计算机科学
COMPUTER SCIENCE

基于对抗样本和自编码器的鲁棒异常检测

李沙沙，邢红杰

引用本文

李沙沙，邢红杰. 基于对抗样本和自编码器的鲁棒异常检测[J]. 计算机科学，2024，51(5)：363-373.

LI Shasha, XING Hongjie. Robust Anomaly Detection Based on Adversarial Samples and AutoEncoder [J]. Computer Science, 2024, 51(5): 363-373.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

面向流动追踪的多层鲁棒Excel文档水印

Robust and Multilayer Excel Document Watermarking for Source Tracing

计算机科学, 2024, 51(5): 374-381. https://doi.org/10.11896/jsjkx.230300192

CheatKD:基于毒性神经元同化的知识蒸馏后门攻击方法

CheatKD:Knowledge Distillation Backdoor Attack Method Based on Poisoned Neuronal Assimilation

计算机科学, 2024, 51(3): 351-359. https://doi.org/10.11896/jsjkx.221200035

一种基于变分多跳图注意力编码器的深层协同真值发现

Deep Collaborative Truth Discovery Based on Variational Multi-hop Graph Attention Encoder

计算机科学, 2024, 51(3): 109-117. https://doi.org/10.11896/jsjkx.221200063

基于隐空间扩散模型的差分隐私数据合成方法研究

Differential Privacy Data Synthesis Method Based on Latent Diffusion Model

计算机科学, 2024, 51(3): 30-38. https://doi.org/10.11896/jsjkx.230700177

一种抗屏摄攻击的DCT域深度水印方法

Screen-shooting Resilient DCT Domain Watermarking Method Based on Deep Learning

计算机科学, 2024, 51(2): 343-351. https://doi.org/10.11896/jsjkx.221200121

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300153

基于对抗样本和自编码器的鲁棒异常检测

李沙沙　邢红杰
河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室　河北 保定 071002
(lss96133@163.com)

摘　要　基于自编码器的异常检测方法仅利用正常样本进行训练,因此可以有效地重构正常样本,但不能较好地对异常样本进行重构。另外,当基于自编码器的异常检测方法受到对抗攻击时,往往会取得错误的检测结果。为了解决上述问题,提出了一种基于对抗样本和自编码器的鲁棒异常检测(Robust Anomaly Detection Based on Adversarial Samples and AutoEncoder,RAD-ASAE)方法。RAD-ASAE 由两个参数共享的编码器和一个解码器构成。首先,对正常样本施加微小的扰动以生成对抗样本,利用正常样本和对抗样本同时对模型进行训练,以提高模型的对抗鲁棒性;其次,在样本空间中最小化对抗样本的重构误差以及正常样本与对抗样本的重构样本之间的均方误差,同时在潜在空间中最小化正常样本和对抗样本的潜在特征之间的均方误差,以提高自编码器的重构能力。在 MNIST,Fashion-MNIST,CIFAR-10 数据集上进行实验,结果表明,与 7 种相关方法相比,RAD-ASAE 展现了更优的异常检测性能。

关键词:自编码器;对抗样本;异常检测;对抗攻击;鲁棒性
中图分类号　TP391.4

Robust Anomaly Detection Based on Adversarial Samples and AutoEncoder

LI Shasha and XING Hongjie
Hebei Key Laboratory of Machine Learning and Computational Intelligence,College of Mathematics and Information Science,Hebei University,Baoding,Hebei 071002,China

Abstract　The anomaly detection method based on AutoEncoder only uses normal samples for training,so it can effectively reconstruct normal samples,but cannot reconstruct abnormal samples. In addition,when the anomaly detection method based on AutoEncoder is attacked by adversarial attacks,it often obtains wrong detection results. In order to solve the above problems,robust anomaly detection based on adversarial samples and AutoEncoder(RAD-ASAE) method is proposed. RAD-ASAE consists of two parameter-shared encoders and a decoder. First,the normal sample is perturbed slightly to generate the adversarial sample,and normal samples and adversarial samples are used to train the model at the same time to improve the adversarial robustness of the model. Second,the reconstruction error of the adversarial samples is minimized in the sample space,and the mean square error between the normal samples and the reconstructed samples of the adversarial samples is minimized. At the same time,the mean square error between the latent features of the normal samples and the adversarial samples is minimized in the latent space to improve the reconstruction ability of the AutoEncoder. Experimental results on MNIST,Fashion-MNIST and CIFAR-10 show that RAD-ASAE demonstrates better detection performance in comparison with 7 related methods.

Keywords　AutoEncoder,Adversarial samples,Anomaly detection,Adversarial attack,Robustness

1　引言

在实际生活中,异常检测模型被广泛地应用于信息安全[1]、医疗[2]、工业生产[3]等领域。异常检测问题可被视为“单类分类”任务,即在训练阶段,仅使用正常样本训练异常检测模型,在测试阶段,利用训练好的模型将测试样本分类为正常样本或异常样本。Pimentel 等[4]将传统的浅层异常检测算法划分为以下 5 类。

1)基于概率的异常检测算法,主要包括参数化方法和非参数化方法。如 Ilonen 等[5]利用近似方法计算高斯混合模型中概率密度函数的置信度,并根据置信度进行异常检测。Zhang 等[6]提出了基于核密度波动的异常检测算法,该算法结合密度波动思想,将核密度波动因子用作异常检测的评价指标。但该类算法在规模较小的数据集上性能较差。

到稿日期:2023-03-18　返修日期:2023-07-09
基金项目:国家自然科学基金(61672205);河北省自然科学基金(F2017201020);河北大学高层次人才科研启动项目(521100222002)
This work was supported by the National Natural Science Foundation of China(61672205),Natural Science Foundation of Hebei Province,China(F2017201020) and High-Level Talents Research Start-Up Project of Hebei University(521100222002).
通信作者:邢红杰(hjxing@hbu.edu.cn)

2)基于距离的异常检测算法,主要包括最近邻方法和基于聚类的方法。如 Ding[7] 提出了基于 k 近邻隔离森林的异常检测算法,该算法采用超球体作为隔离工具,利用 k 近邻构建隔离森林,并采用基于距离的异常值计算方法。Fan 等[8] 提出了基于特征选择的 K-means 聚类异常检测算法,该算法选取距离样本中心最远的样本作为初始的聚类中心,并采用信息熵属性权重计算方法对样本进行划分。该类算法的时间复杂度过高,因此不适用于处理大规模高维数据。

3)基于区域的异常检测算法,主要包括支持向量数据描述和单类支持向量机。如 Turkoz 等[9] 提出了广义支持向量数据描述算法,引入广义的贝叶斯算法为每个类构建超球体,实现多模态情况下的异常检测。Kittidachanan 等[10] 提出了基于网格搜索单类支持向量机分类的异常检测算法,该算法使用网格搜索方法估计单类支持向量机的超参数。该类算法较难选取合理控制边界区域大小的参数。

4)基于重构的异常检测算法,主要包括基于神经网络的方法和基于子空间的方法。如 Singh 等[11] 提出了一种用于图像区域分类的异常检测算法,该算法使用特征选择机制对图像区域进行编码,并将多层感知器用作分类器。Li 等[12] 提出了鲁棒监督子空间学习算法,该算法对输出预测和输入重构进行联合优化,用于构造鲁棒的子空间,通过从目标函数中分离出稀疏项来抑制异常样本的影响。该类算法的计算成本较高,并且它们的性能对参数的不同取值尤为敏感。

5)基于信息论的异常检测算法。如 Callegari 等[13] 提出了网络通信量中的信息论异常检测算法,该算法通过三维可逆草图对通信量进行聚合,计算不同通信量描述符的熵以进行异常检测。这类算法在处理异常样本较少的数据集时检测效果较差。

为了弥补浅层异常检测方法的不足,深度学习在异常检测领域得到了广泛应用。在基于深度学习的异常检测方法中,自编码器(AutoEncoder,AE)[14] 是一种常用的网络结构,它由编码器和解码器组成,输入样本经过编码器后获得潜在特征,潜在特征再经过解码器得到重构样本。基于自编码器的异常检测方法将输入样本和重构样本之间的重构误差用作异常得分,异常得分高的样本被视为异常样本,异常得分低的样本被视为正常样本。虽然基于自编码器的异常检测方法取得了较优的检测性能,但它在正常样本上所取得的重构误差与部分异常样本的重构误差非常接近,导致这些异常样本很难被正确检测[15]。为了解决该问题,相关学者提出了许多改进方法。Xia 等[16] 提出了用于无监督异常值去除的判别重构算法,该算法在自编码器的训练过程中逐步添加判别信息,使正常样本和异常样本更具有区分性,通过重构误差进行异常检测。Gong 等[17] 提出了记忆增强自编码器(Memory-Augmented AutoEncoder,MemAE),MemAE 将深度自编码器和内存模块相结合,在训练过程中利用内存模块存储更多正常样本的潜在特征,测试时从内存模块中检索出与待测样本的潜在特征相关性最大的存储项并进行重构,通过重构误差对正常样本和异常样本进行分类。An 等[18] 提出了基于变分自编码器的重构概率异常检测算法,该算法利用变分自编码器对样本进行训练,使用样本分布的概率度量替代传统自编码器的重构误差,以区分正常样本和异常样本。Zhou 等[19] 提出了一种鲁棒自编码器,将训练集拆分为正常样本子集和包含异常值和噪声的样本子集,提取正常样本的特征,调整目标函数中包含异常值和噪声的样本子集的稀疏程度,以提高自编码器的鲁棒性。

