DOI: 10.19650/j.cnki.cjsi.J2312130

基于球面正则化的支持向量描述视觉异常检测*

邓诗卓[1,2],滕 达[1],李晓红[1],陈佳祺[1],陈东岳[1,2]
(1. 东北大学信息科学与工程学院 沈阳 110819; 2. 东北大学佛山研究生创新学院 佛山 528311)

摘 要:异常检测作为视觉领域中一项独特而关键的任务,在医疗、安保等领域具有广泛的前景。异常检测目前受限于大规模异常数据标注,因此现有方法集中在单类分类和弱监督学习,深度支持向量描述(Deep SVDD)是实现单类分类的常见方法。然而,传统 Deep SVDD 在开展异常检测时往往面临球体崩塌。针对这一问题,提出了基于球面正则化的 SVDD 异常检测算法,通过引入软间隔损失与支持向量的思想,优化模型学习流程。进一步地,面向可标注样本,提出了基于 SVDD 的弱监督异常检测方法。在公开数据集 MNIST 和 CIFAR-10 上进行消融和对比实验,实验证明,相比于有监督算法,在 MNIST 数据集上,SR-WSVDD 的性能提高了 3.7%,而在 CIFAR-10 数据集上则提高了 16.7%。此外,与其他弱监督算法相比,SR-WSVDD 在 CIFAR-10 数据集上提升了 1.8%。所提出的 SR-SVDD 异常检测算法,弥补 Deep SVDD 容易发生球体崩塌的缺陷,使模型异常检测结果更加准确。

关键词: 计算机视觉;单类分类;弱监督学习;异常检测;自编码器;支持向量
中图分类号: TP391.4 TH701 **文献标识码**: A **国家标准学科分类代码**: 510.99

Spherical regularized support vector description for visual anomaly detection

Deng Shizhuo[1,2], Teng Da[1], Li Xiaohong[1], Chen Jiaqi[1], Chen Dongyue[1,2]
(*1. College of Information Science and Engineering, Northeastern University, Shenyang 110819, China*;
2. Foshan Graduate School of Innovation, Northeastern University, Foshan 528311, China)

Abstract: Anomaly detection is an important task in the computer vision, such as medical, security. One of the challenges in anomaly detection is not easy to obtain large-scale annotated anomalous data. Existing methods focus on one-class classification and weakly supervised learning. Deep support vector data Description (Deep SVDD) is an important method to realize one-class anomaly detection. However, previous Deep SVDD often encounter the hypersphere collapse when constructing the model of the hypersphere. To solve this problem, support vector data description based on spherical regularization (SR-SVDD) is proposed in this paper. SR-SVDD applies the idea of support vectors to optimize the learning process by introducing slack terms. Furthermore, this paper proposes weakly supervised support vector data description based on spherical regularization (SR-WSVDD), which utilizes small amounts of labeled data. Ablation experiments and comparison experiments are carried out on MNIST and CIFAR-10. Experimental results show that, compared with supervised algorithms, the performance of SR-WSVDD is improved by 3.7% on the MNIST, and 16.7% on the CIFAR-10. In addition, compared with other weakly supervised algorithms, SR-WSVDD improves by 1.8% on CIFAR-10 dataset. The proposed SR-SVDD solves the spherical collapse of previous Deep SVDD, and makes the anomaly detection results more accurate.

Keywords: computer vision; one-class classification; weakly supervised learning; anomaly detection; autoencoder; support vector

0 引 言

异常检测作为机器学习领域有挑战性的研究课题,在理论和实践方面具有重要的研究意义。异常检测涵盖范围很广,不同的异常检测任务所涉及到的研究对象不同,例如,时间序列数据、文本数据以及多模态数据等,因此具有不同的挑战。对于时间序列数据来说,其异常数

收稿日期:2023-11-08 Received Date: 2023-11-08
*基金项目:国家自然科学基金(62202087)、广东省基础与应用基础研究基金(2024A1515010244,2021B1515120064)项目资助

据常具有长期依赖关系,这使得传统模型在捕捉这种模式上面临挑战,需要使用适应于长期依赖的模型,并且时间序列可能包含周期性模式与噪声,模型需要考虑这些因素以准确捕捉异常;文本数据通常是高维度且稀疏的,并且具有语义复杂性,相同的文本数据的语义可以有多种解释,使得定义异常变得复杂;多模态数据中不同模态之间可能存在复杂的关联关系,为了有效地进行异常检测,需要考虑跨模态之间的关系,确保模型能够综合利用多模态信息。本文将图像异常检测作为研究任务,考虑到图像数据的空间相关性,本文在深度学习模型中选择了以卷积操作为基础的自编码器,更适合进行图像领域的特征提取工作。不同于其他机器学习任务,异常检测存在以下特性[1]。1)异常检测任务中样本分布不平衡,正常样本数量占比大,标注的异常样本数量稀少甚至没有;2)异常事件通常未知且不可预测,并且异常事件具有不规则性,这导致各个异常事件间具有较大的差异性。基于以上特性,传统的有监督学习算法难以有效解决大多数场景下的异常检测任务。

近年来,由于视觉设备普及以及存储设备成本下降,图像与视频样本数量迅速提升,为图像异常检测的开展提供了数据支撑。基于视觉的异常检测任务通常指存在于图像中区别于正常场景的小概率事件。在监控图像中,正常行驶的车流是正常事件,车辆堵塞停滞被判定为异常事件。本文提出的算法作为基于图像领域的异常检测算法,可以解决很多现实生活中的异常检测问题。例如,在交通监测中用于检测道路上交通事故、逆行、堵车等不符合正常车流的异常情况;在医学领域中,本算法可以用于检测医学图像的异常结构或者病变;在农业领域可以用于检测作物的生长情况。

基于视觉的异常检测任务的核心问题是异常样本的稀缺性。由于异常样本通常难以观测和标注,因此,研究人员经常采用无监督学习和弱监督学习方法来应对异常检测的挑战。其中,单类分类是一种常用的无监督学习方法[2-9],在单类分类的训练过程中,不需要明确的异常样本,而是利用大多数正常样本的特征来构建超平面,以分隔正常和异常样本。近年来的理论及实验证明,传统单类分类算法在处理异常检测任务中图像与视频的高维数据时会出现内存占用大以及特征提取不充分等问题。相较而言,面向高维数据时,深度学习比传统算法更加有效[10],因此出现了很多基于深度学习模型的单类分类框架,通过将神经网络模型引入到原有的单类分类算法中以获得更好的性能,主流算法包括深度支持向量数据描述(deep support vector data description, Deep SVDD)[11],该算法通过训练神经网络拟合超球体从而为正常样本与异常样本划分边界,已经成为主要的深度单类分类算法。然而,在实际的超球体拟合过程中,模型存在着超球体崩塌现象,即构建得到的超球体不能很好地包围正常样本或者球体半径很小,甚至半径趋近于 0。这个现象导致原有的 Deep SVDD 不能成功地识别异常样本。

