时序异常检测综述整理(2020-2021)

Ai 宅码 2022-08-09 22:53 发表于广东

最近阅读几篇异常检测综述，这里整理分享给大家，推荐阅读：5星。不足之处，还望批评指正。

赵越博士的异常检测库Python Outlier Detection (PyOD) [1]写的很好，还提供了关于异常检测的学习资料 ，我阅读了几篇综述，个人比较推荐以下三篇：

• **2020 | Anomaly detection in univariate time-series: A survey on the state-of-the-art** ：全文偏基础，介绍了异常类型、时序模式、异常检测常见的模型方法，还有公开数据集和评估指标。适合入门阅读。推荐指数：4星；

• **2021 | Deep Learning for Anomaly Detection: A Review**[4]：作者介绍了异常检测面临的挑战、模型分类、每类模型的优缺点、公开数据集和未来可发展方向。还是很全面和详细的。推荐指数：5星；

•**2021 | Revisiting Time Series Outlier Detection: Definitions and Benchmarks** ：作者提出新的异常分类标准和各类异常数据人工合成方法。然后模型测试得出深度网络劣于传统模型。推荐指数：4星。

另外，其他异常检测综述要么是关于图像的异常检测，要么是基于图的异常检测方法总结。这里不是本文旨在分享的范畴，有兴趣的朋友可自行阅读[2]。本文将从以下6个方面介绍：

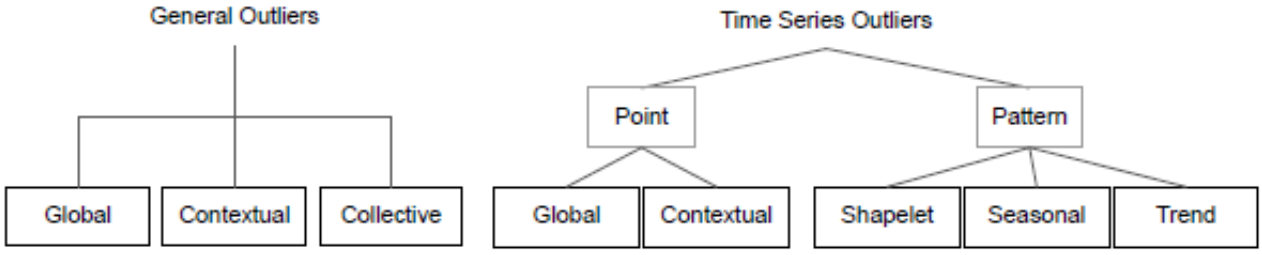
一、异常分类
二、异常检测的挑战
三、异常检测的模型分类

一、异常分类

以前传统关于异常检测的分类如图1(a)所示，分为：

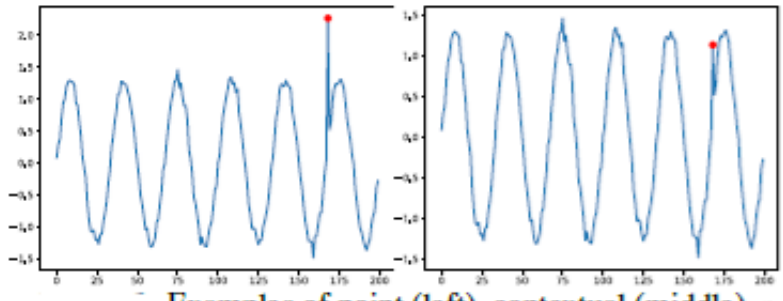
- **点异常值**：相对于全局其他数据的异常实例。
- **上下文异常值**：上下文异常值通常在它们自己的上下文中具有相对较大/较小的值，但不是全局的。
- **集体异常值**：被定义为相对于整个数据集异常的相关异常数据实例的集合。

但这种分类方式常因为上下文定义边界模糊，导致集体异常值和上下文异常值的定义边界也模糊。上下文异常值的上下文在不同文献中通常非常不同。它们可以是一个小窗口，包含相邻点或在季节性方面具有相似相对位置的点。比如图2中的集体异常值，如果以季节性方面的上下文考虑，其实也能看做是上下文异常。