此外,当基于深度学习的异常检测方法受到对抗攻击时,往往会取得错误的检测结果。对抗样本[20] 指在数据集中通过故意添加细微的扰动所形成的样本,该扰动是通过求解生成对抗样本的优化问题,得到使模型产生错误分类结果的最小扰动。利用添加对抗扰动的训练集对网络进行训练,会使模型以高置信度产生错误输出,如图 1 所示,攻击者利用快速梯度符号法[21](Fast Gradient Sign Method,FGSM)向食蜂鸟的原始样本中添加扰动,图像识别模型把生成的对抗样本以较高的置信度错误地识别为海豚,且原始样本与对抗样本在视觉上难以区分。相关研究表明,基于自编码器的异常检测方法容易受到被恶意添加到输入样本中的扰动所影响,导致网络产生错误的输出。为了解决该问题,Salehi 等[22] 提出了自编码器改进异常检测的对抗性鲁棒训练(Adversarially Robust Training of AutoEncoder Improves Novelty Detection,ARAE),该方法首先生成对抗样本,然后利用对抗样本和正常样本一起训练自编码器,以提高模型的鲁棒性,但该方法忽略了自编码器对输入样本的重构约束,导致检测效果欠佳。此外,由于训练集仅由正常样本构成,因此自编码器无法学习到异常样本。当异常样本与正常样本较为相似,或者异常样本的重构误差较小时,该方法的检测性能较差。

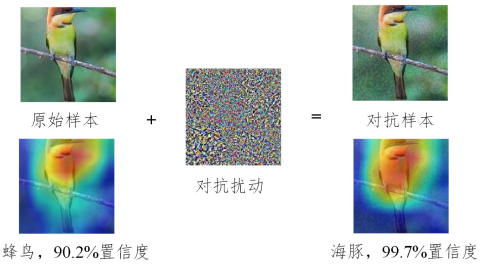


图 1 原始样本与 FGSM 生成对抗样本的置信度

Fig. 1 Confidence of original sample and adversarial sample generated by FGSM

为了避免基于自编码器的异常检测方法受到对抗攻击的影响,提高该类方法的对抗鲁棒性和检测性能,因此采用对抗攻击和防御技术,提出了基于对抗样本和自编码器的鲁棒异常检测(Robust Anomaly Detection Based on Adversarial Samples and AutoEncoder,RAD-ASAE)方法。所提方法仅采用正常样本生成对抗样本,无需使用异常样本,降低了数据集的复杂性。利用正常样本和对抗样本进行双空间联合训练,充分学习样本的多层次表示以提高自编码器的重构能力,从而进行有效地异常检测。RAD-ASAE 的训练过程包括两个阶段:第一阶段,通过使正常样本与对抗样本的潜在特征距离最远来构造优化问题并对其进行求解,以获得每个正常样本的最优扰动,生成相应的对抗样本;第二阶段,对样本空间中对抗样本的重构误差、正常样本与对抗样本的重构样本之间的均方误差,以及潜在空间中正常样本和对抗样本的潜在

特征之间的均方误差同时进行最小化，以获得最优的网络参数，对双空间进行联合训练以提高自编码器的重构能力。所提 RAD-ASAE 的主要贡献为：

1）通过求解相应的优化问题，为每个正常样本构造最佳的对抗样本，利用正常样本与对抗样本所组成的训练集对模型进行训练，能有效提高模型的对抗鲁棒性。

2）模型在样本空间中训练能使对抗样本与其重构样本、正常样本与对抗样本的重构样本更加接近，在潜在空间中训练能使正常样本的潜在特征与对抗样本的潜在特征更加接近。此外，在双空间中对模型进行联合训练，能使模型充分学习样本的多层次表示，有效提高自编码器的重构能力。

3）为模型开展参数选取分析实验、消融实验和鲁棒性分析实验，并将所提方法与现有的异常检测方法进行了实验比较，结果表明所提方法具有更优的对抗鲁棒性和检测性能。

本文第 2 章简要介绍了基于自编码器的异常检测方法、对抗攻击和防御方法；第 3 章详细介绍了所提方法的模型结构、问题公式化和算法实现；第 4 章通过实验验证了所提方法的有效性；最后对所提方法进行了总结并对未来的研究工作进行了展望。

2 相关工作

2.1 基于自编码器的异常检测方法

自编码器（AE）是一种由编码器和解码器构成的前馈神经网络，具有较强的特征提取能力。AE 使用其编码器将输入样本映射到低维空间以获得潜在特征，再经过解码器得到重构样本，且在训练过程中以尽可能小的误差对输入样本进行重构。AE 的训练过程如图 2 所示。

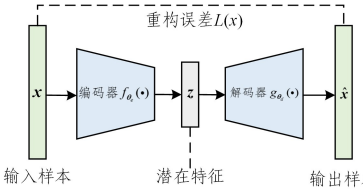


图 2 自编码器的训练过程

Fig. 2 Training process of autoencoder

对于给定的输入样本 x，编码器输出的潜在特征 z 可以表示为：

$$\boldsymbol{z}=f_{\theta_e}(\boldsymbol{W}_e\boldsymbol{x}+\boldsymbol{b}_e) \tag{1}$$

其中，编码器的网络参数为 $\boldsymbol{\theta}_e=\{\boldsymbol{W}_e,\boldsymbol{b}_e\}$，$\boldsymbol{W}_e$ 和 $\boldsymbol{b}_e$ 分别为编码器的权重和偏差项，$f_{\theta_e}(\cdot)$ 为编码器的激活函数。

解码器输出的重构样本 $\hat{\boldsymbol{x}}$ 可以表示为：

$$\hat{\boldsymbol{x}}=g_{\theta_d}(\boldsymbol{W}_d\boldsymbol{z}+\boldsymbol{b}_d) \tag{2}$$

其中，解码器的网络参数为 $\boldsymbol{\theta}_d=\{\boldsymbol{W}_d,\boldsymbol{b}_d\}$，$\boldsymbol{W}_d$ 和 $\boldsymbol{b}_d$ 分别为解码器的权重和偏差项，$g_{\boldsymbol{\theta}_d}(\cdot)$ 为解码器的激活函数。

自编码器重构样本的能力主要取决于学习到的潜在特征，学习到的潜在特征越好，重构样本的能力越强。通过随机梯度下降（Stochastic Gradient Descent，SGD）[23] 算法最小化 $\boldsymbol{x}$ 和 $\hat{\boldsymbol{x}}$ 之间的重构损失 $L(\boldsymbol{x})$，以更新网络参数，从而使自编码器网络重构出与输入样本 $\boldsymbol{x}$ 尽可能相似的样本 $\hat{\boldsymbol{x}}$，即：

$$\min_{\boldsymbol{\theta}_e,\boldsymbol{\theta}_d} L(\boldsymbol{x})=\min_{\boldsymbol{\theta}_e,\boldsymbol{\theta}_d}\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2 \tag{3}$$

其中，$\|\cdot\|_2$ 表示 ℓ_2 范数。

在测试阶段，计算每个待测样本的重构误差。因为仅有正常样本参与训练，所以 AE 只能学习正常样本的潜在特征。对于异常样本，AE 的重构能力相对较差，因此会产生较大的重构误差。据此，基于 AE 的异常检测方法将重构误差作为待测样本的异常得分，即：

$$L(\boldsymbol{x})=\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|^2 \tag{4}$$

当 $L(\boldsymbol{x})$ 的值大于预先设定的阈值时，则认为 $\boldsymbol{x}$ 是异常样本；否则认为 $\boldsymbol{x}$ 是正常样本。

2.2 对抗攻击

对抗样本指在样本中施加细微的扰动，使目标模型产生错误的分类结果。对抗攻击可以被看作是生成对抗样本的过程，根据攻击者对目标模型的体系结构和参数的了解情况，对抗攻击可以分为白盒攻击和黑盒攻击[24]。

白盒攻击指攻击者完全了解目标模型的全部信息。目前主流的白盒攻击算法包括基于梯度的攻击算法和基于优化的攻击算法。

1）基于梯度的攻击算法。该算法指目标模型根据正常样本的梯度信息获得使模型损失值最大的扰动，将扰动添加到正常样本中使模型产生错误的输出结果。如 FGSM 算法沿正常样本梯度的反方向添加使模型损失最大的扰动，从而降低分类置信度。在 FGSM 算法中对抗样本定义为：

$$\boldsymbol{x}'=\boldsymbol{x}+\alpha\cdot sign[\nabla_x J(\boldsymbol{\theta},\boldsymbol{x},y)] \tag{5}$$

其中，$\boldsymbol{x}'$ 为对抗样本，α 为对抗扰动的强度，∇_x 为关于 $\boldsymbol{x}$ 的梯度，J 为损失函数，$\boldsymbol{\theta}$ 为模型参数，y 为 $\boldsymbol{x}$ 的目标输出。

