Deep Semi-Supervised Anomaly Detection

摘要

异常检测，就是在数据集中发现那些不符合预期行为或模式的数据点的过程。这些异常点可能是由于数据系统故障、欺诈行为、罕见事件或其他特殊情况所导致。这对工业，金融等行业十分重要。本文对 Lukas Ruff 等人发表在 ICLR 2020 的 Deep SAD 进行解读复现其结果，充分认识了异常检测的相关方法理论。本文首先基于 Lukas Ruff 等人发表在 ICML 2018 的论文介绍无监督学习的方法，该论文引入超球面来对异常数据进行清洗。紧接着介绍 Deep SAD 模型。最后，通过通过对方法的复现以及大量的实验结果充分地体现了 Deep SAD 与其他方法相比，即使提供少量的标记数据也能带来显著的性能提升。

关键词：异常检测；半监督学习；表示学习；自监督学习

1 引言

随着深度学习的快速发展，深度异常检测方法在大型复杂数据集上显示出比浅层方法更好的效果。通常，异常检测被视为无监督学习问题。然而在实践中，除了大量未标记样本外，人们还可以访问一小部分标记样本。例如，由某些领域专家验证为正常或异常的部分。半监督异常检测方法旨在利用此类标记样本，但大多数提出的方法仅限于包括标记的正常样本。只有少数方法利用了标记异常，现有的深度方法都是特定于领域的。因此，为应对更多情况提高模型的泛化能力，Lukas Ruff 等人提出了 Deep SAD。它能够在异常样本较多的情况下也能很好的分辨出异常和正常样本。

2 相关工作

异常检测 (AD) 是识别数据中异常样本的任务。通常，异常检测方法试图以无监督的方式学习数据的“紧凑”描述，假设大多数样本都是正常的（即非异常）。例如，在一类分类中 [20])，目标是找到一组包含大多数数据的小测量值，不包含在该集合中的样本被视为异常。浅层无监督异常检测方法，例如单类 SVM [22]、核密度估计 [23] 或孤立森林 [11] 通常需要手动特征工程才能对高维数据有效，并且其对大型数据集的可扩展性有限。这些局限性引起了人们对开发新型无监督异常检测深度方法的极大兴趣 [18]。