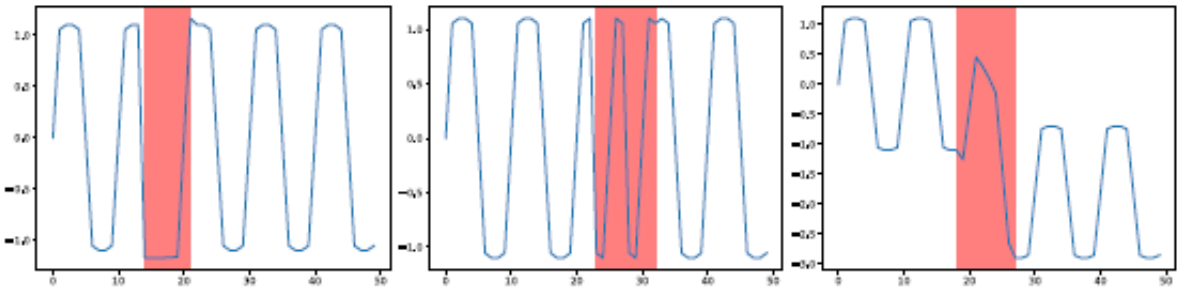


(a) Existing Taxonomy (b) Behavior-Driven Taxonomy

综述　提出了新的异常分类法，如图1(b)所示。具体的样例如下：



Examples of point (left), contextual (middle)



(a) Shapelet Outliers (b) Seasonal Outlier (c) Trend Outlier

关于3类Pattern异常，可以基于shapelet函数来定义：

$$X_{i,j} = \rho(2\pi\omega T_{i,j}) + \tau(T_{i,j})$$

其中， $\rho(2\pi T, \omega) = \sum_n [A\sin(2\pi\omega_n T) + B\cos(2\pi\omega_n T)]$ ，X是由多个不同+频率的波的值相加得到的。 $\tau(\cdot)$ 为趋势项，例如线性函数 $\tau(T) = T$ 。如果s为相似度度量函数，那么以上3种异常类型可以分别定义为：

- **shapelet outliers (异常的局部子序列)**: $s(\rho(.), \hat{\rho}(.)) > \delta$ 。
- **seasonal outliers (异常周期性的局部子序列)**: $s(\omega, \hat{\omega}) > \delta$ 。
- **trend outliers (异常趋势的局部子序列)**: $s(\tau(.), \hat{\tau}(.)) > \delta$ 。

其中， δ 为异常判定的阈值。

二、异常检测的挑战

综述　介绍了深度异常检测解决的主要挑战：

- **CH1：异常检测召回率低。**由于异常非常罕见且异质，因此很难识别所有异常。
- **CH2：**异常通常在低维空间中表现出明显的异常特征，而在**高维空间中变得隐藏且不明显。**
- **CH3：正常/异常的数据高效学习。**利用标记数据来学习正常/异常的表征，对于准确的异常检测至关重要。
- **CH4：抗噪异常检测。**许多弱/半监督异常检测方法假设标记的训练数据是干净的，这可能容

易受到被错误标记为相反类别标签的噪声实例的影响。

• **CH5：复杂异常的检测。**现有的大多数方法都是针对点异常的，不能用于条件异常和组异常，因为它们表现出与点异常完全不同的行为。

• **CH6：异常解释。**在许多安全关键领域中，如果将异常检测模型直接用作黑盒模型，则可能存在一些重大风险。

图3展示了传统方法和深度方法在不同能力上的区别，以及不同能力对解决哪些挑战至关重要：

Method	End-to-end Optimization	Tailored Representation Learning	Intricate Relation Learning	Heterogeneity Handling
Traditional	×	×	Weak	Weak
Deep	✓	✓	Strong	Strong
Challenges	CH1-6	CH1-6	CH1, CH2, CH3, CH5	CH3, CH5