在深度单类分类算法中构建并引入合理的神经网络模型是实现基于视觉的异常检测任务的另一个核心要素。自编码器作为 Deep SVDD 引入的神经网络之一,具有学习恒等函数、提取数据降维的中间表示的功能。为了解决不同场景下的异常检测任务,研究人员开发了很多的基于自编码器的异常检测算法,包括去噪自编码器[12]、稀疏自动编码器[13],变分自编码器[14]等等。基于其优秀的特性与功能,自编码器已经是用于深度异常检测算法的主流神经网络模型[11,15]。

此外,在实际应用中,部分异常样本是能够获取与标注的,因此运用少量有标注的异常样本数据的弱监督异常检测算法是提升检测效果的解决方案之一。在弱监督环境下,数据集通常由大量的正常样本和少量的有标注异常样本组成。为了发挥标注异常样本的作用,研究人员设计了半监督[16]、不完全监督[17]、不确切监督算法[18-20]等学习模式。由于异常事件具有未知性以及不规则性,因此在训练过程中引入异常样本的信息是提高异常检测算法性能的重要途径。

基于上述分析,本文针对无异常样本和少量异常样本参与训练的两类异常检测任务提出两种异常检测算法,分别是用于无异常样本场景下的基于球面正则化的支持向量描述(support vector data description based on spherical regularization, SR-SVDD),与面向少量异常样本场景的基于支持向量数据描述的弱监督异常检测方法(weak supervision of support vector data description based on spherical regularization, SR-WSVDD)。本文总结了原有 Deep SVDD 算法存在的球体崩塌现象的原因;在提出的算法中,将支持向量机(SVM)思想与序列最小优化相结合,引入软间隔项与支持向量,进行损失重构,避免发生模型训练过程中球体崩塌现象;基于球体崩塌现象的诱因,对模型结构进行改进,实现更适用于超球体构建的自编码器模型。实验结果表明,本文提出的方法具有合理性,并成功解决原有 Deep SVDD 算法存在球体崩塌现象的弊端,并且在多个图像数据集上具有更好的鲁棒性与检测性能。

1 相关工作

鉴于视觉异常检测任务在实际应用中经常难以准确发现和标记异常样本,有监督的异常检测算法实施过程中常常面临重重困难。相比之下,单类异常检测作为一种无监督学习方法,仅依赖正常样本进行训练,在应对现实世界中异常数据稀缺问题上表现抢眼。经典传统的单

类异常检测方法包括单类支持向量机[3]、核密度估计[4]、隐马尔可夫模型[5],马尔可夫随机场[6]、高斯混合模型[7]和基于字典的重构模型[8-9]等。在构建异常检测模型时,尤其是在高维数据丰富的场景中,由于计算可扩展性不足以及维度灾难的困扰,这些方法往往表现不佳甚至在模型构建阶段面临失败的问题。

相对而言,深度学习为异常检测任务提供了一种可扩展性和泛化能力表现更出色的途径。这种方法与单类分类的结合能够应对视觉数据在复杂场景中的异常检测任务,并已经取得了更具竞争力的检测效果。其中,深度自编码器[15]为深度学习应用在异常检测的重要形式。Hawkins 等[12]提出了去噪自编码器,使用叠加噪声的原始数据进行重构学习,提高了特征提取的鲁棒性。Vincent 等[13]提出稀疏自动编码器,通过增加稀疏约束,使得神经网络在神经元较多的情况下依然能够提取特征。Makhzani 等[14]提出变分自编码器,以概率的方式描述对潜在空间的观察结果,更好地实现了数据生成。Masci 等[21]较早地在算法中融合卷积,提出卷积自编码器。Chong 等[22]提出在卷积自编码器中插入长短期记忆网络,构成时间自编码器。Xu 等[23]提出了重构视频帧与光流图的去噪自编码器。Zhao 等[24]提出时空自编码器,使用 3D 卷积提取时空特征并引入重构损失和未来帧预测进行异常检测任务。

此外,在部分应用场景中,研究人员可以获取与标注少量的异常样本。针对这类情况,仅使用单类分类则会忽略这些异常信息,因此需要引入弱监督异常检测算法来提取异常样本特征。面向不完全监督,Ruff 等[25]提出深度半监督异常检测算法,此方法结合信息论和熵分布知识对 Deep SVDD 进行改进,实现了端到端的不完全监督异常检测模型。Pang 等[16]提出了基于强化学习的不完全监督算法,该方法积极寻找超出标记训练数据范围的新异常类别并利用现有数据模型不断探索新的异常类型。另外,Pang 等[17]将算法制定为成对关系预测任务来解决无监督中的高误报率问题,利用有限数量的标记异常数据实现新的不完全监督模型。Pang 等[26]还提出了一个基于排名的算法,此算法使用少量标记数据作为先验知识学习更具表现力和应用相关的知识,并将表征学习与异常检测相互结合以学习低维表达,解决异常值不稳定的情况。Sultani 等[27]针对不确切监督任务提出了基于多示例学习的弱监督异常检测分类器,此模型使用三维卷积神经网络进行时空特征的提取并利用深度异常排序模型进行异常检测。此后的许多研究[18-20]都是基于多示例框架进行创新。Zhang 等[18]使用时间卷积网络代替原始多示例学习中的三维卷积神经网络,考虑每个包内示例之间的差距,并重新定义了新的损失函数约束不确切监督问题的函数空间。Tian 等[19]训练了一个特征级学习函数用来有效识别异常样本,提高多示例框架的鲁棒性。Feng 等[20]则将自训练与多示例学习进行结合提出了一种多实例自训练框架。

2　先验算法说明

2.1　Deep SVDD 算法原理分析

在以往的 Deep SVDD 算法中,训练数据 $x_1, \cdots, x_n \in \boldsymbol{X}$,其中 $n \in \boldsymbol{N}$ 表示样本数据的数量,$\boldsymbol{X}$ 表示数据集,输出空间 $\boldsymbol{F}$,$\phi(\cdot;\boldsymbol{W})$ 表示将输入空间 $\boldsymbol{X}$ 映射到输出空间 $\boldsymbol{F}$ 的自编码器,c 为超球体区域中心,R 为超球体半径。此自编码器具备 $L \in \boldsymbol{N}$ 个隐藏层和权重 $\boldsymbol{W} = \{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^L\}$。Deep SVDD 旨在通过目标函数学习权重 $\boldsymbol{W}$,从而最小化输出空间 F 的超球体。该超球体的内侧为正常区域,外侧为异常区域。根据以上说明,基于软间隔的深度支持向量数据描述算法的目标函数如下:

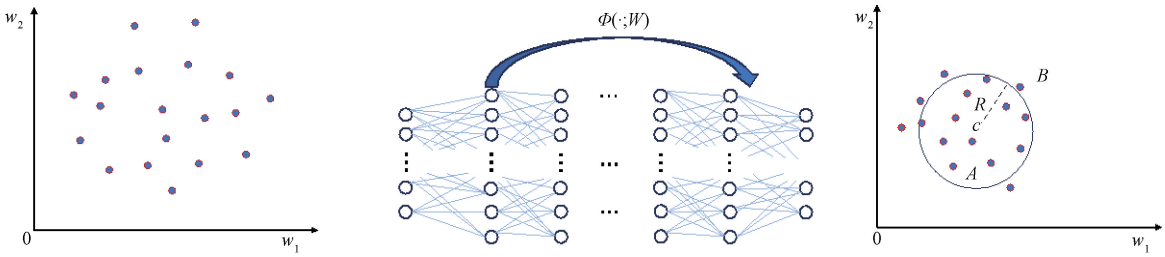
$$\min_{R,W} R^2 + \frac{1}{vn}\sum_{i=1}^{n}\max\{0, \|\phi(\boldsymbol{x}_i;\boldsymbol{W}) - \boldsymbol{c}\|^2 - R^2\} + \frac{\lambda}{2}\sum_{l=1}^{L}\|\boldsymbol{W}^l\|_F^2 \tag{1}$$

该目标函数的第 1 项旨在最小化球体半径 R^2,以达到精炼超球体体积、对正常数据进行准确描述的目标;第 2 项被视为软边界项,容许部分数据映射至球体外围,这种设置能够显著减缓算法的过拟合问题,超参数 $v \in (0,1]$ 被用于平衡超球体体积与松弛项之间的关系;第 3 项为网络增添正则化约束,缓解过拟合现象。

如图 1 所示,自编码器学习样本数据中的共同元素,进而使得数据中的正常数据被紧密地映射到中心点 c 周围。在进行测试时,正常数据就会被映射到超球体中心附近,而异常点被映射到超球体之外的区域。但 Deep SVDD 在训练过程会出现球体崩塌现象,深究该现象的科学原因,其根源在于 Deep SVDD 使用的损失函数(式(1))。虽然该损失函数达到了精炼超球体与将正常样本限制在超球体内的目标,但却与两类常见网络结构产生了冲突。当这两类网络结构出现时,Deep SVDD 构建的超球体可能出现球体崩塌现象,导致检测任务失败,下面分析这两类结构引起球体崩塌的原因。

2.2　Deep SVDD 问题描述

为了避免超球体崩塌现象,Deep SVDD 使用的网络结构必须遵循非零权重与无偏置项的条件。当网络中线性层权重 W_0 为 0 时,线性层操作 $z^l(\boldsymbol{x}) = \sigma^l(\boldsymbol{W}^l \cdot z^{l-1}(\boldsymbol{x}) + b^l)$ 退化为 $z^l(\boldsymbol{x}) = \sigma^l(b^l)$,则对于任意样本输入 $\boldsymbol{x}$,网络输出不受样本输入影响,均保持为经过激活函数操作后的偏置项,即 $\phi(\boldsymbol{x};\boldsymbol{W}_0) = \sigma^l(b^l) = \boldsymbol{c}_0$。基于原有损失函数进行优化,网络可以轻松学习错误心值 $\boldsymbol{c} = \boldsymbol{c}_0$。即网络



(a) 样本原始空间
(a) Original space

(b) 自编码器结构
(b) The autoencoder architecture

(c) 球体特征空间
(c) Sphere feature space

图 1 Deep SVDD 算法实验样本示意图

Fig. 1 Schematic diagram of the experimental sample of the Deep SVDD algorithm

学习到与样本无关的常值函数,且该函数值为球心值,对于每个样本,网络会将样本特征映射到球心。这意味着任何样本的特征点都集中在中心点 c,而不是聚集在中心点附近,从而导致球体崩塌现象的发生。因此,在进行实际的神经网络设置时,应当采用无偏置项的神经网络解决这一问题。

另一类限制为网络结构中不可使用单调有界的激活函数。单调有界的激活函数存在着一组上下限,其至少存在一个特征使一组激活函数的输出具有共同符号,那么非零上下限就可以在这组输出上被逐渐逼近,则具有有界激活函数的网络单元可能会饱和,从而在后续层中模拟偏置项,这又会导致球体坍塌。因此,在 Deep SVDD 中应首选无界激活函数或仅以 0 为界的函数,以避免因学习模拟偏置项而导致球体坍塌。

综合上述,为防止超球体崩塌现象,原有算法要求神经网络不仅不能含有偏置项,且仅可采用无界激活函数,以防止超球体崩溃问题的出现。然而,这两项限制对神经网络的泛化能力与模型扩展能力施加了一定的制约。因此,在后续的章节中,将致力于解决上述问题,推动其在实际应用中的表现。

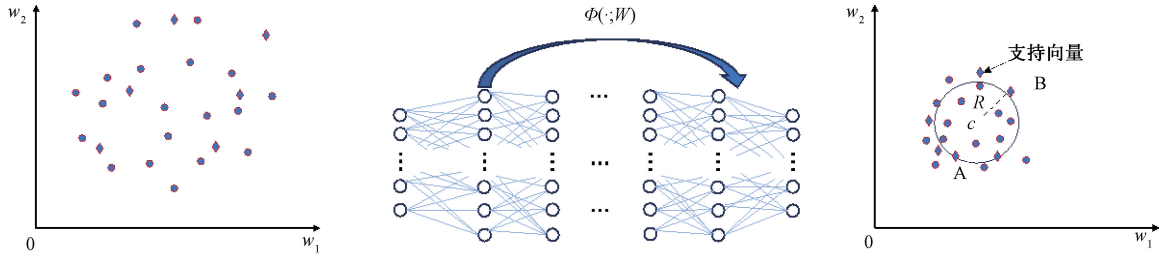
3 SR-SVDD 算法

本文基于 SVM 算法中支持向量定义,对 Deep SVDD 中支持向量的定义进行重新构想,从而设计出将数据收敛于超球面的算法——SR-SVDD。在学习策略方面,SR-SVDD 算法采用支持向量样本作为调控超球体体积的样本。该算法在支持样本的选择方式类似于求解 SVM 问题时所采用的序列最小优化算法。在每次训练迭代中,使用中位数函数挑选距离超球体最远的若干样本,这些样本作为支持向量参与训练。这种逐步迭代的学习方式有助于更全面地掌握样本的共性。例如,找出在 100 个样本中的共同特性较为困难,但在 10 个样本中找到共同特性则相对容易,进而基于 10 组共同特性更细致地描述正常样本的特征。