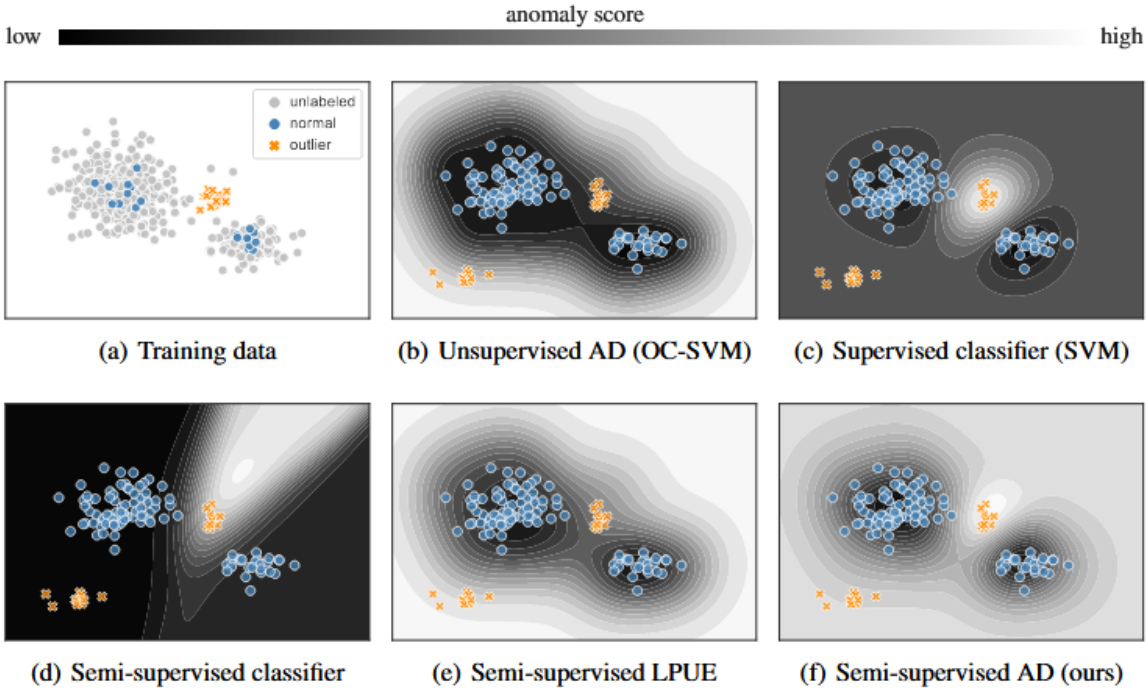


图 1. 半监督异常检测的必要性：训练数据（如图 (a) 所示）包括（主要是正常的）未标记数据（灰色），以及一些标记的正常样本（蓝色）和标记的异常（橙色）。图 (b)-(f) 展示了在测试时各种学习范式的决策边界，以及每个图中左下角出现的新型异常。半监督异常检测方法利用了所有训练数据：未标记样本、标记的正常样本以及标记的异常。这在单类学习和分类之间取得了平衡。

2.1 基于核的单分类

可能最突出的核基础单类分类方法是单类支持向量机（One-Class SVM，OC-SVM）。OC-SVM 的目标是在特征空间中找到一个最大间隔超平面 $w \in \mathcal{F}_k$，它最好地将映射后的数据与原点分离。给定数据集 $\mathcal{D}_n = \{x_1, \ldots, x_n\}$，其中 $x_i \in \mathcal{X}$，OC-SVM 解决以下原始问题：

$$\min_{w,\rho,\xi} \frac{1}{2}\|w\|^2_{\mathcal{F}_k} - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \tag{1}$$

$$\text{s.t.} \quad \langle w, \phi_k(x_i)\rangle_{\mathcal{F}_k} \geq \rho - \xi_i, \xi_i \geq 0, \quad \forall i. \tag{2}$$

这里 ρ 是从原点到超平面 w 的距离。非负松弛变量 $\xi = (\xi_1, \ldots, \xi_n)^T$ 允许边界是软的，但违反 ξ_i 会受到惩罚。$\|w\|^2_{\mathcal{F}_k}$ 是超平面 w 的正则化项，其中 $\|\cdot\|_{\mathcal{F}_k}$ 是由 $\langle\cdot,\cdot\rangle_{\mathcal{F}_k}$ 诱导的范数。超参数 $\nu \in (0, 1]$ 控制目标中的权衡。在特征空间中将数据与原点分离转化为找到一个半空间，其中大部分数据位于其中，而位于这个半空间之外的点，即 $\langle w, \phi_k(x)\rangle_{\mathcal{F}_k} < \rho$，被认为是异常的。

支持向量数据描述（SVDD）是一种与 OC-SVM 相关的技术，它使用超球面而不是超平面来分离数据。SVDD 的目标是在特征空间 $\mathcal{F}_k$ 中找到包含大多数数据的最小超球面，其中心为 $c \in \mathcal{F}_k$，半径 $R > 0$。SVDD 的原始问题由以下公式给出：

$$\min_{R,c,\xi} R + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \tag{3}$$

$$\text{s.t. } \|\phi_k(x_i) - c\|^2_{\mathcal{F}_k} \le R^2 + \xi_i, \xi_i \ge 0, \quad \forall i. \tag{4}$$

同样，松弛变量 $\xi_i \ge 0$ 允许软边界，超参数 $\nu \in (0,1]$ 控制惩罚 ξ_i 和球体体积之间的权衡。落在球体之外的点，即 $\|\phi_k(x) - c\|^2_{\mathcal{F}_k} > R^2$，被认为是异常的。