具体到模型上的挑战，会在下面进行详细讲解。

三、异常检测的模型分类

不同综述对时序异常检测的模型分类方式也挺不同的，比如：

• 综述　：分为统计方法，经典机器学习方法和使用神经网络的异常检测方法。

• 综述　：分为基于预测偏差的方法，基于时序表征分类的方法，基于子序列不一致性分析的方法。

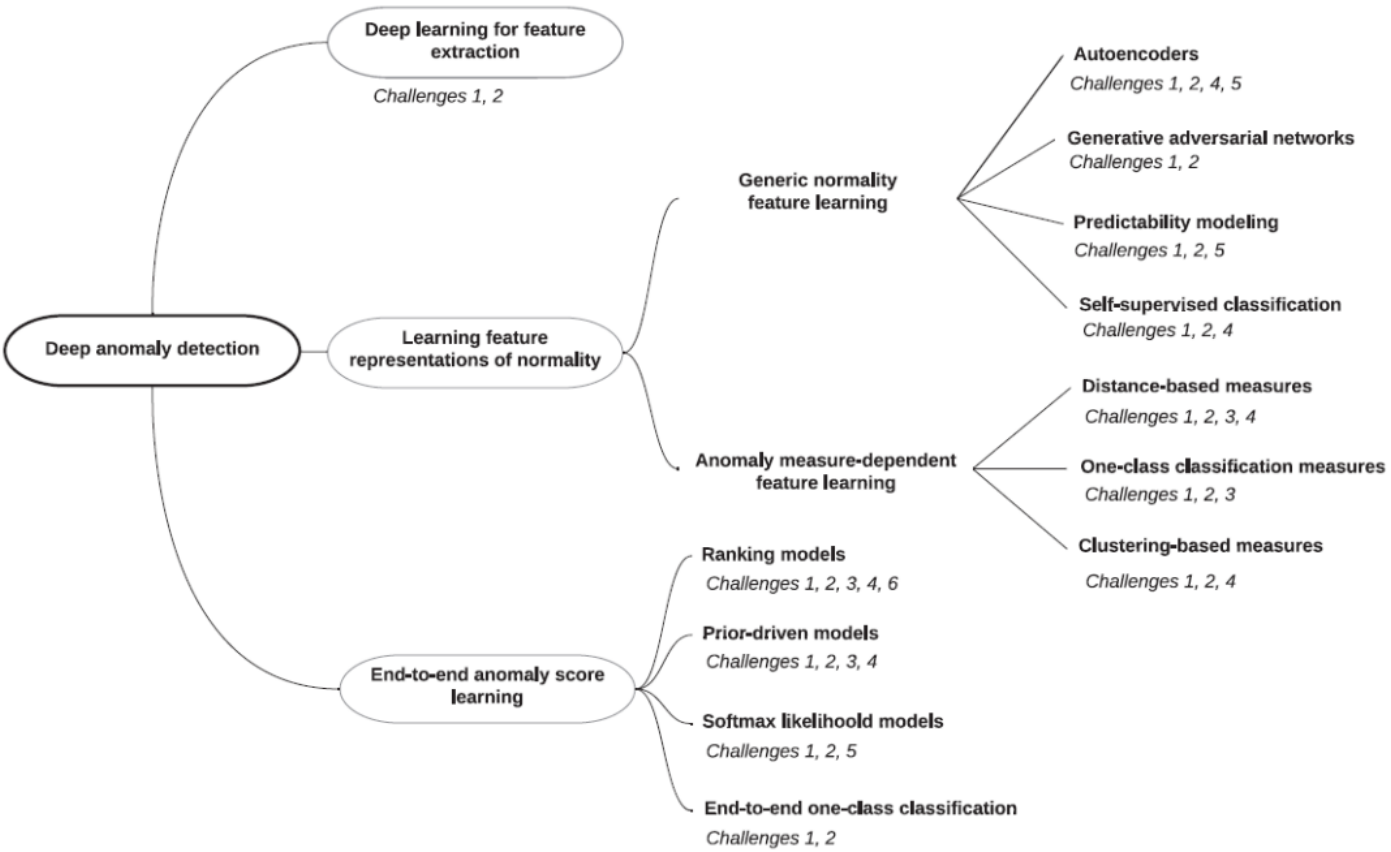
• 综述　：针对神经网络算法，分为特征提取的方法，学习常态特征表征的方法，端对端学习异常分数的方法。

