709	0.689
IAAE	0.708	0.717	0.712

表 5 6 种模型在 Arrhythmia 数据集上的性能对比

模型	精确率	召回率	F1 值
OC-SVM	0.622	0.689	0.654
AE-AD	0.735	0.710	0.722
VAE-AD	0.734	0.757	0.745
GAN-AD	0.715	0.796	0.753
DAGMM	0.862	0.823	0.842
IAAE	0.843	0.861	0.852

表 6 6 种模型在 Cardio 数据集上的性能对比

模型	精确率	召回率	F1 值
OC-SVM	0.827	0.818	0.822
AE-AD	0.671	0.653	0.662
VAE-AD	0.774	0.769	0.634
GAN-AD	0.679	0.711	0.695
DAGMM	0.751	0.734	0.742
IAAE	0.805	0.824	0.813

斯分布, 同时使用神经互信息估计器最大化正常样本与其低维隐向量表示之间的互信息, 得到了高质量的正常样本表示.

4 结束语

提出了一种融合互信息估计和对抗自编码器的半监督异常检测模型, 所提模型在编码器的训练目标函数中引入训练样本及其表示的互信息正则化项、增设了互信息最大化估计网络, 克服了对抗自编码器无法显示学习训练样本及其表示之间互信息的问题. 实验结果表明, 所提模型能提高正常样本的表示学习质量, 更好地学习样本分布. 下一步将把所提模型拓展应用于时间序列、图像等更复杂类型的数据.

参考文献:

[1] Pang Guansong, Shen Chunhua, Cao Longbing, et al. Deep learning for anomaly detection: a review[J]. ACM Computing Surveys, 2021, 54(2): 1-38.

[2] Ribeiro M, Lazzaretti A E, Lopes H S. A study of deep convolutional auto-encoders for anomaly detection in videos[J]. Pattern Recognition Letters, 2018, 105: 13-22.

[3] Zhou Chong, Paffenroth R C. Anomaly detection with robust deep autoencoders[C]//23rd ACM SIGKDD International Conference on Knowledge Discovery and Data

- Mining. Halifax: ACM, 2017: 665-674.
- [4] Zong Bo, Song Qi, Min M R, et al. Deep autoencoding Gaussian mixture model for unsupervised anomaly detection [C] // International Conference on Learning Representations. Vancouver: IEEE, 2018: 1-19.
- [5] An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability [J]. Special Lecture on IE, 2015, 2(1): 1-18.
- [6] Yao Rong, Liu Chongdang, Zhang Linxuan, et al. Unsupervised anomaly detection using variational auto-encoder based feature extraction [C] // 2019 IEEE International Conference on Prognostics and Health Management (ICPHM). San Francisco: IEEE, 2019: 1-7.
- [7] Schlegl T, Seeböck P, Waldstein S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery [C] // International Conference on Information Processing in Medical Imaging. Boone: Springer Cham, 2017: 146-157.
- [8] Zenati H, Romain M, Foo C S, et al. Adversarially learned anomaly detection [C] // 2018 IEEE International Conference on Data Mining (ICDM). Singapore: IEEE, 2018: 727-736.
- [9] Li Chunyuan, Liu Hao, Chen Changyou, et al. ALICE: towards understanding adversarial learning for joint distribution matching [C] // Advances in Neural Information Processing Systems. Long Beach: MIT, 2017: 5495-5503.
- [10] Pidhorskyi S, Almohsen R, Doretto G. Generative probabilistic novelty detection with adversarial autoencoders [C] // Advances in Neural Information Processing Systems. Montréal: MIT, 2018: 6822-6833.
- [11] Ruff L, Vandermeulen R A, Görnitz N, et al. Deep semi-supervised anomaly detection [C] // International Conference on Learning Representations. Addis Ababa: IEEE, 2020: 1-13.
- [12] Crescimanna V, Graham B. An information theoretic approach to the autoencoder [C] // INNS Big Data and Deep Learning Conference. Genoa: Springer Cham, 2019: 99-108.
- [13] Belghazi M I, Baratin A, Rajeshwar S, et al. Mutual information neural estimation [C] // International Conference on Machine Learning. Stockholm: PMLR, 2018: 531-540.
- [14] Rezaabad A L, Vishwanath S. Learning representations by maximizing mutual information in variational autoencoders [C] // 2020 IEEE International Symposium on Information Theory (ISIT). Los Angeles: IEEE, 2020: 2729-2734.
- [15] Marco S. Financial fraud detection dataset: version2 [EB/OL]. (2018-05-11) [2021-01-15]. <https://github.com/gitiHubi/deepAD>.
- [16] Shebuti R. Outlier detection datasets [EB/OL]. (2016-01-01) [2021-01-15]. New York: Stony Brook University. <http://odds.cs.stonybrook.edu>.
- [17] Erfani S M, Rajasegarar S, Karunasekera S, et al. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning [J]. Pattern Recognition, 2016, 58: 121-134.