OC-SVM 和 SVDD 密切相关。两种方法都可以通过它们各自的对偶问题来解决，这些对偶问题都是二次规划 [24]，可以通过多种方法解决，例如序列最小优化。在广泛使用的高斯核的情况下，两种方法等价，并且是渐近一致的密度水平集估计器 [25]。使用超参数 $\nu \in (0,1]$ 来制定原始问题（如公式 (1) 和 (2)）是一个方便的参数选择，因为 $\nu \in (0,1]$ 是（i）异常分数的上界，以及（ii）支持向量（位于边界上或边界外的点）的分数的下界。这个结果被称为 ν-属性，它允许将关于训练数据中异常分数的先验信念纳入模型。

除了需要执行显式特征工程 [15] 之外，上述方法的另一个缺点是由于核的构建和操作导致的计算扩展性差。核方法在样本数量上的扩展性至少是二次的，除非使用某种近似技术 [16]。此外，使用核方法进行预测需要存储支持向量，这可能需要大量的内存。下面将介绍的深度支持向量数据描述（Deep SVDD）不受这些限制的影响。

2.2 深度支持向量数据描述

基于核基础的 SVDD 和最小体积估计，通过寻找最小尺寸的封闭数据的超球体来构建。为此，我们采用一个神经网络 [19]，该网络被联合训练以将数据映射到最小体积的超球体中。

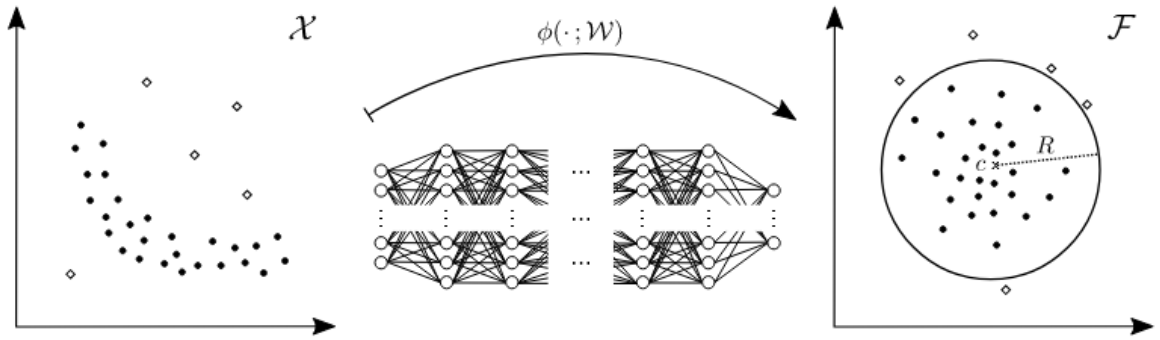


图 2. Deep SVDD 学习了一个神经网络变换 $d(\cdot;\mathcal{W})$，其权重为 $\mathcal{W}$，从输入空间 $\mathcal{X} \subseteq \mathbb{R}^d$ 到输出空间 $\mathcal{F} \subseteq \mathbb{R}^p$，试图将大多数数据网络表示映射到一个以中心 c 和最小体积半径 R 为特征的超球体中。正常样本的映射落在超球体内部，而异常的映射落在超球体外部。

对于输入空间 $\mathcal{X} \subseteq \mathbb{R}^D$ 和输出空间 $\mathcal{Z} \subseteq \mathbb{R}^d$，令 $\phi(\cdot;\mathcal{W}) : \mathcal{X} \to \mathcal{Z}$ 为一个具有 L 个隐藏层和相应权重集 $\mathcal{W} = \{W^1, \ldots, W^L\}$ 的神经网络。Deep SVDD 的目标是训练神经网络 ϕ 学习一种变换，该变换最小化输出空间 $\mathcal{Z}$ 中以预定点 c 为中心的数据封闭超球体的体积。给定 n 个（未标记的）训练样本 $x_1, \ldots, x_n \in \mathcal{X}$，单类 Deep SVDD 目标为

$$\min_{\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n}\|\phi(x_i;\mathcal{W}) - c\|^2 + \frac{\lambda}{2}\sum_{\ell=1}^{L}\|W^\ell\|_F^2, \quad \lambda > 0. \tag{5}$$