Kurakin 等[25] 提出的迭代快速梯度符号方法（Iterative Fast Gradient Sign Method，I-FGSM）在 FGSM 基础上进行多次迭代，在 I-FGSM 算法中对抗样本定义为：

$$\begin{aligned}&\boldsymbol{x}_0'=\boldsymbol{x}\\&\boldsymbol{x}'_{t+1}=Clip_\varepsilon\{\boldsymbol{x}_t'+\alpha\cdot sign[\nabla_x J(\boldsymbol{\theta},\boldsymbol{x},y)]\}\end{aligned} \tag{6}$$

其中，$Clip_\varepsilon\{\}$ 为逐像素的裁剪函数，使对抗样本位于 $\boldsymbol{x}$ 的 ε 邻域内，t 为迭代次数。

Xiao 等[26] 提出了压缩迭代快速梯度符号算法，该算法通过减少对现有累积梯度实体的依赖，消除了迭代累积机制形成的冗余扰动，从而生成像素变化较少的对抗样本。但该算法在目标性对抗攻击中性能较差。

Madry 等[27] 提出的投影梯度下降（Projected Gradient Descent，PGD）算法同样进行多次迭代，与 I-FGSM 相比，PGD 算法使用均匀分布的随机噪声作为初始扰动，且对梯度进行投影操作而非裁剪操作。在 PGD 算法中对抗样本定义为：

$$\boldsymbol{x}'_{t+1}=proj_{x+S}\{\boldsymbol{x}_t'+\alpha\cdot sign[\nabla_x J(\boldsymbol{\theta},\boldsymbol{x},y)]\} \tag{7}$$

其中，S 为一组扰动，$proj_{x+S}$ 为投影函数。

2）基于优化的攻击算法。该算法指通过优化目标函数来保证正常样本与对抗样本的距离最小，对正常样本添加最小的扰动生成对抗样本。如 Carlini 等提出了 C&W[28] 算法，该算法根据不同的范数生成尽可能小的扰动以得到对抗样本。在 C&W 算法中，对抗样本可通过求解下面的优化问题得到。

$$\min_{\delta} \|\delta\|_p + c \cdot f(\boldsymbol{x}+\delta)$$
$$\text{s.t. } \boldsymbol{x}+\delta \in [0,1]^n \tag{8}$$

其中，$\|\cdot\|_p$ 表示 p 范数，δ 为扰动，c 为超参数。值得注意的是，C&W 算法非常耗时，因此并未得到广泛应用。

黑盒攻击指攻击者不了解目标模型的体系结构和参数情况。黑盒攻击的算法主要包括基于迁移的攻击算法，攻击者通过对目标模型进行查询和访问来构建合成模型，攻击者在合成模型上使用白盒攻击方法生成对抗样本，最后将该对抗样本迁移到目标模型上进行攻击。如 Lin 等[29]提出了敏感区域感知的黑盒对抗攻击算法，该算法使用可转移的黑盒灵敏度图生成方法定位正常样本的敏感区域，采用 l_0 和 l_∞ 动态调整策略以及自适应进化策略，进行有效且难以察觉的对抗攻击。Papernot 等[30]提出了本地替代模型攻击算法，该算法通过生成替代模型来模拟目标模型的决策边界，使用替代模型生成对抗样本，且在训练过程中使用雅克比矩阵减少目标模型的查询次数。此外，与以往攻击神经网络输出层的方法不同，Zhou 等[31]选择对中间特征层进行攻击，在特征空间中最大化正常样本与对抗样本的距离，引入扰动的平滑正则化，以进行有效攻击。

2.3 对抗防御

自对抗攻击被提出以来，相关学者提出了许多关于对抗样本的防御方法。目前将现有的防御方法分为以下 3 类[32]。

1)数据预处理。在对抗样本输入之前对其进行修改，消除或减轻对抗扰动，以缓解对抗攻击的影响。主要包括对抗样本去噪、数据压缩和像素偏移。Liao 等[33]设计了一种高级表示去噪引导器来消除对抗样本中不可察觉的噪声，以达到更好的防御效果。Liu 等[34]提出了基于对抗训练和图像去噪的对抗样本防御算法，该算法将非局部均值去噪算法与对抗训练相结合，在训练阶段对正常样本和对抗样本进行扰动去噪。Jia 等[35]提出了一种图像压缩模型来防御对抗样本，该模型使用压缩卷积神经网络以保持原始图像的结构信息和净化对抗扰动，并使用重构卷积神经网络对原始图像进行高质量重构。Prakash 等[36]提出了像素偏移的防御机制，该算法重新分配像素值来局部破坏图像，使用小波变换平滑对抗样本，减小对抗攻击的影响。该类算法会造成输入样本关键信息的丢失，导致产生错误的分类结果。

2)增强模型的鲁棒性。在深度神经网络中添加更多的层、子网络或者增强神经网络的泛化能力来提高模型鲁棒性，主要包括对抗训练和防御蒸馏。Jia 等[37]提出了可学习攻击策略的对抗训练，该方法通过学习自动生成攻击策略，在不同的训练阶段采用不同的攻击策略以增强模型的鲁棒性。Papernot 等[38]在蒸馏方法的基础上提出了防御蒸馏，能够有效地防御白盒攻击。该类算法需要重新训练网络，计算成本较大，并且面对精心设计的对抗攻击时，防御效果欠佳。

3)检测对抗样本。使用神经网络通过阈值策略区分正常样本和对抗样本。如在 Samangouei 等[39]提出的 Defense-GAN 网络中，生成模型 G 模拟数据的分布，对抗模型 A 判断输入的样本是正常样本还是对抗样本。Meng 等[40]提出的 MagNet 网络利用检测器区分正常样本和对抗样本，利用重组网络重构对抗样本，以提高网络的分类能力。Chen 等[41]提出了深度神经网络对抗防御的显著特征提取算法，该算法利用生成式对抗网络模型对正常样本和对抗样本的显著特征和平凡特征进行耦合提取，通过特征分离以及比较显著特征和平凡特征之间的差值，来实现对正常样本和对抗样本的检测。

3 基于对抗样本和自编码器的鲁棒异常检测

本章将从生成对抗样本、对抗训练与测试、算法实现 3 方面对所提 RAD-ASAE 进行详细描述。

3.1 生成对抗样本

如前文所述，当受到对抗攻击时，基于自编码器的异常检测方法往往会取得错误的检测结果。为了解决该问题，可以为正常样本集 $\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^N$ 构造相应的对抗样本集 $\boldsymbol{X}^*=\{\boldsymbol{x}_i^*\}_{i=1}^N$。正常样本 $\boldsymbol{x}_i$ 和样本 $\boldsymbol{x}_i+\delta_i$ 经过编码器后所得的潜在特征表示为：

$$\boldsymbol{z}_i = f_{\boldsymbol{\theta}_e}(\boldsymbol{W}_e\boldsymbol{x}_i+\boldsymbol{b}_e) \tag{9}$$
$$\boldsymbol{z}_i^* = f_{\boldsymbol{\theta}_e}(\boldsymbol{W}_e(\boldsymbol{x}_i+\delta_i)+\boldsymbol{b}_e) \tag{10}$$

其中，编码器的网络参数为 $\boldsymbol{\theta}_e=\{\boldsymbol{W}_e,\boldsymbol{b}_e\}$，$\boldsymbol{W}_e$ 和 $\boldsymbol{b}_e$ 分别表示权重和偏差项，$f_{\boldsymbol{\theta}_e}(\cdot)$为编码器的激活函数，$\delta_i$ 为扰动。

微小的扰动 δ_i 可以保证正常样本 $\boldsymbol{x}_i$ 与样本 $\boldsymbol{x}_i+\delta_i$ 在视觉上高度相似，但样本 $\boldsymbol{x}_i+\delta_i$ 会使分类器产生错误的分类结果，以实现有效攻击，如图 1 所示。通过使 $\boldsymbol{x}_i$ 经编码器所得潜在特征 $\boldsymbol{z}_i$ 与 $\boldsymbol{x}_i+\delta_i$ 经编码器所得潜在特征 $\boldsymbol{z}_i^*$ 相距最远，得到最小扰动 δ_i。基于此，可得如下优化问题：

$$\max_{\delta_i} L_z = \|\boldsymbol{z}_i-\boldsymbol{z}_i^*\|_2^2$$
$$\text{s.t. } \|\delta_i\|_\infty \leqslant \varepsilon_i \tag{11}$$

其中，$\|\cdot\|_\infty$ 为 ℓ_∞ 范式，ε_i 为攻击强度。

利用投影梯度下降算法对优化问题(9)进行求解，在满足 $\|\delta_i\|_\infty \leqslant \varepsilon_i$ 时，使 $\boldsymbol{z}_i$ 与 $\boldsymbol{z}_i^*$ 的距离最远从而得到扰动 δ_i。该扰动是添加到正常样本中的噪声，使其成为对抗样本，此时获得的 $\boldsymbol{x}_i+\delta_i$ 为样本 $\boldsymbol{x}_i$ 相应的最优对抗样本 $\boldsymbol{x}_i^*$，即：

$$\boldsymbol{x}_i^* = \boldsymbol{x}_i+\delta_i \tag{12}$$

由此获得对抗样本集 $\boldsymbol{X}^*=\{\boldsymbol{x}_i^*\}_{i=1}^N$。

3.2 对抗训练与测试

在 RAD-ASAE 对抗训练过程中，利用正常样本与对抗样本所组成的训练集对模型进行训练，以提高其对抗鲁棒性。RAD-ASAE 在双空间进行联合训练，通过对样本空间和潜在空间中的损失函数进行最小化来充分学习样本中所蕴含的多层次表示。RAD-ASAE 的对抗训练过程如图 3 所示。