我总结如下：

在之前的文章里，统计方法和经典机器学习的方法基本都已经介绍过了，这边就不重复介绍了。

方法类别	方法名称	异常标准
基于分布	3sigma	值>mu+3sigma或值<mu-3sigma
	Z-score	z_score>3
	boxplot	值>q3+1.5IQR或值<q1-1.5IQR
	Grubbs假设检验	z-score>Grubbs临界值
基于距离	KNN	K近邻平均距离>阈值
基于密度	LOF	LOF>阈值
	COF	COF>阈值
	SOS	异常概率>阈值
基于聚类	DBSCAN	无法被聚类成簇（label=-1）
基于树	iForest	异常得分>阈值
基于降维	PCA	低维空间在所有方向上的偏离程度>阈值
	AutoEncoder	重建样本和原始样本的误差>阈值
基于分类	One-Class SVM	超平面外部的点（label=-1）
基于预测	Moving Average、ARIMA等	预测值和真实值的残差+基于分布的方法

这里，主要基于综述 ，介绍下神经网络下的模型分类，其实zero在知乎已经整理了这篇综述内容，写的很好，强烈建议阅读文章 。



神经网络下的模型分类如下：

1. 特征提取：deep learning和anomaly detection是分开的，deep learning只负责特征提取。

2. 常态特征表征学习：deep learning和anomaly detection是相互依赖的，一起学习正常样本的有效表征。

• **通用常态特征表征学习：**这类方法最优化一个特征学习目标函数，该函数不是为异常检测而设计的，但学习到的高级特征能够用于异常检测，因为这些高级特征包含了数据的隐藏规律。

• **依赖异常度量的特征表征学习：**该类方法直接将现有的异常评价指标嵌入表征学习的优化目标中。

3. 端对端异常分数学习：deep learning和anomaly detection是完全一体的，通过端到端的学习，直接输出异常分数。

1. 特征提取

旨在利用深度学习从高维和/或非线性可分离数据中提取低维特征表征，用于下游异常检测。特征提取和异常评分完全不相交且彼此独立。因此，深度学习组件仅作为降维工作。

优点：

• 很容易获得大量先进的预训练深度模型和现成的异常检测器做特征提取和异常检测；

• 深度特征提取比传统线性方法更有效。

缺点：

• 特征提取和异常评分是独立分开的，通常会导致次优的异常评分；

• 预训练的深度模型通常仅限于特定类型的数据。（感觉更适用于图像，因为图像可以做分类预训练，个人对时序预训练了解的不是很多）。

2. 通用常态特征表征学习

这类方法最优化一个特征学习目标函数，该函数不是为异常检测而设计的，但学习到的高级特征能够用于异常检测，因为这些高级特征包含了数据的隐藏规律。例如：AutoEncoder、GAN、预测模型。

优点：

• AE：方法简单，可用不同AE变种；

• GAN：产生正常样本的能力很强，而产生异常样本的能力就很弱，因此有利于进行异常检

测；

• 预测模型：存在大量序列预测模型，能学到时间和空间的依赖性。

缺点：

• AE：学习到的特征表征可能会因为“训练数据中不常见的规律、异常值或噪声“而产生偏差；

• GAN：训练可能存在多种问题，比如难以收敛，模式坍塌。因此，基于异常检测的 GANs 训练或难以进行；

• 预测模型：序列预测的计算成本高。

另外，以上方法都有两个共性问题：

• **都假设训练集是正常样本**，但若训练集中混入噪声或异常值，会给模型表征学习能力带来偏差；

• **没有将异常评价纳入到模型优化的目标当中**，最后检测的结果可能是次优的。

3. 依赖异常度量的特征表征学习

该类方法直接将现有的异常评价指标嵌入表征学习的优化目标中，解决了通用常态特征表征学习中第二个共性问题。例如Deep one-class SVM，Deep one-class Support Vector Data Description (Deep one-class SVDD)等。