惩罚映射样本到超球体中心 c 的平均平方距离迫使网络提取数据集中最稳定的共同变化因素。因此，正常数据点倾向于被映射到超球体中心附近，而异常点则被映射到更远的地。第二项是标准的权重衰减正则化项。Deep SVDD 通过使用反向传播的随机梯度下降（SGD）进行优化。对于初始化，首先预训练一个自编码器，然后使用编码器的收敛权重初始化网络 ϕ 的权重 $\mathcal{W}$。初始化后，超球体中心 c 被设置为从数据的初始前向传播中获得的网络输出的平均值。一旦网络训练完成，测试点 x 的异常分数由 $\phi(x;\mathcal{W})$ 到超球体中心的距离给出：

$$s(x) = \|\phi(x;\mathcal{W}) - c\|. \tag{6}$$

2.3 半监督异常检测

半监督异常检测一词已被用来描述两种不同的异常检测设置。大多数现有的半监督异常检测方法，无论是浅层的 [2,4,13] 还是深层的 [1,3,21]，仅结合了标记的正常样本，而没有标记的异常样本，即它们更精确地是学习正样本（即正常样本）和未标记样本（LPUE）[26] 的实例。一些工作 [7] 已经研究了一般半监督 AD 设置，其中也利用了标记的异常样本，然而现有的深度学习方法是特定领域或数据类型的 [6,9,12]。

深度半监督学习的研究几乎完全集中在分类作为下游任务 [5,8,14,17]。这种半监督分类器通常假设相似的点可能属于同一类，这被称为聚类假设 [10]。然而，这种假设仅适用于 AD 中的“正常类”，但对于“异常类”则至关重要，因为异常不一定彼此相似。相反，半监督异常检测方法必须找到一个紧凑的正常类描述，同时正确区分标记的异常 [7]。

3 本文方法

3.1 本文方法概述

本文方法与上述 Deep SVDD 一脉相承。也是为了找到一个超球面，使得正常样本映射到球面内，异常样本映射到球面外。对于图像数据（如 MNIST、Fashion-MNIST 和 CIFAR-10），该方法使用了卷积神经网络（CNN）作为编码器和解码器的核心组件。对于表格或其他非图像数据，使用了全连接网络（Fully Connected Network, FCN）作为编码器和解码器。

3.2 损失函数定义

假设除了 n 个未标记样本 $x_1,\ldots,x_n \in \mathcal{X}$ 其中 $\mathcal{X} \subseteq \mathbb{R}^D$ 之外，还可以访问 m 个标记样本 $(\tilde{x}_1,\tilde{y}_1),\ldots,(\tilde{x}_m,\tilde{y}_m) \in \mathcal{X}\times\mathcal{Y}$，其中 $\mathcal{Y}=\{-1,+1\}$，$\tilde{y}=+1$ 表示已知的正常样本，而 $\tilde{y}=-1$ 表示已知的异常。Deep SAD 目标如下：

$$\min_{\mathcal{W}} \frac{1}{n+m}\sum_{i=1}^{n}\|\phi(x_i;\mathcal{W})-c\|^2 + \frac{\eta}{n+m}\sum_{j=1}^{m}\left(\|\phi(\tilde{x}_j;\mathcal{W})-c\|^2\right)^{\tilde{y}_j} + \frac{\lambda}{2}\sum_{\ell=1}^{L}\|W^\ell\|_F^2 \tag{7}$$

在 Deep SAD 目标中为未标记数据采用了与 Deep SVDD 相同的损失项，因此在没有标记训练数据可用（$m=0$）的特殊情况下，Deep SAD 和 Deep SVDD 是一样的。

对于标记数据，引入了一个新的损失项，该项通过超参数 $\eta>0$ 加权，这控制了标记项和未标记项之间的平衡。设置 $\eta>1$ 更加强调标记数据，而 $\eta<1$ 强调未标记数据。对于标

记的正常样本（$\tilde{y} = +1$），还对映射点到中心 c 的距离施加了二次损失，从而意图整体学习一个集中正常数据的潜在分布。同样，可能会考虑 $\eta > 1$ 来强调标记的正常样本超过未标记样本。对于标记的异常（$\tilde{y} = -1$），惩罚距离的倒数，使得异常必须映射到离中心更远的地方。