此外，对抗样本的潜在特征 $\boldsymbol{z}_i^*$ 经过解码器后所得的重构样本为：

$$\hat{\boldsymbol{x}}_i^* = g_{\boldsymbol{\theta}_d}(\boldsymbol{W}_d\boldsymbol{z}_i^*+\boldsymbol{b}_d) \tag{13}$$

其中，解码器的网络参数为 $\boldsymbol{\theta}_d=\{\boldsymbol{W}_d,\boldsymbol{b}_d\}$，$\boldsymbol{W}_d$ 和 $\boldsymbol{b}_d$ 分别为权重和偏差项，$g_{\boldsymbol{\theta}_d}(\cdot)$为解码器的激活函数。

为了提高 AE 的重构能力，充分学习样本空间中的样本表示，可将对抗样本 $\boldsymbol{x}_i^*$ 与经过自编码器后得到的对抗样本的重构样本 $\hat{\boldsymbol{x}}_i^*$ 之间的均方误差作为重构损失，通过最小化该重构损失来提高样本空间中对抗样本与对抗样本的重构样本之间的相似性。因此，样本空间中对抗样本的

重构损失 L_c 表示为：

$$L_c=\frac{1}{N}\sum_{i=1}^{N}\| \boldsymbol{x}_i^*-\hat{\boldsymbol{x}}_i^* \|_2^2 \tag{14}$$

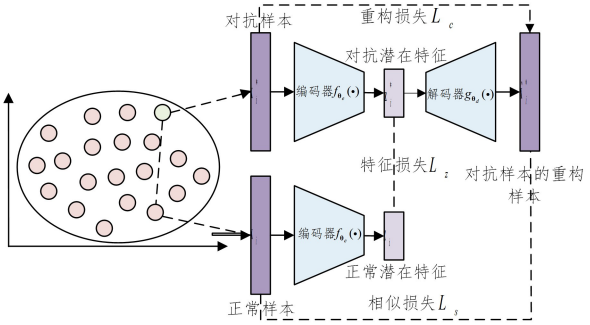


图 3 RAD-ASAE 的训练过程

Fig. 3 Training process of RAD-ASAE

在样本空间中，正常样本与对抗样本不仅要在视觉上具有相似性，还应该与对抗样本的重构样本更加相似。因此，为了提高样本空间中各样本间的相似性，在 RAD-ASAE 中引入相似损失 L_s，即计算正常样本与对抗样本的重构样本之间的均方误差，表示为：

$$L_s=\frac{1}{N}\sum_{i=1}^{N}\| \boldsymbol{x}_i-\hat{\boldsymbol{x}}_i^* \|_2^2 \tag{15}$$

此外，潜在空间中包含更为丰富的样本特征，为了使正常样本经过编码器后得到的潜在特征 $\boldsymbol{z}_i$ 与对抗样本经过编码器后得到的潜在特征 $\boldsymbol{z}_i^*$ 更为相似，在潜在空间中引入特征损失 L_z，即计算正常样本与对抗样本所得潜在特征之间的均方误差，表示为：

$$L_z=\frac{1}{N}\sum_{i=1}^{N}\| \boldsymbol{z}_i-\boldsymbol{z}_i^* \|_2^2 \tag{16}$$

综上，采用样本空间和潜在空间的 3 个损失 L_c，L_s，L_z 对模型进行约束，能有效提高自编码器的重构能力，并且有效提高 RAD-ASAE 的对抗鲁棒性和检测性能。最终 RAD-ASAE 进行双空间联合训练的总损失为：

$$L_{\text{RAD-ASAE}}=\alpha L_c+\beta L_z+\gamma L_s \tag{17}$$

其中，α，β，γ 为 $L_{\text{RAD-ASAE}}$ 的平衡参数，$\alpha+\beta+\gamma=1$，在经过实验分析后，MNIST 和 Fashion-MNIST 设置为 $\alpha=\gamma=0.4$ 和 $\beta=0.2$，CIFAR-10 设置为 $\alpha=\gamma=0.1$ 和 $\beta=0.8$。

在 RAD-ASAE 测试过程中，仅需使用训练好的编码器和解码器对待测样本进行分类。假设给定的待测样本是正常样本，编码器先将其编码为正常样本的潜在特征，再经过解码器重构成正常样本，使正常样本获得较小的重构误差。假设给定的待测样本是异常样本，编码器先将其编码为异常样本的潜在特征，其次经过解码器获得重构样本，异常样本会产生较大的重构误差。因此，所提方法将待测样本的重构误差作为异常得分对其进行分类。

RAD-ASAE 的测试过程如图 4 所示，给定待测样本 $\boldsymbol{Y}=\{\boldsymbol{y}_i\}_{i=1}^{M}$，经过编码器和解码器后得到重构样本 $\hat{\boldsymbol{Y}}=\{\hat{\boldsymbol{y}}_i\}_{i=1}^{M}$，计算待测样本与重构样本的重构误差，并使用该重构误差作为异常得分 $S(\boldsymbol{y}_i)$ 对待测样本进行分类，异常得分的计算式可表示为：

$$S(\boldsymbol{y}_i)=\frac{1}{M}\sum_{i=1}^{M}\| \boldsymbol{y}_i-\hat{\boldsymbol{y}}_i \|_2^2 \tag{18}$$

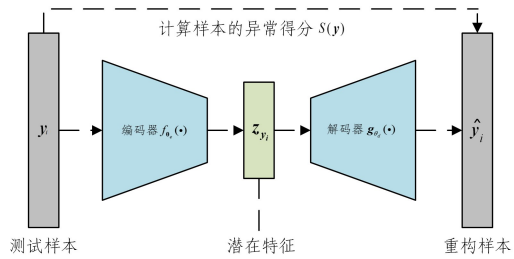


图 4 RAD-ASAE 的测试过程

Fig. 4 Testing process of RAD-ASAE

3.3 算法实现

对抗样本的生成过程、RAD-ASAE 的对抗训练过程和 RAD-ASAE 的测试过程分别如算法 1、算法 2 和算法 3 所示。

算法 1 对抗样本的生成过程

输入：正常样本 $\mathbf{X}=\{\mathbf{x}_i\}_{i=1}^{N}$

输出：对抗样本 $\mathbf{X}^*=\{\mathbf{x}_i^*\}_{i=1}^{N}$

初始化：编码器 $f_{\boldsymbol{\theta}_e}(\cdot)$ 和解码器 $g_{\boldsymbol{\theta}_d}(\cdot)$ 的参数 $\boldsymbol{\theta}_e$，$\boldsymbol{\theta}_d$

1. for i←1 to N do
2. 通过式(9)和式(10)生成对抗样本 $\mathbf{X}^*=\{\mathbf{x}_i^*\}_{i=1}^{N}$
3. 使用 PGD 更新，对式(9)进行优化
4. end for

算法 2 RAD-ASAE 的对抗训练过程

输入：正常样本 $\mathbf{X}=\{\mathbf{x}_i\}_{i=1}^{N}$ 和对抗样本 $\mathbf{X}^*=\{\mathbf{x}_i^*\}_{i=1}^{N}$

输出：编码器 $f_{\boldsymbol{\theta}_e}(\cdot)$ 和解码器 $g_{\boldsymbol{\theta}_d}(\cdot)$

初始化：编码器 $f_{\boldsymbol{\theta}_e}(\cdot)$ 和解码器 $g_{\boldsymbol{\theta}_d}(\cdot)$ 的参数 $\boldsymbol{\theta}_e$，$\boldsymbol{\theta}_d$

1. for i←1 to N do
2. 通过式(11)和式(12)计算 $\mathbf{x}_i$ 和 $\mathbf{x}_i^*$ 经过各自编码器后的潜在特征 $\mathbf{z}_i$ 和 $\mathbf{z}_i^*$
3. 通过式(13)计算 $\mathbf{z}_i^*$ 经过解码器后的重构向量 $\hat{\mathbf{x}}_i^*$
4. 通过式(14)－式(16)计算 L_c，L_s 和 L_z
5. 通过式(17)得到 $L_{\text{RAD-ASAE}}$
6. 使用 PGD 更新 $f_{\boldsymbol{\theta}_e}(\cdot)$ 和 $g_{\boldsymbol{\theta}_d}(\cdot)$ 的参数 $\boldsymbol{\theta}_e$，$\boldsymbol{\theta}_d$
7. end for

算法 3 RAD-ASAE 的测试过程

输入：待测样本 $\mathbf{Y}=\{\mathbf{y}_i\}_{i=1}^{M}$，编码器 $f_{\boldsymbol{\theta}_e}(\cdot)$ 和解码器 $g_{\boldsymbol{\theta}_d}(\cdot)$，阈值 δ