优化：

• 基于距离的度量：比起传统方法，能处理高维空间数据，有丰富的理论支持；

• 基于one-class分类的度量：表征学习和one-class模型能一起学习更好的特征表示，同时免于手动选择核函数；

• 基于聚类的度量：对于复杂数据，可以让聚类方法在深度专门优化后的表征空间内检测异常点。

缺点：

• 基于距离的度量：计算量大；

• 基于one-class分类的度量：在正常类内分布复杂的数据集上，该模型可能会无效；

• 基于聚类的度量：模型的表现严重依赖于聚类结果。也受污染数据的影响。

以上缺点在于：**没办法直接输出异常分数。**

3. 端对端异常分数学习

通过端到端的学习，直接输出异常分数。个人对这部分的了解是一片空白，只能初略转述下综述中的内容，有兴趣的朋友可以阅读原文跟进相关工作。

优点：

- 排名模型：利用了排序理论；
- 先验驱动模型：将不同的先验分布嵌入到模型中，并提供更多解释性；
- Softmax似然模型：可以捕捉异常的特征交互信息；
- 端到端的one-class分类模型：端到端式的对抗式优化，GAN有丰富的理论和实践支持。

缺点：

- 排名模型：训练数据中必须要有异常样本；
- 先验驱动模型：没法设计一个普遍有效的先验，若先验分布不能很好地拟合真实分布，模型的效果可能会变差；
- Softmax似然模型：特征交互的计算成本很大，而且模型依赖负样本的质量；
- 端到端的one-class分类模型：GAN具有不稳定性，且仅限于半监督异常检测场景。

深度相关的30个代表性模型：

Method	Ref.	Sup.	Objective	DA	DP	PT	Archit.	Activation	# layers	Loss	Data
OADA	[65] (4)	Semi	Reconstruction	Yes	No	No	AE	ReLU	3	MSE	Video
Replicator	[57] (5.1.1)	Unsup.	Reconstruction	No	No	No	AE	Tanh	2	MSE	Tabular
RandNet	[29] (5.1.1)	Unsup.	Reconstruction	No	Yes	Yes	AE	ReLU	3	MSE	Tabular
RDA	[175] (5.1.1)	Semi	Reconstruction	No	No	No	AE	Sigmoid	2	MSE	Tabular
UODA	[91] (5.1.1)	Semi	Reconstruction	No	No	Yes	AE & RNN	Sigmoid	4	MSE	Sequence
AnoGAN	[138] (5.1.2)	Semi	Generative	No	No	No	Conv.	ReLU	4	MAE	Image
EBGAN	[170] (5.1.2)	Semi	Generative	No	No	No	Conv. & MLP	ReLU/lReLU	3-4	GAN	Image & Tabular
FFP	[86] (5.1.3)	Semi	Predictive	Yes	No	Yes	Conv.	ReLU	10	MAE/MSE	Video
LSA	[1] (5.1.3)	Semi	Predictive	No	No	No	Conv.	lReLU	4-7	MSE & KL	video
GT	[48] (5.1.4)	Semi	Classification	Yes	Yes	No	Conv.	ReLU	10-16	CE	Image
E^3Outlier	[157] (5.1.4)	Semi	Classification	Yes	Yes	No	Conv.	ReLU	10	CE	Image