再次通过映射点到中心 c 的距离定义 Deep SAD 异常分数，如公式 (6) 所示，并通过使用 SGD 优化 Deep SAD 目标 (7)。

4 复现细节

4.1 与已有开源代码对比

Deep SAD 方法的官方代码已开源[1]，该代码采用的是 Pytorch 框架。基于算力受限，本文通过 MNIST，Fashion-MNIST，CIFAR-10 数据集进行训练测试[2]，通过尝试不同的 ratio_known_outlier（训练数据中标记为异常样本的比例），不同的 ratio_pollution（未标记训练数据中异常样本的比例），不同的 n_known_outlier_classes（标记训练数据中异常类别的种类数量）对模型进行评估，基本复现了论文中的结果。

4.2 实验环境搭建

本文在复现过程中所使用的实验环境如下:

- CPU:AMD Ryzen 7 4800H with Radeon Graphics
- GPU:NVIDIA GTX 1660ti 6GB
- python 3.10
- 框架:Pytorch 2.5.1
- 包管理：Anaconda
- CUDA:12.4

4.3 创新点

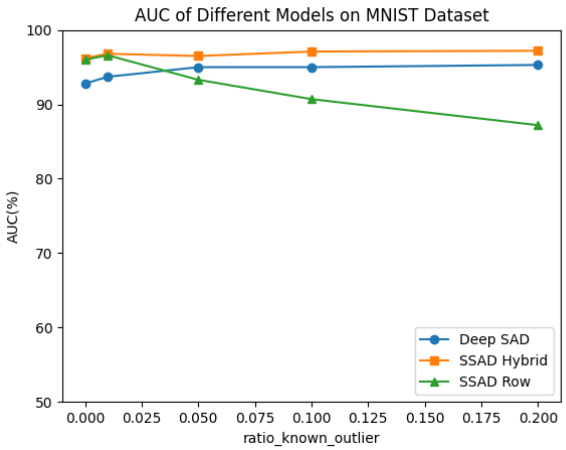
无

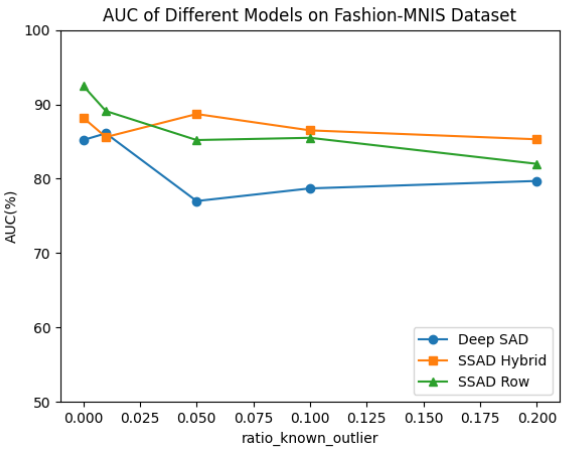
5 实验结果分析

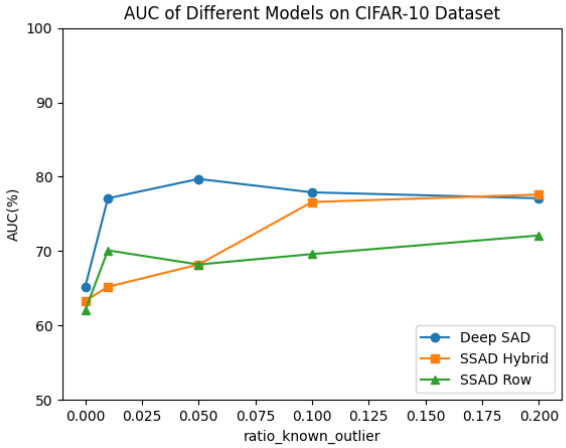
如下展示的是在 ratio_known_outlier 不同的情况下，对比 Deep SAD 与浅层，混合层方法在三个数据集上准确度（AUC）的表现。