输出：Y 的异常得分 $S(\mathbf{y}_i)$ 及其类别

初始化：编码器 $f_{\boldsymbol{\theta}_e}(\cdot)$ 和解码器 $g_{\boldsymbol{\theta}_d}(\cdot)$ 的参数 $\boldsymbol{\theta}_e$，$\boldsymbol{\theta}_d$

1. for i←1 to M do X
2. 通过式(12)和式(13)得到待测样本 $\mathbf{Y}=\{\mathbf{y}_i\}_{i=1}^{M}$ 的重构样本 $\hat{\mathbf{Y}}=\{\hat{\mathbf{y}}_i\}_{i=1}^{M}$
3. 通过式(18)计算待测样本的异常得分 $S(\mathbf{y}_i)$
4. end for
5. if $S(\mathbf{y}_i)\geqslant\delta$：
6. then
7. $\mathbf{y}_i$ 是异常样本；
8. else
9. $\mathbf{y}_i$ 是正常样本；
10. end if

4 实验结果

为了验证所提 RAD-ASAE 的性能，在 3 个公共数据集上将其与 7 种相关方法进行了实验比较。首先，详细介绍了所

选取的公共数据集和相关方法；其次，对实验安排和评价指标进行了描述；再次，对损失函数参数的选取进行了分析；然后，将 RAD-ASAE 与相关方法进行了比较；最后，通过消融研究和鲁棒性分析进一步验证了 RAD-ASAE 的有效性。

4.1 数据集及相关方法

在以下实验中选用 MNIST[42]，Fashion-MNIST[43]和 CIFAR-10[44]3 组数据集，每个数据集的示例样本如图 5 所示。这些数据集的大小、图像形状等都不相同。MNIST 数据集包含 70000 个 28×28 灰度手写数字，范围为 0～9。Fashion-MNIST 数据集类似于 MNIST 数据集，包含了 10 个时尚产品类别的 70000 张 28×28 灰度图像。CIFAR-10 数据集包含 10 个类别的 60000 张 32×32 彩色图像。Fashion-MNIST 数据集和 CIFAR-10 数据集的标注编号如表 1 所列。



图 5 从 3 个数据集中选取的示例样本

Fig. 5 Samples selected from three datasets

表 1 Fashion-MNIST 和 CIFAR-10 中的各类别标注的编号

Table 1 Label the categories in Fashion-MNIST and CIFAR-10

数据集	编号									
	0	1	2	3	4	5	6	7	8	9
Fashion-MNIST	T-shirt/top	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
CIFAR-10	Airplane	Car	Bird	Cat	Deer	Dog	Fog	Horse	Ship	Truck

实验过程中选取了 7 种相关方法进行对比实验，分别为 iForest[45]，DSVDD[46]，fAnoGAN[47]，GANomaly[48]，MemAE，DAGMM[49]，ARAE。

iForest 是一种快速异常检测方法，具有线性时间复杂度和高精准度，但其不适用特别高维的数据且离群点敏感。参数设置同文献[45]。

DSVDD 是一种将深度神经网络和传统 SVDD 相结合的异常检测方法，利用深度神经网络提取样本特征，同时将输出空间优化为包含样本的超球，以此来判定异常，但由于使用待测样本与超球中心的距离作为异常检测的判定准测，因此对特征提取的质量有较高的要求。参数设置同文献[46]。

fAnoGAN 是一种利用编码器将待测样本快速映射到生成对抗网络的潜在空间并使用 WGAN 做异常检测的方法，但其只能检测出部分异常，存在较高的漏报率。参数设置同文献[47]。

GANomaly 使用编码器-解码器的结构作为生成器，极大地提高了异常检测的速度，并取得了很好的性能。参数设置同文献[48]。

MemAE 是一种基于自编码器结构且具有存储功能的异常检测的方法，在训练阶段将有代表性的潜在特征存储到记忆模块中，在测试阶段将待测样本的潜在特征在记忆模块中寻找最接近的特征组合。与传统基于自编码器做异常检测的方法相比，弥补了由于自编码器优秀的泛化性能而无法检测异常的缺点。参数设置同文献[17]。

DAGMM 巧妙地将 DAE 的降维和密度估计过程结合到一起，通过多轮训练达到最大似然函数的效果。参数设置同文献[49]。

ARAE 首先生成对抗样本，然后利用对抗样本和正常样本一起训练自编码器，以提高模型的鲁棒性，但该方法忽略了自编码器对输入样本的重构约束，导致效果欠佳。参数设置同文献[22]。

4.2 实验安排及评价指标

RAD-ASAE 模型使用 3 层全连接网络，网络层大小为(512,256,128)，使用 Sigmoid 激活函数，权重和偏差项的初始化方式与文献[50]相同。此外，使用 Adam 优化器[51]进行优化。RAD-ASAE 在 MNIST 和 CIFAR-10 数据集上的最大训练迭代次数为 1600，扰动设置为 0.2，在 Fashion-MNIST 数据集上的最大训练迭代次数为 3000，扰动设置为 0.05。实验过程如下。

首先，针对损失函数 $L_{\text{RAD-ASAE}}$ 中的权重参数 α，β 和 γ 的不同取值，对所提 RAD-ASAE 检测性能的影响进行实验分析。由式(17)可知，损失函数 $L_{\text{RAD-ASAE}}$ 由 L_c，L_s 和 L_z 构成，其中 L_c 与 L_s 分别是样本空间中的重构损失和相似损失，L_z 是潜在空间中的特征损失。由于 L_c 和 L_s 都是样本空间中的损失函数。因此 $\alpha=\gamma$，此外，令 $\alpha+\beta+\gamma=1$，且 β 在{0.2,0.4,0.6,0.8}中选取。

其次，为了验证 RAD-ASAE 的异常检测性能，在 MNIST，Fashion-MNIST 和 CIFAR-10 数据集上，将 RAD-ASAE 与 iForest，DSVDD，fAnoGAN，GANomaly，MemAE，DAGMM 以及 ARAE 这 7 种相关方法进行实验比较。相关方法选取的参数设置同相关参考文献。

然后，为了分析样本空间和潜在空间中的约束对 RAD-ASAE 检测性能的影响，对损失函数 $L_{\text{RAD-ASAE}}$ 的选取方式进行消融实验分析。

最后，为了验证 RAD-ASAE 的对抗鲁棒性，将攻击参数 ε 设置为{0,0.05,0.1,0.2}进行实验分析。

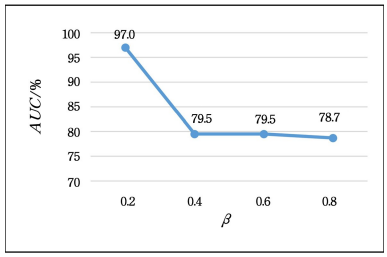
由于实验量较大，在进行参数变化实验、消融实验与对抗鲁棒性实验时，选取 MNIST，CIFAR-10 和 Fashion-MNIST 数据集中类别为“2”和“8”的样本作为训练样本进行分析。

为了衡量各种异常检测模型的性能，本文使用 ROC 曲线下的面积(AUC)作为性能度量指标。ROC 的横坐标为假正率(FPR)，表示预测为正常样本但实际为异常样本占所有异常样本的比例。ROC 的纵坐标为真正率(TPR)，表示预测为正常样本且实际为正常样本占所有正常样本的比例。AUC%值是判断预测模型优劣的标准，其特性是无论数据集正常样本和异常样本是否存在不平衡，随机猜测的基线始终

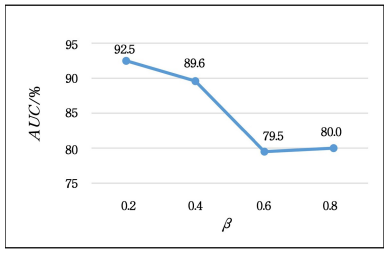
是 50[52]。AUC%值越接近 1，说明其相应方法的检测性能越好。

4.3 参数分析

在不同数据集中，为了分析参数变化对异常检测性能的影响，对损失函数 $L_{\text{RAD-ASAE}}$ 中的权重参数进行实验，实验结果如图 6－图 8 所示。



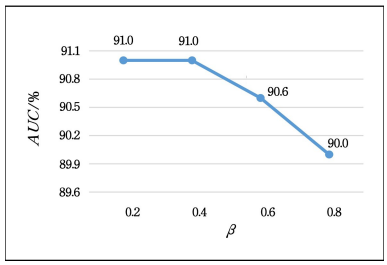
(a)正常类为 2



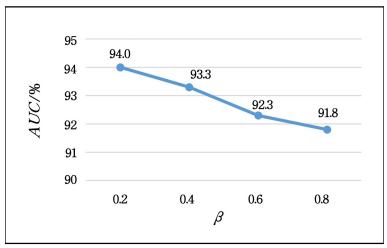
(b)正常类为 8

图 6 当 β 取不同值时，RAD-ASAE 在 MNIST 数据集类别为 2 和类别为 8 时取得的 AUC 值

Fig. 6 AUC values of RAD-ASAE in MNIST dataset class 2 and class 8 with different β