在目标函数方面,如图 2 所示,令输入空间 $X \subset \boldsymbol{R}^d$,输出空间 $\boldsymbol{F} \subset \boldsymbol{R}^p$,$\phi(\cdot;\boldsymbol{W})$ 表示从输入空间 $\boldsymbol{X}$ 映射到输出空间 $\boldsymbol{F}$ 的神经网络,神经网络有 $L(L \in \boldsymbol{N})$ 层隐藏层和一组权重 $\boldsymbol{W}(\boldsymbol{W} = \{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^L\})$,$R$ 表示超球体的半径,其中 $R > 0$,c 表示超球体的中心。SR-SVDD 通过改进的损失函数最小化数据输出空间 $\boldsymbol{F}$ 的超球体并保证神经网络在使用偏置项时不会造成超球体崩塌现象。算法经过训练将判定空间分为球外空间和球内空间,当进行异常性判断时,若特征点落在球内空间中,则判定数据为正常数据,若特征点落在球外空间中,则判定数据为异常数据。SR-SVDD 的损失函数 L_S 如下:

$$
\begin{aligned}
L_S &= L_1 + L_2 + L_3 + L_4 \\
L_1 &= \frac{1}{n_S}\sum_{i=1}^{n_S} \| \phi(\boldsymbol{x}_{is};\boldsymbol{W}) - \boldsymbol{c} \|^2 \\
L_2 &= C_1 \frac{1}{n_1}\sum_{i=1}^{n_1} \max(0, R^2 - \| \phi(\boldsymbol{x}_{i1};\boldsymbol{W}) - \boldsymbol{c} \|^2) \\
L_3 &= C_2 \frac{1}{n_2}\sum_{i=1}^{n_2} \max(0, \| \phi(\boldsymbol{x}_{i2};\boldsymbol{W}) - \boldsymbol{c} \|^2 - R^2) \\
L_4 &= \frac{\lambda}{2}\sum_{l=1}^{L} \| \boldsymbol{W}^l \|_F^2
\end{aligned}
\tag{2}
$$

式中:L_1 损失表示最小化支持向量到超球体中心的距离;其中 x_{is} 表示支持向量样本;n_s 表示选择的支持向量的数量值;L_2 软间隔项损失表示部分支持向量样本 x_{i1} 可以进入到超球体内;n_1 表示进入到超球体内的支持向量数量值;C_1 是超参数用于平衡与其他几项之间的关系;L_3 也是软间隔损失项表示超球体内部的部分非支持向量样本 x_{i2} 可以超出超球体范围;n_2 表示在超球体外侧的非支持向量数量值;C_2 用于平衡各损失项之间关系的超参数;L_4 正则化损失项用于缓解过拟合问题。SR-SVDD 算法引入了两个松弛项,以减小出现过拟合现象的可能性。通过支持向量与超球体内部向量的相互作用,当神经网络带有偏置项时,目标函数的最优解中的最优半径 R 不再趋近零,因此避免了超球体发生坍塌的现象。



(a) 样本原始空间
(a) Original space

(b) 自编码器结构
(b) The autoencoder architecture

(c) 球体特征空间
(c) Sphere feature space

图 2　SR-SVDD 算法实验样本分布示意图

Fig. 2　Distribution of experimental samples of SR-SVDD algorithm

基于信息传输极大原则理论对 SR-SVDD 进行解释,SR-SVDD 算法互信息主体是正常样本与球体特征,目标是最大化输入数据 $\boldsymbol{X}$ 与潜在表示 $\boldsymbol{Z}$ 之间的互信息。

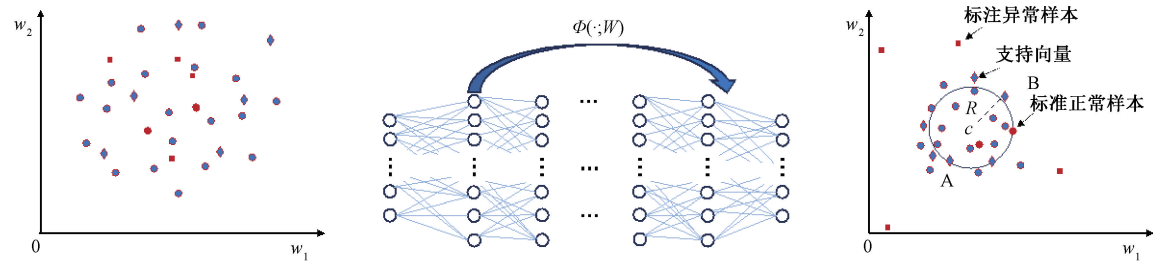
$$\max_{p(z\mid x)} I(\boldsymbol{X};\boldsymbol{Z}) + \beta R(\boldsymbol{Z}) \tag{3}$$

式中:$R(\boldsymbol{Z})$ 表示附加约束或者是正则化项例如稀疏性、潜在先验分布距离等;$\beta(\beta > 0)$ 用于平衡互信息与约束项之间的关系。根据信息传输极大原则对深度学习技术的要求[28],神经网络需要提取样本中的最具信息量的特征。遵循这项要求,SR-SVDD 算法目的是使超球体囊括的特征与正常样本的互信息最大化。为了实现互信息最大化,本文算法引入 SVM 的优化思想,在每次迭代过程中将最不符合正常特征的样本作为训练目标,去除不符合正常样本特征的冗余信息,使超球体特征排除异常样本信息,并包含正常样本共性信息。因此 SR-SVDD 算法具有基于信息传输极大原则的可解释性。

4　SR-WSVDD 算法

在特定条件下可以获取并标注少量异常样本,因此在弱监督学习场景中基于正常样本和少量异常样本构成的数据集,改进 SR-SVDD 单类算法,进而提出 SR-WSVDD 弱监督算法,以提升异常检测算法的准确性。

如图 3 所示,在弱监督场景中,数据集中有 n 个未标注样本 $x_1, \cdots, x_l \in X, X \subset \boldsymbol{R}^d$,$m$ 个标记样本 $(\tilde{\boldsymbol{x}}_1, \tilde{y}_1), \cdots, (\tilde{\boldsymbol{x}}_m, \tilde{y}_m) \in (\boldsymbol{X}, Y), \boldsymbol{X} \in \mathrm{R}^D, Y = \{-1, +1\}$,其中 $\tilde{y} = +1$ 表示正常样本标签,$\tilde{y} = -1$ 表示异常样本标签。SR-WSVDD 算法通过改进的损失函数指导神经网络最小化正常样本组成的超球体体积,最大化异常样本到中心点 c 的距离。其中 R 表示超球体的半径($R>0$)。



(a) 样本原始空间
(a) Original space

(b) 自编码器结构
(b) The autoencoder architecture

(c) 球体特征空间
(c) Sphere feature space

图 3　SR-WSVDD 实验样本示意图

Fig. 3　Schematic diagram of experimental sample of SR-WSVDD

SR-WSVDD 算法损失函数如式(4)所示。对于不完全监督异常检测场景中未标记数据,将使用与 SR-SVDD 算法相同的损失项,即 L_S 损失项(式(2)、(4))。对于标记数据,此不完全监督异常检测算法引入了一个新的损失项,即 L_W 损失项(式(4)),该损失项通过超参数 $\eta(\eta > 0)$ 加权,控制标记损失项和未标记损失项之间的平衡,将 η 设置的更大表明更注重标记样本,将 η 设置的更小则更加强调未标注数据。当没有可用的有标记训练数据时,此不完全监督目标函数就会退化为 SR-SVDD 算法的无监督目标函数。

$$\begin{aligned} L &= L_S + L_W \\ L_W &= \frac{\eta}{m}\sum_{j=1}^{m}\left(\left\|\phi(\tilde{\boldsymbol{x}}_j;\boldsymbol{W}) - c\right\|^{\tilde{y}_j}\right) \end{aligned} \tag{4}$$