[1] https://github.com/lukasruff/Deep-SAD-PyTorch

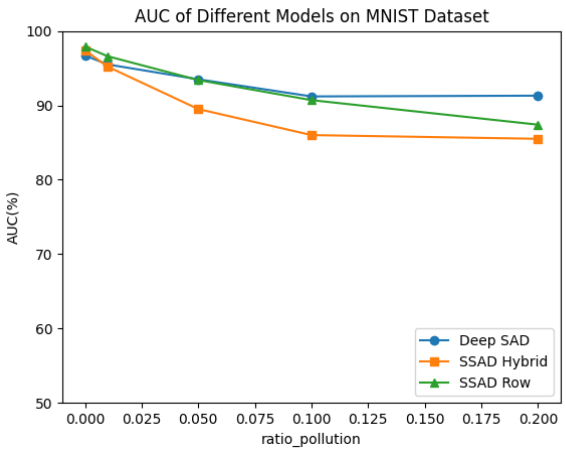
[2] https://github.com/lukasruff/Deep-SAD-PyTorch/blob/master/src/main.py

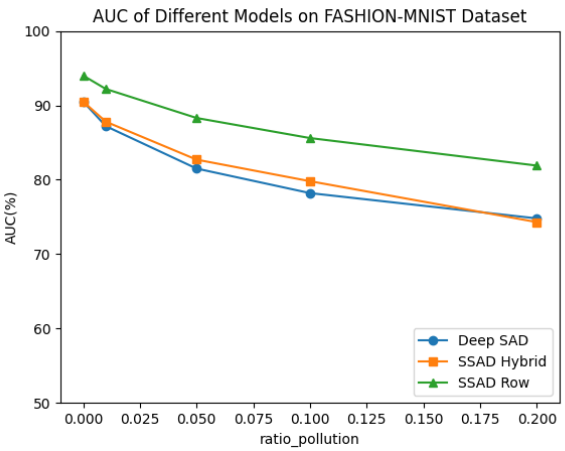


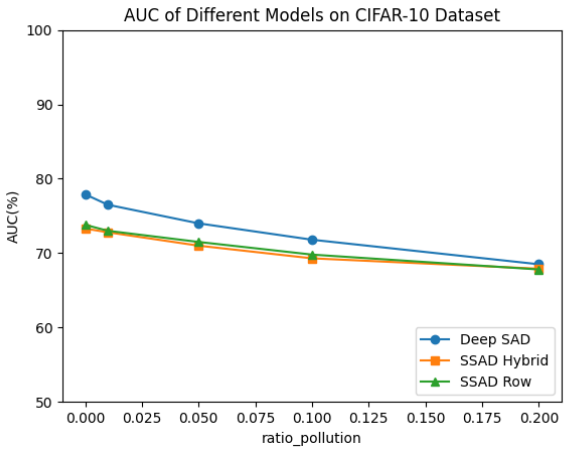




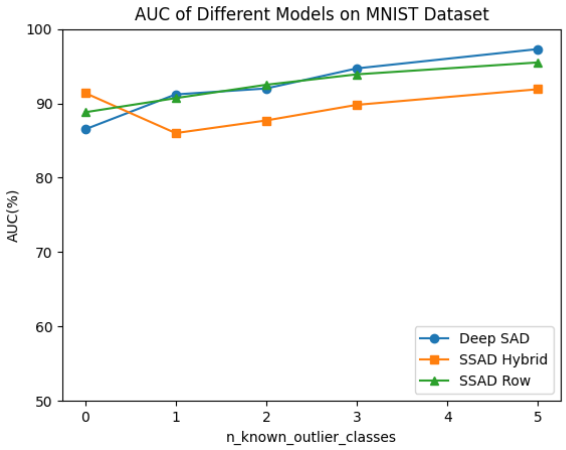
如下展示的是在 ratio_pollution 不同的情况下，对比 Deep SAD 与浅层，混合层方法在三个数据集上准确度（AUC）的表现。