REPEN	[112] (5.2.1)	Unsup.	Distance	No	No	No	MLP	ReLU	1	Hinge	Tabular
RDP	[155] (5.2.1)	Unsup.	Distance	No	No	No	MLP	lReLU	1	MSE	Tabular
AE-1SVM	[104] (5.2.2)	Unsup.	One-class	No	No	No	AE & Conv.	Sigmoid	2-5	Hinge	Tabular & image
DeepOC	[161] (5.2.2)	Semi	One-class	No	No	No	3D Conv.	ReLU	5	Hinge	Video
Deep SVDD	[132] (5.2.2)	Semi	One-class	No	No	Yes	Conv.	lReLU	3-4	Hinge	Image
Deep SAD	[133] (5.2.2)	Semi	One-class	No	No	Yes	Conv. & MLP	lReLU	3-4	Hinge	Image & Tabular
DEC	[162] (5.2.3)	Unsup.	Clustering	No	Yes	Yes	MLP	ReLU	4	KL	Image & Tabular
DAGMM	[179] (5.2.3)	Unsup.	Clustering	No	Yes	No	AE & MLP	Tanh	4-6	Likelihood	Tabular
SDOR	[117] (6.1)	Unsup.	Anomaly scores	No	No	Yes	ResNet & MLP	ReLU	50 + 2	MAE	Video
PReNet	[114] (6.1)	Weak	Anomaly scores	Yes	No	No	MLP	ReLU	2-4	MAE	Tabular
MIL	[145] (6.1)	Weak	Anomaly scores	No	Yes	Yes	3DConv. & MLP	ReLU	18/34 + 3	Hinge	Video
PUP	[107] (6.2)	Unsup.	Anomaly scores	No	No	No	MLP	ReLU	3	Likelihood	Sequence
DevNet	[115] (6.2)	Weak	Anomaly scores	No	No	No	MLP	ReLU	2-4	Deviation	Tabular
APE	[30] (6.3)	Unsup.	Anomaly scores	No	No	No	MLP	Sigmoid	3	Softmax	Tabular
AEHE	[45] (6.3)	Unsup.	Anomaly scores	No	No	No	AE & MLP	ReLU	4	Softmax	Graph
ALOCC	[135] (6.4)	Semi	Anomaly scores	Yes	No	No	AE & CNN	lReLU	5	GANs	Image
OCAN	[174] (6.4)	Semi	Anomaly scores	No	No	Yes	LSTM-AE & MLP	ReLU	4	GANs	Sequence
Fence GAN	[103] (6.4)	Semi	Anomaly scores	No	Yes	No	Conv. & MLP	lReLU/Sigmoid	4-5	GANs	Image & Tabular
OCGAN	[120] (6.4)	Semi	Anomaly scores	No	No	No	Conv.	ReLU/Tanh	3	GANs	Image