针对新增的有标注数据,目标函数中以标签为权重指数,另数据点到超球体中心 c 的距离平方为底,对数据进行精准惩罚。对于标注的正常数据,通过最小

化数据点到超球体中心 c 的距离平方,以实现正常样本的共性学习,将其映射到中心点 c 附近。对于标注的异常数据,该算法通过采用数据点到超球体中心 c 距离平方的倒数作为惩罚项,使得异常样本点被映射到更远离中心点的位置,这契合了异常数据不集中的一般性假设。通过新增的损失项 L_W,算法能够在学习阶段从有标签的数据中获得更多信息,增强了异常检测的能力。

在弱监督场景中,引入了异常样本增加了信息的不确定性。因此结合信息熵与互信息最大化传输原则对 SR-WSVDD 进行了原理性解释(式(5))。此算法输入涵盖了无标注样本 $\boldsymbol{X}$ 和有标注样本。经过特征提取,得到正常样本特征 $\boldsymbol{Z}$,标注异常样本特征 $\boldsymbol{Z}^-$ 以及标注正常样本特征 $\boldsymbol{Z}^+$。对于无标注正常数据仍然使用信息最大传输理论进行解释,即式(5)的第 1 项。第 2 项表明了,对于有标注样本,根据信息熵的概念,使正常样本分布更加有秩序,使异常样本更加错综复杂。从几何角度来看,正常样本特征更加紧凑在超球体中,而异常样本特征更加分散在超球体外。

$$\max_{p(z\mid x)} I(\boldsymbol{X};\boldsymbol{Z}) + \beta(H(\boldsymbol{Z}^-) - H(\boldsymbol{Z}^+)) \tag{5}$$

5 整体流程

SR-SVDD 与 SR-WSVDD 算法中的网络权重对于目标函数来说通常是非凸的,这在深度学习中非常常见,对于计算的有效优化,算法依靠梯度下降法使用反向传播来优化网络权重。首先建立用于构建超球体的自编码器,通过数据前向传播提取特征,进而利用特征的平均值更新超球体的中心。接着,算法运用分位数函数计算超球体半径。最终,基于损失函数对网络权重进行调整。通过循环迭代上述过程,逐步建立超球体。

6 实验验证

6.1 数据集

为解释本文模型的工作原理,在鸢尾花数据集上进行本模型的原理实验;为验证本文模型的性能,将本文算法与最先进的异常检测模型在 MNIST[29] 与 CIFAR-10[30] 两个基准数据集上进行对比,体现了本文算法在异常检测任务中的优越性。

鸢尾花数据集包含了 3 个不同种类的鸢尾花,每个类别各 50 个样本。每个样本具有 4 个特征:花萼长度、花萼宽度、花瓣长度和花瓣宽度。本文将使用该数据集进行原理实验,从而清晰地展示 SR-SVDD 算法收敛与进行异常检测任务的过程。MNIST 为手写数字图像数据集,涵盖了从 0~9 这 10 个数字。每个数字类别有约 70 000 个 28×28 pixels 的灰度图,表示了手写数字的笔迹样本,总共包含 60 000 个训练图像和 10 000 个测试图像。CIFAR-10 包含了来自 10 个不同类别的 60 000 张彩色图像,每个类别有 6 000 张 32×32 pixels 的彩色图像,包括飞机、汽车、鸟类、猫、狗等类别。

6.2 实验设置

实验在 2080Ti 型号 GPU 上进行训练和测试。在神经网络模型方面,本算法基于 PyTorch 框架搭建神经网络模型,使用 LeNet 类型的自编码器进行实验。实验通过理论分析和实验研究进行超参数设置。首先两个算法都选用 Adam 作为优化器,初始学习率设置为 0.000 1 并以一定的速度进行衰减,SR-SVDD 算法的批数据数量为 200,SR-WSVDD 的批数据数量为 128。然后 SR-SVDD 与 SR-WSVDD 在算法执行前向传播后,将超球体中心坐标点 c 设置为映射数据的平均值。经过多次实验,将算法中用于计算球体半径的分位数定为 0.5,将损失函数中的 C_1 设置为 1,C_2 设置为 10,将 SR-WSVDD 中的 $\eta(\eta > 0)$ 设置为 1。

由于实验使用的数据集中不存在正常类别与异常类别,因此需要对数据集的正常类别与异常类别进行定义。在进行原理性实验时,将 setosa 类鸢尾花作为正常类别,将其他两类作为异常类别。在消融实验与对比实验中,对 MNIST 和 CIFAR-10 数据集进行划分。每次实验额外添加一类有标注的异常样本,有标注的异常样本数量占总样本数量的 0.01。

在 MNIST 数据集的实验中,本文使用 8×(5×5×1) 和 4×(5×5×1) 的两层卷积核,连接全连接层以及相应的解码器;对于 CIFAR-10 数据集,编码器部分使用 32×(5×5×3),64×(5×5×3) 和 128×(5×5×3) 的 3 层卷积网络,连接全连接层以及相应的解码器。根据参数量与运算量计算公式进行计算,应用在 MNIST 数据集上的模型参数量与运算量分别为 4 073 和 5 292 240,属于小型模型,CIFAR-10 数据集上应用的模型参数量分别为 3 325 955 和 181.97×10^6,属于中等模型。这样的参数量与运算量在大模型时代可以高效地进行异常检测任务。

6.3 原理实验

为了更加清晰地表达 SR-SVDD 与 SR-WSVDD 算法收敛和异常检测的过程,本文使用可视化的方法在鸢尾花数据集上进行模拟验证。

1)SR-SVDD 原理实验

为了可视化地展示模型收敛过程,实验通过神经网络将鸢尾花数据集的 4 个特征压缩为两个关键特征,从而得出迭代过程中的样本分布情况,具体呈现如图 4 所示。图 4 中分别为不同迭代轮次中训练样本与测试样本

的分布情况。可以发现随着迭代次数的增加,正常样本分布逐渐具有规律性,异常样本与正常样本间的差异程度也在逐步增加。通过原理实验结果可以发现,在新损失函数的作用下,SR-SVDD 算法的网络结构中即使包含偏置项,同样顺利实现了球体构建。由此可以说明,新损失函数中利用向量间相互作用的思想可以帮助具有偏置项网络避免球体崩塌现象,即避免了学习零权重以及零半径的最优解。