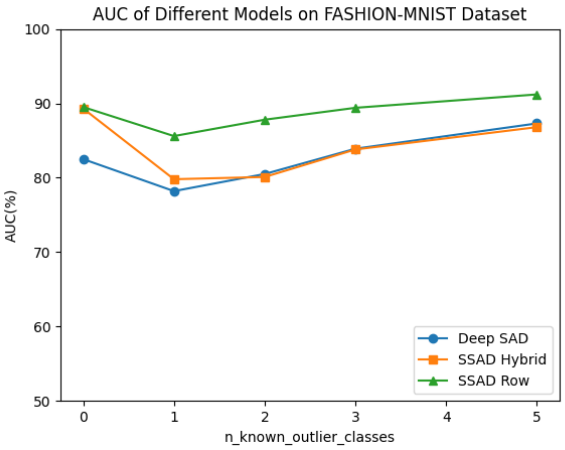


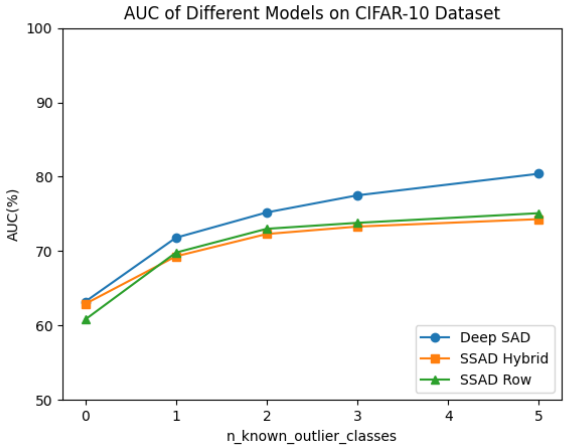




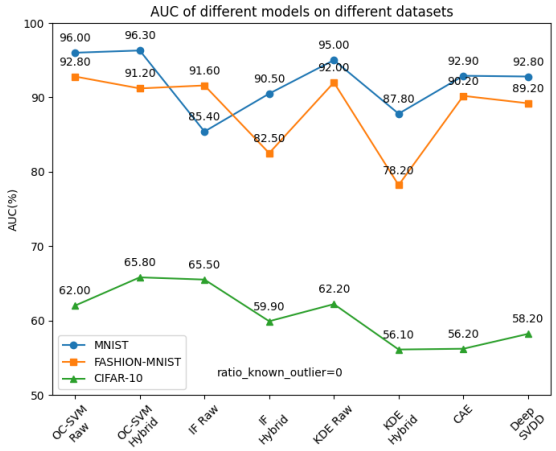
如下展示的是在 n_known_outlier_classes 不同的情况下，对比 Deep SAD 与浅层，混合层方法在三个数据集上准确度（AUC）的表现。

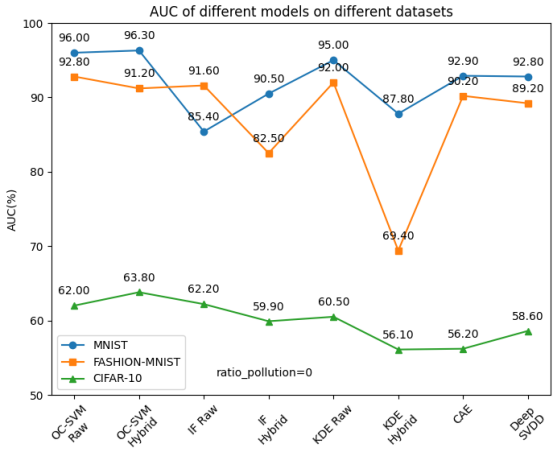


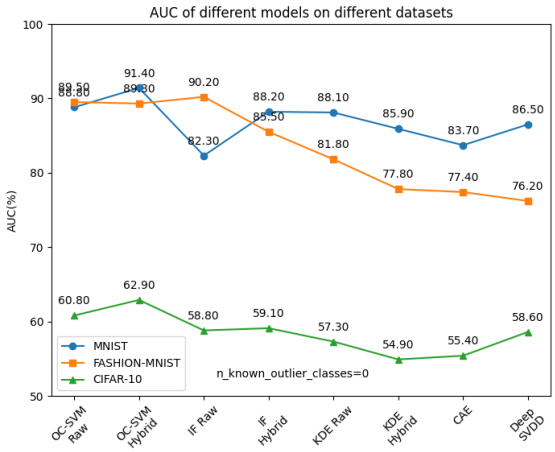




如下展示的是在 ratio_known_outlier，ratio_pollution，n_known_outlier_classes 均为 0 的情况下，其他各类方法在三个数据集上准确度（AUC）的表现。