(a)正常类为 2

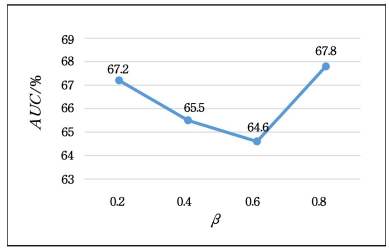


(b)正常类为 8

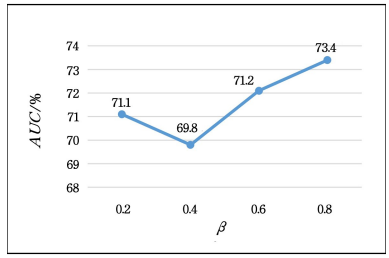
图 7 当 β 取不同值时，RAD-ASAE 在 Fashion-MNIST 数据集类别为 2 和类别为 8 时取得的 AUC 值

Fig. 7 AUC values of RAD-ASAE in Fashion-MNIST dataset class 2 and class 8 with different β

图 6(a)和图 6(b)分别给出了 RAD-ASAE 在 MNIST 数据集上选取正常类为“2”和“8”时不同的参数值 β 得到的 AUC%值。当 $\beta=0.2$ 时，RAD-ASAE 模型取得最大 AUC%值，分别为 97 和 92.5，于是选取 $\alpha=\gamma=0.4$，$\beta=0.2$ 作为 RAD-ASAE 在 MNIST 数据集上损失函数 $L_{\text{RAD-ASAE}}$ 最佳的权重参数设置。图 7(a)中，当 $\beta=0.2$ 与 $\beta=0.4$ 时，RAD-ASAE 模型在 Fashion-MNIST 数据集中选取正常类为“2”时取得最大 AUC%值，为 91。图 7(b)中，当 $\beta=0.2$ 时，RAD-ASAE 模型在 Fashion-MNIST 数据集中选取正常类为“8”时取得最大 AUC%值，为 94，结合图 7(a)和图 7(b)，RAD-ASAE 模型在 Fashion-MNIST 数据集中选取的最佳权重参数设置为 $\alpha=\gamma=0.3$，$\beta=0.4$。图 8(a)和图 8(b)分别给出了 RAD-ASAE 在 CIFAR10 数据集上选取正常类为“2”和“8”时不同的 β 参数得到的 AUC%值，由图可知，当 $\beta=0.8$ 时，取得最大 AUC%值，分别为 67.8 和 73.4。因此，在 CIFAR10 数据集中，RAD-ASAE 选取的最佳权重参数设置为 $\alpha=\gamma=0.1$，$\beta=0.8$。



(a)正常类为 2



(b)正常类为 8

图 8 当 β 取不同值时，RAD-ASAE 在 CIFAR-10 数据集类别为 2 和类别为 8 时取得的 AUC 值

Fig. 8 AUC values of RAD-ASAE in CIFAR-10 dataset class 2 and class 8 with different β

综上，由图 6－图 8 的实验结果可知，针对不同的数据集，β 需要选取合适的参数值才能使 RAD-ASAE 取得最优的异常检测性能。

4.4 相关方法比较分析

为了分析 RAD-ASAE 的异常检测性能，在 MNIST，Fashion-MNIST 和 CIFAR-10 数据集上，将它与 iForest，DSVDD，fAnoGAN，GANomaly，MemAE，DAGMM 以及 ARAE 这 7 种相关方法进行实验比较。表 2－表 4 分别列出了 RAD-ASAE 和相关方法在 MNIST，Fashion-MNIST 和 CIFAR-10 图像数据集构建的 10 个异常检测数据集上的对比结果。

表 2 中，除类别“8”外，RAD-ASAE 在 MNIST 数据集的其余 9 个类别上均取得了优于其他 7 种相关方法的 AUC%值，且平均 AUC%值最高，为 97.8。表 3 中，RAD-ASAE 方法在正常类为“1”“2”“4”“6”“7”和“8”的数据集上均取得了优于其他 7 种方法的 AUC%值，且平均 AUC%值最高，为 93.4。表 4 中，虽然 RAD-ASAE 只有在正常类为“2”“3”和“6”时取得最大的 AUC%值，但获得了最大的平均 AUC%

值，为 63.5。此外，ARAE 方法在 MNIST 数据集中正常类为“0”和“1”以及 Fashion-MNIST 数据集中正常类为“0”“1”“7”，展现出来较高的 AUC%值，但 ARAE 并未有效提高自编码器的重构能力，导致平均 AUC 值较低，异常检测性能较差。fAnoGAN 在 Fashion-MNIST 数据集中正常类为“5”，取得了较高的 AUC%值，但 fAnoGAN 方法只能检测出部分异常，因此整体效果并不是很好。GANomaly 在 MNIST 数据集中正常类为“1”，在 Fashion-MNIST 数据集中正常类为“3”和“5”，取得了较高的 AUC 值，但 GANomaly 对训练数据质量要求较高。在 CIFAR10 数据集中，MemAE 和 RAD-ASAE 具有一定的竞争性，是因为 MemAE 可以抑制泛化能力，使重构样本更加接近于正常样本。

表 2 不同方法在 MNIST 数据集上取得的 AUC 值

Table 2 AUC values of different methods on MNIST dataset

数据集	方法							
	iForest	DSVDD	fAnoGAN	GANomaly	MemAE	DAGMM	ARAE	RAD-ASAE
MNIST(0)	88.1	98.4	97.8	99.0	97.0	78.3	**99.8**	**99.8**
MNIST(1)	96.1	99.6	99.7	**99.9**	99.0	93.4	**99.9**	**99.9**
MNIST(2)	66.3	86.1	83.8	90.1	81.3	70.1	95.1	**97.0**
MNIST(3)	73.3	90.1	83.2	94.8	84.6	62.2	97.0	**97.4**
MNIST(4)	76.2	95.9	89.5	97.1	89.0	77.1	97.3	**97.8**
MNIST(5)	66.4	84.4	92.3	97.4	81.0	67.5	96.9	**97.5**
MNIST(6)	77.3	97.5	96.4	99.2	91.3	65.4	99.5	**99.6**
MNIST(7)	80.2	95.3	96.4	96.8	91.7	72.3	97.3	**97.8**
MNIST(8)	66.9	93.1	82.0	87.6	80.0	68.8	92.1	92.5
MNIST(9)	77.2	95.2	93.5	98.1	90.3	80.4	98.2	**98.4**
AVG	76.9	93.6	91.5	96.0	88.5	73.6	97.3	**97.8**

表 3 不同方法在 Fashion-MNIST 数据集上取得的 AUC 值

Table 3 AUC values of different methods on Fashion-MNIST dataset

数据集	方法							
	iForest	DSVDD	fAnoGAN	GANomaly	MemAE	DAGMM	ARAE	RAD-ASAE
Fashion-MNIST(0)	80.5	87.5	86.6	91.1	82.0	74.5	**93.0**	92.8
Fashion-MNIST(1)	93.4	98.3	98.5	98.2	91.0	94.3	**99.2**	**99.2**
Fashion-MNIST(2)	81.2	84.5	90.4	82.6	86.6	82.2	90.8	**91.0**
Fashion-MNIST(3)	86.6	92.5	93.2	**94.9**	84.3	82.4	94.1	93.9
Fashion-MNIST(4)	82.8	91.2	90.3	90.0	88.0	76.4	91.5	**91.6**
Fashion-MNIST(5)	86.8	87.8	**93.0**	**93.0**	88.8	87.4	91.4	91.5
Fashion-MNIST(6)	71.0	77.2	82.2	80.6	81.4	59.3	83.0	**83.1**
Fashion-MNIST(7)	93.7	98.4	98.4	98.2	92.0	95.3	**98.9**	**98.9**
Fashion-MNIST(8)	80.1	92.1	90.3	88.2	82.0	77.2	93.5	**94.0**
Fashion-MNIST(9)	92.3	**98.6**	97.6	97.1	89.7	92.6	98.0	98.1
AVG	84.8	90.8	92.0	91.4	86.6	82.2	93.3	**93.4**

表 4 不同方法在 CIFAR-10 数据集上取得的 AUC 值

Table 4 AUC values of different methods on CIFAR-10 dataset

数据集	方法							
	iForest	DSVDD	fAnoGAN	GANomaly	MemAE	DAGMM	ARAE	RAD-ASAE
CIFAR-10(0)	55.2	64.1	70.2	75.6	**79.3**	57.8	71.2	73.1
CIFAR-10(1)	49.1	**60.3**	47.8	58.4	45.2	49.3	42.5	46.9
CIFAR-10(2)	53.9	47.8	66.3	56.2	65.4	55.6	67.2	**67.8**
CIFAR-10(3)	50.3	57.8	58.2	53.9	53.3	51.4	55.0	**59.3**
CIFAR-10(4)	60.3	52.3	74.8	56.9	**77.4**	54.3	74.5	77.2
CIFAR-10(5)	53.1	**64.9**	59.1	56.1	57.2	51.7	48.1	58.7
CIFAR-10(6)	58.6	58.2	72.7	67.2	71.8	54.6	69.3	**72.9**
CIFAR-10(7)	52.5	**60.9**	51.1	53.4	53.2	55.4	51.2	59.2
CIFAR-10(8)	57.3	77.1	73.1	73.1	**79.4**	52.1	71.5	73.4
CIFAR-10(9)	53.5	68.2	47.8	59.5	52.9	54.2	39.2	46.6
AVG	54.4	61.2	62.1	61.0	63.5	53.6	59.0	**63.5**