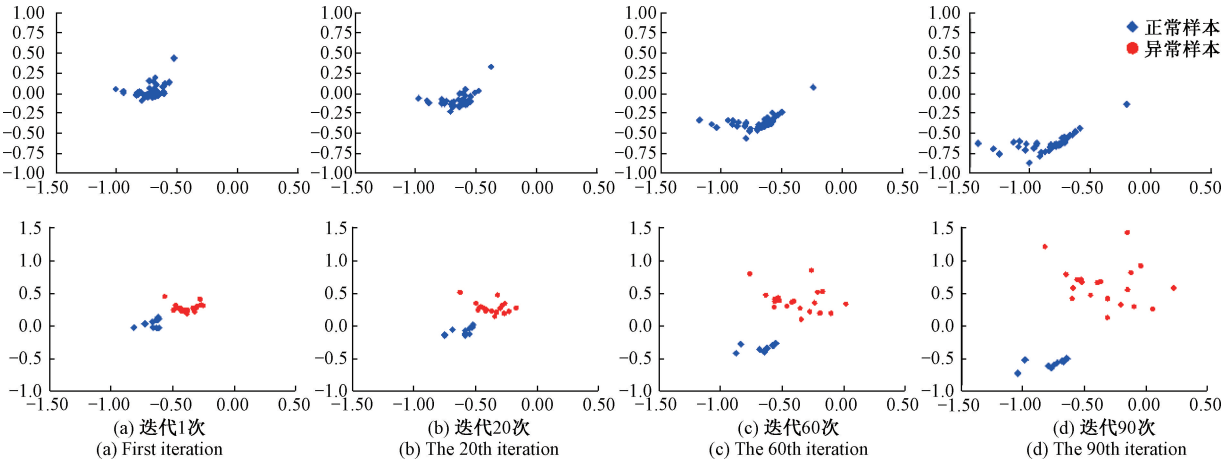


图 4　SR-SVDD 算法原理实验样本分布

Fig. 4　SR-SVDD algorithm principle experimental sample distribution

为验证在 SR-SVDD 算法中,异常检测模型可以得到良好的训练,SR-SVDD 算法在 MNIST 数据集上训练过程中损失值的变化过程如图 5 所示。损失值顺利下降的趋势说明模型得到了良好的训练。

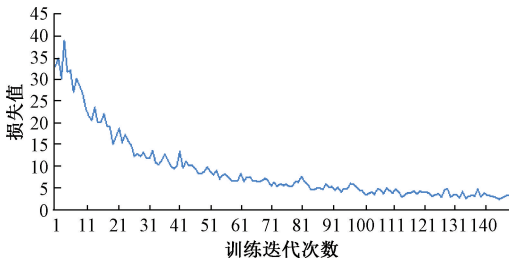


图 5　SR-SVDD 在 MNIST 数据集上的收敛情况

Fig. 5　Convergence of SR-SVDD on MNIST dataset

为进一步验证 SR-SVDD 算法在图像异常检测领域的有效性,本文在公共安全自采数据集上进行异常检测验证实验,该数据包含采集人员手持 6 类物品的图像,包括锤子、刀锯、斧子和棍棒等物品。验证过程中将锤子设定为正常类别,其他 4 类作为异常物品,并根据样本特征与球心距离进行统计,即 $\|\phi(\boldsymbol{x};\boldsymbol{W}) - \boldsymbol{c}\|$。实验结果如图 6 所示,包含各类物品与球心距离均值。根据图 6 可以发现,作为正常样本的锤子,其距离均值低于其他物品,即更趋向于球心,而其他物品分布均远离球心,因此更容易被识别为异常,从而验证了 SR-SVDD 在图像异常检测方面的有效性。

2)SR-WSVDD 原理实验

不同迭代次数下训练样本和测试样本分布的变化如图 7 所示。正常样本的分布逐渐趋于球体,而异常样本则在迭代次数增加的过程中逐渐远离正常样本的聚集区域。这些观察结果充分证明了算法的有效性与适用性。

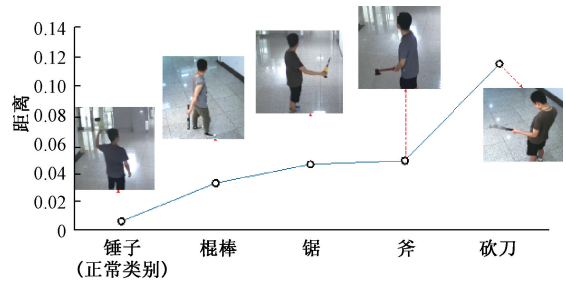


图 6　SR-SVDD 在公共安全自采数据集上的异常检测结果

Fig. 6　Detection results of SR-SVDD on public security self-collected dataset

6.4　消融实验

为了探究 SR-SVDD 相较于 Deep SVDD 的优越性,本研究在 MNIST 和 CIFAR-10 上进行了一系列实验。实验选用了 SR-SVDD、Deep SVDD 以及软间隔 DeepSVDD 3 类算法,以保持主干神经网络、参数等设置的一致性,确保消融实验的公平性。实验的结果如图 8 所示,其中 x 轴表示正常样本类别,y 轴表示异常检测任务的 AUC 值。由图 8 可以看出,SR-SVDD 算法在异常检测性能上具有明显的优势。

相对于软间隔 Deep SVDD,在绝大多数情况下,SR-SVDD 算法都呈现出更为好的检测效果,与 Deep SVDD 算法相比也有一定的性能领先。针对 MNIST 数

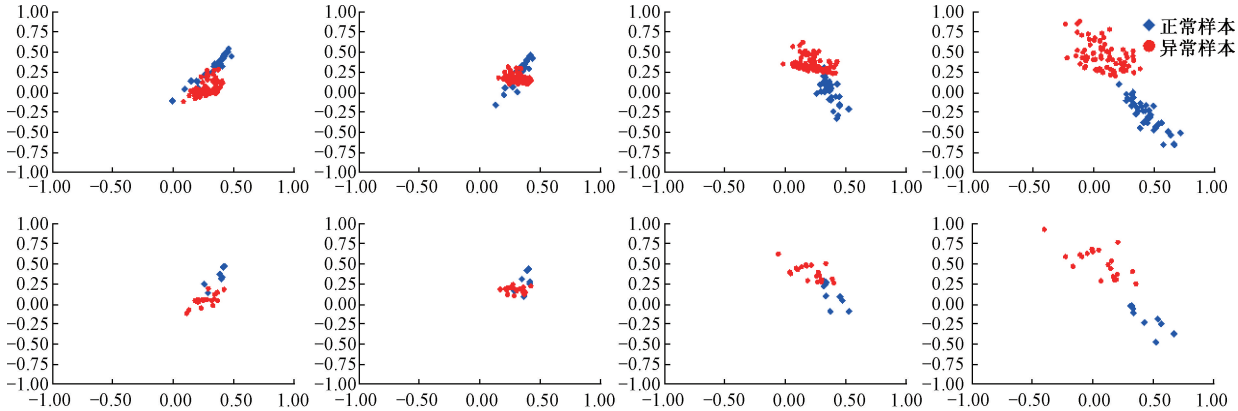


图 7 SR-WSVDD 算法原理实验样本分布

Fig. 7 SR-WSVDD algorithm principle experimental samples distribution

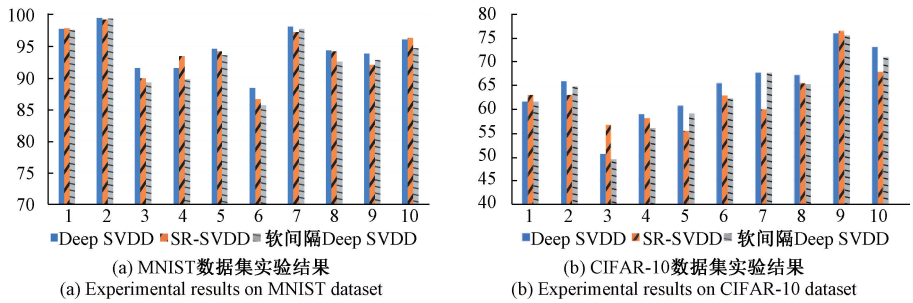


图 8 损失函数不同时数据集的 AUC 值

Fig. 8 The AUC of the dataset when the loss function is different

据集的实验中,当将 7 类数字标定为正常样本时,SR-SVDD 算法在异常检测方面胜过了软间隔 Deep SVDD。同样,在 CIFAR-10 的实验中,当将 6 类不同种类定义为正常样本时,SR-SVDD 的异常检测成绩也显著优于软间隔 Deep SVDD 算法。这一系列实验证明了 SR-SVDD 在异常检测领域的优秀性能。