6 总结与展望

本文对 Deep SAD 方法的进行了解读，并且从原理上解释了此异常检测方法的精髓之处，即是在单类 SVM 的启发下引入超球面，以此来区分各类样本的属性。并且鉴于目前的方法大多都是正常的，在预训练阶段都是假设所有样本都是正常样本来初始化参数，但现实的工业或各领域中难免会有其他异常较多的情况，该模型很好的解决了这一点。但受限于算力资源，本文仅采用三个较小数据集进行验证。因此在未来的工作中可以采用更多的数据集，并且尝试改性算法思想和模型以此来提高泛化能力。

参考文献

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer, 2019.

[2] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009, 2010.

[3] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[4] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[5] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Russ R Salakhutdinov. Good semi-supervised learning that requires a bad gan. *Proc. Conf. on Neural Information Processing Systems*, 30, 2017.

[6] Tolga Ergen and Suleyman Serdar Kozat. Unsupervised anomaly detection with lstm neural networks. *IEEE transactions on neural networks and learning systems*, 31(8):3127–3141, 2019.

[7] Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*, 46:235–262, 2013.

[8] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Proc. Conf. on Neural Information Processing Systems*, 27, 2014.

[9] B Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, 4(2):36, 2018.

[10] Semi-Supervised Learning. Semi-supervised learning. *CSZ2006. html*, 5:2, 2006.

[11] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[12] Erxue Min, Jun Long, Qiang Liu, Jianjing Cui, Zhiping Cai, and Junbo Ma. Su-ids: A semi-supervised and unsupervised framework for network intrusion detection. In *Cloud Computing and Security: 4th International Conference, ICCCS 2018, Haikou, China, June 8–10, 2018, Revised Selected Papers, Part III 4*, pages 322–334. Springer, 2018.

[13] Jordi Mũnoz-Marí, Francesca Bovolo, Luis Gómez-Chova, Lorenzo Bruzzone, and Gustavo Camp-Valls. Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE transactions on geoscience and remote sensing*, 48(8):3188–3197, 2010.

[14] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Proc. Conf. on Neural Information Processing Systems*, 31, 2018.

[15] Mahesh Pal and Giles M Foody. Feature selection for classification of hyperspectral data by svm. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2297–2307, 2010.

[16] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Proc. Conf. on Neural Information Processing Systems*, 20, 2007.

[17] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Proc. Conf. on Neural Information Processing Systems*, 28, 2015.

[18] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[19] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[20] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[21] Hongchao Song, Zhuqing Jiang, Aidong Men, and Bo Yang. A hybrid semi-supervised anomaly detection model for high-dimensional data. *Computational intelligence and neuroscience*, 2017(1):8501683, 2017.

[22] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54:45–66, 2004.

[23] Robert Vandermeulen and Clayton Scott. Consistency of robust kernel density estimators. In *Conference on Learning Theory*, pages 568–591. PMLR, 2013.

[24] Sreekanth Vempati, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Generalized rbf feature maps for efficient detection. In *Proc. British Machine Vision Conf.*, pages 1–11, 2010.

[25] Régis Vert, Jean-Philippe Vert, and Bernhard Schölkopf. Consistency and convergence rates of one-class svms and related algorithms. *Journal of Machine Learning Research*, 7(5), 2006.

[26] Bangzuo Zhang and Wanli Zuo. Learning from positive and unlabeled examples: A survey. In *2008 International Symposiums on Information Processing*, pages 650–654. IEEE, 2008.