整体来看，RAD-ASAE 在 MNIST，Fashion-MNIST 和 CIFAR-10 数据集上取得了最大的 AUC%值，该方法通过对双空间进行联合训练，来充分学习正常样本在样本空间中的样本信息以及潜在空间中的潜在特征信息，提高了自编码器的重构能力。在测试阶段，RAD-ASAE 使异常样本的重构更加接近正常样本，与 ARAE 相比，RAD-ASAE 模型能够充分利用自编码器的重构特性，通过对样本空间和潜在空间进行联合训练以及减小输入样本和重构样本之间的重构误差，来提高自编码器的重构能力。

4.5 消融研究

为了分析样本空间和潜在空间中的约束对 RAD-ASAE 检测性能的影响，在 MNIST，Fashion-MNIST 和 CIFAR-10 数据集上对 RAD-ASAE 进行了消融研究，结果如表 5 和表 6 所列。由于实验量较大，在以下实验中，选取 MNIST，Fash-

ion-MNIST 和 CIFAR-10 数据集中正常类为“2”和“8”进行消融研究，研究情况有 3 种需要分析。第 1 种情况只考虑样本空间的约束，即保留 L_c 和 L_s，记为 RAD-ASAE-W/O L_z，在 MNIST，Fashion-MNIST 和 CIFAR-10 数据集中选取正常类为“2”的 AUC 值分别为 95.8，90.7 和 64.4，平均 AUC 值为 83.6，选取正常类为“8”得到的 AUC 值分别为 91.6，93.9 和 67.3，平均 AUC 值为 82.3。第 2 种情况仅使用潜在空间中的特征损失，即对样本空间没有任何约束，记为 RAD-ASAE-W L_z，在 MNIST，Fashion-MNIST 和 CIFAR-10 数据集中选取正常类为“2”时得到的 AUC 值分别为 57.4，76.5 和 61.1，平均 AUC 值为 65，选取正常类为“8”时得到的 AUC 值分别为 76，67.6 和 64.3，平均 AUC 值为 69.3。第 3 种情况为 RAD-ASAE 模型，为了获得较高的重构能力，对样本空间和潜在空间同时施加约束。在 MNIST，Fashion-MNIST 和 CIFAR-10 数据集中选取正常类为“2”时得到的 AUC 值分别为 97，91 和 67.8，平均 AUC 值为 85.3，选取正常类为“8”时得到的 AUC 值为 92.5，94 和 73.4，平均 AUC 值为 86.6。实验结果显示，第 3 种情况的 AUC 值和平均 AUC 值分别取得最大值，这表明同时对样本空间和潜在空间施加约束能够使 RAD-ASAE 展现最优的异常检测性能。

表 5 在 MNIST，Fashion-MNIST 和 CIFAR-10 正常类为“2”的数据集上 RAD-ASAE 使用不同网络结构得到的 AUC 值

Table 5 AUC values obtained by RAD-ASAE using different network structures on datasets MNIST，Fashion-MNIST and CIFAR-10 normal class “2”

数据集	网络结构		
	RAD-ASAE-W/OL_z	RAD-ASAE-WL_z	RAD-ASAE
MNIST(2)	95.8	57.4	97.0
Fahion-MNIST(2)	90.7	76.5	91.0
CIFAR-10(2)	64.4	61.1	67.8
AVG	83.6	65.0	85.3

表 6 在 MNIST，Fashion-MNIST 和 CIFAR-10 正常类为“8”的数据集上 RAD-ASAE 使用不同网络结构得到的 AUC 值

Table 6 AUC values obtained by RAD-ASAE using different network structures on datasets MNIST，Fashion-MNIST and CIFAR-10 normal class “8”

数据集	网络结构		
	RAD-ASAE-W/OL_z	RAD-ASAE-WL_z	RAD-ASAE
MNIST(8)	91.6	76.0	92.5
Fahion-MNIST(8)	93.9	67.6	94.0
CIFAR-10(8)	67.3	64.3	73.4
AVG	82.3	69.3	86.6

由此，RAD-ASAE 模型通过对样本空间中的 L_c 和 L_s 进行约束，不仅可以充分学习输入样本的信息，还可以确保正常样本与对抗样本在视觉上相似。此外，潜在空间中的 L_z 使正常样本与对抗样本的潜在特征更加相似，充分学习输入样本丰富的特征信息。通过对样本空间和潜在空间进行双空间联合训练，来充分学习样本的多层次表示，以获得优异的异常检测效果。

4.6 鲁棒性分析

为了验证 RAD-ASAE 的对抗鲁棒性，选取 MNIST，Fashion-MNIST 和 CIFAR-10 数据集中正常类为“2”和“8”的样本进行鲁棒性分析。

将式(9)中的攻击参数 ε 设置为{0，0.05，0.1，0.2}，当 $\varepsilon=0$ 时，未对训练样本添加扰动，即训练过程只有正常样本，没有生成对抗样本，当 $\varepsilon=0.05$，0.1，0.2 时，即生成对抗样本，且训练数据集由正常样本和对抗样本组成。如表 7 和表 8 所列，在 MNIST 数据集中，选取正常类为“2”时，设置 $\varepsilon=0.2$取得最大 AUC 值，为 97；同样地，选取正常类为“8”时，设置 $\varepsilon=0.2$ 取得最大 AUC 值，为 92.5；在 Fashion-MNIST 和 CIFAR10 数据集中，选取正常类为“2”时，设置 $\varepsilon=0.05$ 取得最大 AUC 值，分别为 91 和 67.8；同样地，选取正常类为“8”时，设置 $\varepsilon=0.05$ 取得最大 AUC 值，分别为 94 和 73.4。因此，在本实验中 MNIST 数据集选取 $\varepsilon=0.2$，Fashion-MNIST 和 CIFAR10 选取 $\varepsilon=0.05$。

表 7 在 MNIST，Fashion-MNIST 和 CIFAR-10 正常类为“2”的数据集上 RAD-ASAE 使用不同攻击参数 ε 得到的 AUC 值

Table 7 AUC values obtained by RAD-ASAE using different attack parameters ε on datasets MNIST，Fashion-MNIST and CIFAR-10 normal class “2”

数据集	攻击参数 ε			
	0	0.05	0.1	0.2
MNIST(2)	95.9	96.1	96.6	97.0
Fahion-MNIST(2)	88.9	91.0	90.1	89.6
CIFAR-10(2)	59.2	67.8	63.2	63.5

表 8 在 MNIST，Fashion-MNIST 和 CIFAR-10 正常类为“2”的数据集上 RAD-ASAE 使用不同攻击参数 ε 得到的 AUC 值

Table 8 AUC values obtained by RAD-ASAE using different attack parameters ε on datasets MNIST，Fashion-MNIST and CIFAR-10 normal class “2”

数据集	攻击参数 ε			
	0	0.05	0.1	0.2
MNIST(8)	95.9	96.1	96.6	97.0
Fahion-MNIST(8)	88.9	91.0	90.1	89.6
CIFAR-10(8)	59.2	67.8	63.2	63.5

此外，由表 7 和表 8 可知，添加扰动比不添加扰动产生的 AUC 值更大，可以得出，当训练样本集中既有正常样本也有对抗样本时，RAD-ASAE 模型仍获得优异的异常检测结果，因此 RAD-ASAE 模型具有较强的对抗鲁棒性。

结束语 为了提高异常检测的对抗鲁棒性，提出了一种基于对抗样本和自编码器的鲁棒异常检测方法(RAD-ASAE)。首先，对给定的正常样本施加扰动后生成对抗样本，采用由对抗样本和正常样本组成的数据集对模型进行训练，以提高模型的对抗鲁棒性；其次，对样本空间和潜在空间同时施加约束，采用双空间联合训练来提高自编码器的重构能力。在 MNIST，Fashion-MNIST，CIFAR-10 数据集上的实验结果表明，与 7 种相关方法相比，RAD-ASAE 展现了更优的检测性能。

尽管 RAD-ASAE 模型在 MNIST，Fashion-MNIST，CIFAR-10 数据集上取得优异的检测性能，但该模型的时间复杂度较高，并且该模型并不一定适用于所有的异常检测数据集。在今后的研究工作中，需要建立更加有效并且快速的模型，用于解决上述问题。

参 考 文 献

[1] WELLER-FAHY D J, BORGHETTI B J, SODEMANN A A. A Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection [J]. IEEE Communications Surveys & Tutorials, 2014, 17(1): 70-91.

[2] GEBREMESKEL G B, YI C, HE Z, et al. Combined Data Mining Techniques Based Patient Data Outlier Detection for Healthcare Safety [J]. International Journal of Intelligent Computing and Cybernetics, 2016, 9(1): 42-68.