为验证 SR-WSVDD 相对于单类分类算法的改进效果,本文对 SR-WSVDD、SR-SVDD 以及 Deep SVDD 算法进行比较,以探究弱监督异常检测算法的卓越性。检测性能 AUC 值如图 9 所示。在实验过程中,为确保对照消融实验的公平性,除数据集和损失函数设置外,3 种算法的其他参数和结构均保持一致。从图 9 可以看出,SR-WSVDD 在 MNIST 数据集上的绝大部分类别的 AUC 值都高于 Deep SVDD 以及 SR-SVDD,在 CIFAR-10 数据集上的每一类的 AUC 值都高于上述两类单类分类算法的 AUC 值。综上所述,此弱监督异常算法的检测能力优于相同设置下单类异常检测算法。

为了验证 SR-WSVDD 算法中损失函数的有效性,将

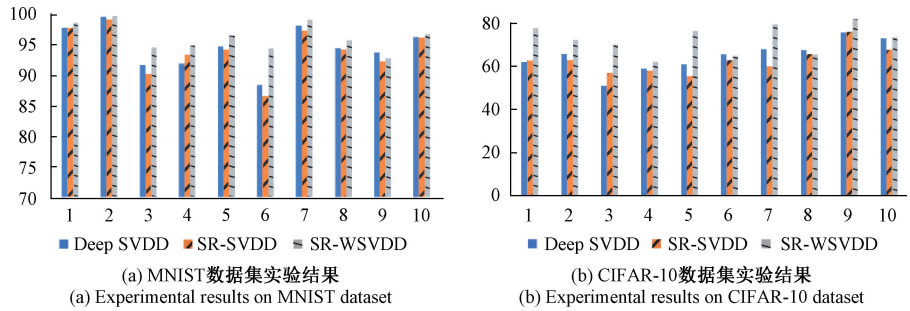


图 9 单类算法与弱监督算法的 AUC 值

Fig. 9 AUC of one-class algorithms and weakly supervised algorithms

此算法与 SAD 算法进行对照实验。两个算法的网络框架,训练策略和数据集设置相同,设置不同的损失函数进行弱监督损失函数有效性验证。两种算法在两个数据集上分别进行 90 次实验后得出的 AUC 平均值如表 1 所示。从表 1 可以看出,在具有额外松弛项的损失函数帮助下,SAD 弱监督应用算法与 SR-WSVDD 算法取得了相似的效果。说明了该弱监督损失函数的有效性。

表 1　不同损失函数下的弱监督算法 AUC 值

Table 1　AUC of weakly supervised algorithms under different loss functions　%

数据集	SR-SVDD	SAD[18]
MNIST	96.45	96.40
CIFAR-10	72.27	72.60

通过上述消融实验,对 SR-SVDD 损失函数及其弱监督算法损失函数的有效性进行了验证,证明了基于支持向量进行样本训练学习理论的正确性。同时,将其理论应用于弱监督领域,其效果优于其单类分类方法,证明了弱监督算法相较于单类算法的优越性。

6.5　对比实验

本文对提出的 SR-SVDD 与经典单类算法模型以及基于深度学习的单类算法模型在 MNIST 和 CIFAR-10 数据集上进行对比分析,如表 2、3 所示。在 MNIST 上,本文算法对各类别的检测平均值优于原有经典单类算法及基于深度学习的单类算法。在 CIFAR-10 数据集中,几乎所有类别的 AUC 指标都优于或接近深度卷积自编码器和最新的改进 GAN 算法。因此,充分证明了 SR-SVDD 在异常检测任务中的优秀表现。

表 2　不同单类异常检测算法在 MNIST 检测结果

Table 2　Results of different one-class anomaly detection algorithms in MNIST　%

传统单类分类算法			深度单类分类算法		
OC-SVM/SVDD[3,31]	KDE[4]	IF[32]	AnoGAN[33]	DCAE[34]	SR-SVDD
91.29	86.93	92.3	91.27	89.65	**94.36**

表 3　不同单类异常检测算法在 CIFAR-10 检测结果

Table 3　Results of different one-class anomaly detection algorithms in CIFAR-10　%

Normal Class	传统单类分类算法			深度单类分类算法		SR-SVDD
	OC-SVM/SVDD[3,31]	KDE[4]	IF[32]	AnoGAN[33]	DCAE[34]	
AIRPLANE	61.6	61.2	60.1	**67.1**	59.1	63.0
AUTOMOBILE	63.8	**64.0**	50.8	54.7	57.4	63.3
BIRD	50.0	50.1	49.2	52.9	48.9	**56.8**
CAT	55.9	56.4	55.1	54.5	**58.4**	58.0
DEER	66.0	**66.2**	49.8	65.1	54.0	55.5
DOG	62.4	62.4	58.5	60.3	62.2	**62.8**
FROG	74.7	**74.9**	42.9	58.5	51.2	60.2
HORSE	62.6	62.6	55.1	62.5	58.6	**65.4**
SHIP	74.9	75.1	74.2	75.8	76.8	76.3
TRUCK	75.9	**76.0**	58.9	66.5	67.3	67.7

经过 SR-SVDD 与单类异常检测算法的比较后,将 SR-WSVDD 与其他弱监督算法、有监督算法以及单类异常检测算法进行对比,对比结果如表 4 所示。由表 4 看出,SR-WSVDD 在性能上相对于有监督分类器表现出显著的提升。在 MNIST 和 CIFAR-10 数据集上,该算法的性能分别提高了 3.7% 和 16.7%。这一结果表明提出的弱监督算法在解决数据不平衡问题方面发挥了重要作用。此外,与其他弱监督算法相比,在相同的神经网络结构下,与基于主动学习的 SSAD 算法相比,SR-WSVDD 算法在 CIFAR-10 数据集上提升了 1.8%,在 MNIST 数据集上取得了类似的效果。与基于生成的 SS-DGM 算法相比,SR-WSVDD 算法的性能提升更为显著,分别在两个数据集上分别提高了 6.6% 和 22.6%。最后,SR-WSVDD 算法相对于先进的单类算法,不仅在 MNIST 上表现出相似的性能,而且在 CIFAR-10 上还超越了表 4 中其他两种先进单类算法。综上所述,通过与有监督学习算法、弱监督学习算法和单类分类算法的综合比较,SR-WSVDD 明确展示了其在异常检测领域的有效性和领先性。

表 4 不同异常算法 AUC 性能对比

Table 4 AUC performance comparison of different anomaly algorithms %

Data	弱监督算法			有监督算法	单类分类算法		SR-WSVDD
	SSAD Raw[26]	SSAD Hybird[26]	SS-DGM[27]	Supervised Classifier	OCGAN[28]	DAGPR[35]	
MNIST	96.6	96.8	89.9	92.8	97.5	96.1	96.5
CIFAR-10	73.0	70.5	49.7	55.6	65.7	66.9	72.3

7 结 论

本文针对异常检测领域常用算法 Deep SVDD 容易出现球体崩塌现象的原因进行深入研究。针对球体崩塌问题,本文提出了 SR-SVDD 和 SR-WSVDD 两种新算法,分别适用于单类和弱监督异常检测两种不同情况。这两个算法结合了 SVM 和支持向量的思想,在提供了合理的学习策略的同时,避免了球体崩塌现象的发生。本文解释了这两个新算法的原理,并通过信息传输极大原则和互信息最大化传输原则对其进行了理论解释。此外,新算法的运行过程为算法的合理性提供了支持。在对比实验中,SR-SVDD 和 SR-WSVDD 两种算法优于现有的大部分异常检测算法,并证实了它们成功地克服了球体崩塌问题。

参考文献

[1] PANG G, SHEN C, CAO L, et al. Deep learning for anomaly detection: A review [J]. ACM computing surveys (CSUR), 2021, 54(2): 1-38.

[2] CHALAPATHY R, MENON A K, CHAWLA S. Anomaly detection using one-class neural networks[J]. ArXiv Preprint, 2018, ArXiv:1802.06360.

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

[4] PARZEN E. On estimation of a probability density function and mode[J]. The Annals of Mathematical Statistics, 1962, 33(3): 1065-1076.

[5] ZHANG D, GATICA-PEREZ D, BENGIO S, et al. Semi-supervised adapted hmms for unusual event detection[C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2005, 1: 611-618.

[6] KIM J, GRAUMAN K. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates[C]. 2009 CVPR. IEEE, 2009: 2921-2928.

[7] LI F, YANG W, LIAO Q. An efficient anomaly detection approach in surveillance video based on oriented GMM[C]. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 1981-1985.

[8] CONG Y, YUAN J, LIU J. Sparse reconstruction cost for abnormal event detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2011: 3449-3456.

[9] LU C W, SHI J P, JIA J Y. Abnormal event detection at 150 FPS in MATLAB[C]. ICCV. IEEE, 2013: 2720-2727.

[10] CHALAPATHY R, CHAWLA S. Deep learning for anomaly detection: A survey[J]. ArXiv Preprint, 2019, ArXiv:1901.03407.

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

[12] HAWKINS S, HE H, WILLIAMS G, et al. Outlier detection using replicator neural networks[C]. DaWak. Berlin, Heidelberg: Springer, 2002: 170-180.

[13] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]. International Conference on Machine Learning, 2008: 1096-1103.

[14] MAKHZANI A, FREY B J. Winner-take-all autoencoders[J]. Advances in Neural Information Processing Systems, 2015, 28.

[15] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[16] PANG G, VAN DEN HENGEL A, SHEN C, et al. Deep reinforcement learning for unknown anomaly detection[J]. ArXiv Preprint, 2020, ArXiv:2009.06847.

[17] PANG G, SHEN C, JIN H, et al. Deep weakly-supervised anomaly detection[J]. ArXiv Preprint, 2019, ArXiv:1910.13601.

[18] ZHANG J, QING L, MIAO J. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection[C]. 2019 ICIP. IEEE, 2019: 4030-4034.

[19] TIAN Y, PANG G, CHEN Y, et al. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning[C]. Proceedings of the ICCV,2021: 4975-4986.

[20] FENG J C, HONG F T, ZHENG W S. Mist: Multiple instance self-training framework for video anomaly detection[C]. Proceedings of the ICCV, 2021: 14009-14018.

[21] MASCI J, MEIER U, CIRESAN D, et al. Stacked convolutional auto-encoders for hierarchical feature extraction[C]. ICANN. Berlin, Heidelberg: Springer, 2011: 52-59.

[22] CHONG Y S, TAY Y H. Abnormal event detection in videos using spatiotemporal autoencoder[C]. ISNN, 2017: 189-196.

[23] XU D, RICCI E, YAN Y, et al. Learning deep representations of appearance and motion for anomalous event detection[C]. BMVC, 2015: 1-12.

[24] ZHAO Y, DENG B, SHEN C, et al. Spatio-temporal autoencoder for video anomaly detection[C]. ACM Multimedia, 2017: 1933-1941.

[25] RUFF L, VANDERMEULEN R A, GÖRNITZ N, et al. Deep semi-supervised anomaly detection[C]. ICLR 2020 Conference Program Chairs, 2020: 854.

[26] PANG G, CAO L, CHEN L, et al. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection[C]. KDD, 2018: 2041-2050.

[27] SULTANI W, CHEN C, SHAH M. Real-world anomaly detection in surveillance videos[C]. CVPR. IEEE, 2018: 6479-6488.

[28] PERERA P, NALLAPATI R, XIANG B. Ocgan: One-class novelty detection using gans with constrained latent representations[C]. CVPR,2019: 2898-2906.

[29] LECUN Y, CORTES C, BURGES C J C. THE MNIST DATABASE of handwritten digits[Z].

[30] JINLIANG N. Cifar10 image classification based on ResNet[J]. System Analysis in Engineering and Control, 2019, 23(1): 412-415.

[31] TAX D M J, DUIN R P W. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.

[32] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]. 8th IEEE International Conference on Data Mining. IEEE, 2008: 413-422.

[33] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery[C]. IPMI. Cham: Springer, 2017: 146-157.

[34] QIAN F, YIN M, LIU X Y, et al. Unsupervised seismic facies analysis via deep convolutional autoencoders[J]. Geophysics, 2018, 83(3): A39-A43.

[35] TSCHANNEN M, DJOLONGA J, RUBENSTEIN P K, et al. On mutual information maximization for representation learning[J]. ArXiv Preprint, 2019, ArXiv:1907.13625.

作者简介

邓诗卓,2013 年于东北大学获得学士学位,2015 年于东北大学获得硕士学位,2020 年于东北大学获得博士学位,现为东北大学讲师,主要研究方向为时序数据分析、小样本学习、机器学习、数据库、视觉异常检测。E-mail:dengshizhuo@mail.neu.edu.cn

Deng Shizhuo received her B.Sc. degree in 2013 from Northeastern University, received her M.Sc. degree in 2015 from Northeastern University, received her Ph.D. degree in 2020 from Northeastern University, now she is lecture in Northeastern University. Her main research interests include time-series data analysis, Few-shot learning, machine learning, databases and visual anomaly detection.