[3] RAMOTSOELA D, ABU-MAHFOUZ A, HANCKE G. A Survey of Anomaly Detection in Industrial Wireless Sensor Networks With Critical Water System Infrastructure as a Case Study [J]. Sensors, 2018, 18(8): 2491.

[4] PIMENTEL M A F, CLIFTON D A, Clifton L, et al. A Review of Novelty Detection [J]. Signal Processing, 2014, 99: 215-249.

[5] ILONEN J, PAALANEN P, KAMARAINEN J K, et al. Gaussian Mixture Pdf in One-Class Classification: Computing and Utilizing Confidence Values [C]// Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong: IEEE, 2006, 2: 577-580.

[6] ZHANG B W, LIU Z, SANG G M. Anomaly Detection Algorithm Based on Kernel Density Fluctuation [J]. Computer Engineering and Applications, 2021, 57(12): 132-136.

[7] DING P L. Anomaly Detection Based on K-Nearest Neighbor Isolation Forest [J]. Computer system application, 2023, 32(2): 199-206.

[8] FAN R, LI N. K-means Clustering Anomaly Detection Method Based on Feature Selection [J]. Network Security Technology and Application, 2018(4): 25-26.

[9] TURKOZ M, KIM S, SON Y, et al. Generalized Support Vector Data Description for Anomaly Detection [J]. Pattern Recognition, 2020, 100: 107-119.

[10] KITTIDACHANAN K, MINSAN W, PORNNOPPARATH D, et al. Anomaly Detection Based on GS-OCSVM Classification [C]// Proceedings of the International Conference on Knowledge and Smart Technology. Phuket: IEEE, 2020: 64-69.

[11] SINGH S, MARKOU M. An Approach to Novelty Detection Applied to the Classification of Image Regions [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(4): 396-407.

[12] LI W, WANG Y. A Robust Supervised Subspace Learning Approach for Output-Relevant Prediction and Detection Against Outliers [J]. Journal of Process Control, 2021, 106: 184-194.

[13] CALLEGARI C, GIORDANO S, PAGANO M. An Information-Theoretic Method for the Detection of Anomalies in Network Traffic [J]. Computers & Security, 2017, 70: 351-365.

[14] HINTON G E, SALAKHUTDINOV R R. Reducing the Dimensionality of Data With Neural Networks [J]. Science, 2006, 313(5786): 504-507.

[15] GONG D, TAN M, ZHANG Y, et al. Blind Image Deconvolution by Automatic Gradient Activation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 1827-1836.

[16] XIA Y, CAO X, WEN F, et al. Learning Discriminative Reconstructions for Unsupervised Outlier Removal [C]// Proceedings of the IEEE International Conference on Computer Vision. Boston: IEEE, 2015: 1511-1519.

[17] GONG D, LIU L, LE V, et al. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection [C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 1705-1714.

[18] AN J, CHO S. Variational Autoencoder Based Anomaly Detection Using Reconstruction Probability [J]. Special lecture on IE, 2015, 2(1): 1-18.

[19] ZHOU C, PAFFENROTH R C. Anomaly Detection With Robust Deep Autoencoders [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax: IEEE, 2017: 665-674.

[20] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing, Properties of Neural Networks [C]// Proceedings of the International Conference on Learning Representations. Banff: IEEE: 2014.

[21] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and Harnessing Adversarial Examples [C]// Proceedings of the International Conference on Learning Representations. San Diego: IEEE, 2015.

[22] SALEHI M, ARYA A, PAJOUM B, et al. Arae: Adversarially Robust Training of Autoencoders Improves Novelty Detection [J]. Neural Networks, 2021, 144: 726-736.

[23] ROBBINS H, MONRO S. A Stochastic Approximation Method [J]. The Annals of Mathematical Statistics, 1951, 22(3): 400-407.

[24] YAN J L, XU Y, ZHANG S C, et al. A Survey of Adversarial Sample Attack and Defense Research on Image Classification Models [J]. Computer Engineering and Application, 2022, 58(23): 24-41.

[25] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial Examples in the Physical World [M]// Artificial Intelligence Safety and Security. Chapman and Hall/CRC, 2018: 99-112.

[26] XIAO Y, PUN C M. Improving Adversarial Attacks on Deep Neural Networks Via Constricted Gradient-Based Perturbations [J]. Information Sciences, 2021, 571: 104-132.

[27] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks [C]// Proceedings of the International Conference on Learning Representations. Vancouver: IEEE, 2018.

[28] CARLINI N, WAGNER D. Towards Evaluating the Robustness of Neural Networks [C]// 2017 IEEE Symposium on Security and Privacy. San Jose: IEEE, 2017: 39-57.

[29] LIN C, HAN S, ZHU J, et al. Sensitive Region-Aware Black-Box Adversarial Attacks [J]. Information Sciences, 2023, 637: 118929.

[30] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical Black-Box Attacks Against Machine Learning [C]// Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. Dallas: ACM, 2017: 506-519.

[31] ZHOU W，HOU X，CHEN Y，et al. Transferable Adversarial Perturbations [C]// Proceedings of the European Conference on Computer Vision. Munich：Springer，2018：452-467.

[32] ZHAO H，CHANG Y K，WANG W J. Overview of Adversarial Attacks And Defense Methods of Deep Neural Networks [J]. Computer Science，2022，49(S2)：662-672.

[33] LIAO F，LIANG M，DONG Y，et al. Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City，IEEE：2018：1778-1787.

[34] LIU S J，ZHANG Z X，ZHANG Q J. Adversarial Sample Defense Algorithm Based on Adversarial Training and Image Denoising [J]. Computer Engineering and Design，2022，43(11)：3093-3100.

[35] JIA X，WEI X，CAO X，et al. Comdefend：An Efficient Image Compression Model to Defend Adversarial Examples [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach：IEEE，2019：6084-6092.

[36] PRAKASH A，MORAN N，GARBER S，et al. Deflecting Adversarial Attacks With Pixel Deflection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：8571-8580.

[37] JIA X，ZHANG Y，WU B，et al. LAS-AT：Adversarial Training With Learnable Attack Strategy [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans：IEEE，2022：13398-13408.

[38] PAPERNOT N，MCDANIEL P，WU X，et al. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks [C]// Proceedings of the IEEE Symposium on Security and Privacy. San Jose：IEEE，2016：582-597.

[39] SAMANGOUEI P，KABKAB M，CHELLAPPA R. Defense-gan：Protecting Classifiers Against Adversarial Attacks Using Generative Models [C]// Proceedings of the International Conference on Learning Representations. Stockholm：ACM，2018.

[40] MENG D，CHEN H. Magnet：A Two-Pronged Defense Against Adversarial Examples [C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas：ACM，2017：135-147.

[41] CHEN R，CHEN J，ZHENG H，et al. Salient Feature Extractor for Adversarial Defense on Deep Neural Networks[J]. Information Sciences，2022，600：118-143.

[42] DENG L. The Mnist Database of Handwritten Digit Images for Machine Learning Research [best of the web[J]. IEEE Signal Processing Magazine，2012，29(6)：141-142.

[43] XIAO H，RASUL K，VOLLGRAF R. Fashion-Mnist：A Novel Image Dataset for Benchmarking Machine Learning Algorithms [J]. arXiv：1708. 07747，2017.

[44] THAKKAR V，TEWARY S，CHAKRABORTY C. Batch Normalization in Convolutional Neural Networks—A Comparative Study With CIFAR-10 Data [C]// Proceedings of the International Conference on Emerging Applications of Information Technology. Islamabad：IEEE，2018：1-5.

[45] LIU F T，TING K M，ZHOU Z H. Isolation Forest [C]// Proceedings of the EEE International Conference on Data Mining. Pisa：IEEE，2008：413-422.

[46] RUFF L，VANDERMEULEN R，GOERNITZ N，et al. Deep One-Class Classification [C]// Proceedings of the International Conference on Machine Learning. Stockholm：ACM，2018：4393-4402.

[47] 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.

[48] AKCAY S，ATAPOUR-ABARGHOUEI A，BRECKON T P. Ganomaly：Semi-Supervised Anomaly Detection Via Adversarial Training [C]// Proceedings of the Asian Conference on Computer Vision. Perth：Springer，2019：622-637.

[49] ZONG B，SONG Q，MIN M R，et al. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection [C]// Proceedings of the International Conference on Learning Representations. Vancouver：IEEE，2018.

[50] HE K，ZHANG X，REN S，et al. Delving Deep into Rectifiers：Surpassing Human-Level Performance on Imagenet Classification [C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago：IEEE，2015：1026-1034.

[51] DIEDERIK K AND JIMMY B. Adam：A Method for Stochastic Optimization [C]// Proceedings of the International Conference on Learning Representations(ICLR). San Diego：IEEE，2015.

[52] FAWCETT T. An Introduction to Roc Analysis [J]. Pattern Recognition Letters，2006，27(8)：861-874.



LI Shasha，born in 1996，postgraduate. Her main research interests include novelty detection，autoencoder and deep learning.



XING Hongjie，born in 1976，Ph.D，professor，Ph. D supervisor. His main research interests include kernel methods，neural networks，novelty detection，and ensemble learning.

（责任编辑：喻藜）