DA, DP, PT, and Archit. are short for data augmentation, dropout, pre-training, and architecture, respectively. # layers account for all layers except the input layer. lReLU represents leaky ReLU.

四、异常检测的数据集

SEQ： 中提出基于shapelet函数，我们可以获取35个合成数据集（可称NeurIPS-TS synthestic datasets or SEQ），其中20个单变量，15个多变量数据集。该数据集覆盖各类异常数据。

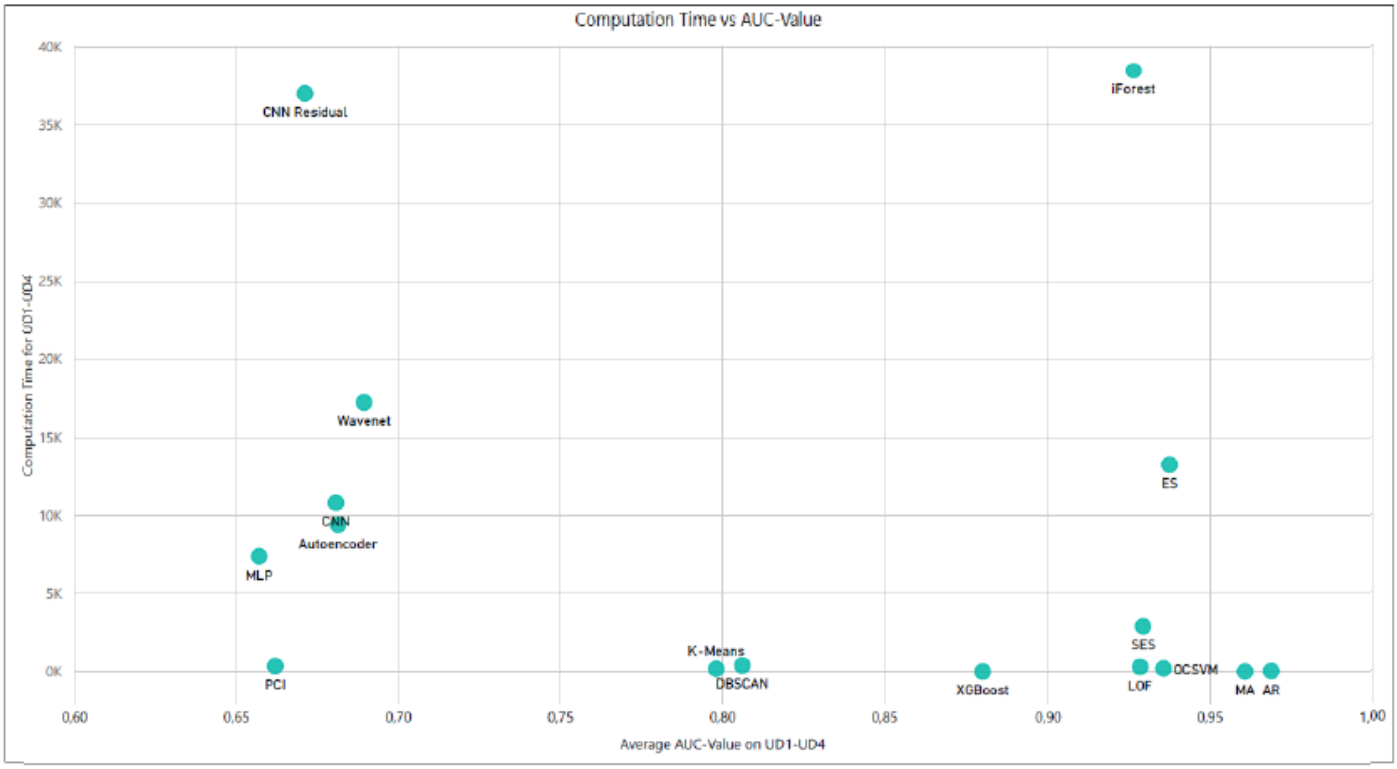
21个开源真实数据集：

Table 3. 21 Publicly Accessible Real-world Datasets with Real Anomalies

Domain	Data	Size	Dimension	Anomaly (%)	Type	Reference
Intrusion detection	KDD Cup 99 [13]	4,091-567,497	41	0.30%-7.70%	Tabular	[57, 103, 104, 179]
Intrusion detection	UNSW-NB15 [100]	257,673	49	≤9.71%	Streaming	[114, 115]
Excitement prediction	KDD Cup 14	619,326	10	6.00%	Tabular	[114, 115]
Dropout prediction	KDD Cup 15	35,091	27	0.10%-0.40%	Sequence	[91]
Malicious URLs detection	URL [93]	2.4m	3.2m	33.04%	Streaming	[112]
Spam detection	Webspam [160]	350,000	16.6m	39.61%	Tabular/text	[112]
Fraud detection	Credit-card-fraud [34]	284,807	30	0.17%	Streaming	[114, 115, 174]
Vandal detection	UMDWikipedia [76]	34,210	N/A	50.00%	Sequence	[174]
Mutant activity detection	p53 Mutants [13]	16,772	5,408	0.48%	Tabular	[112]
Internet ads detection	AD [13]	3,279	1,555	14.00%	Tabular	[112]
Disease detection	Thyroid [13]	7,200	21	7.40%	Tabular	[114, 115, 133, 179]
Disease detection	Arrhythmia [13]	452	279	14.60%	Tabular	[116, 133, 179]
Defect detection	MVTec AD	5,354	N/A	35.26%	Image	[15]
Video surveillance	UCSD Ped 1 [81]	14,000 frames	N/A	28.6%	Video	[117, 161]
Video surveillance	UCSD Ped 2 [81]	4,560 frames	N/A	35.9%	Video	[117, 161]
Video surveillance	UMN [106]	7,739 frames	N/A	15.5%-18.1%	Video	[117]
Video surveillance	Avenue [90]	30,652 frames	N/A	12.46%	Video	[161]
Video surveillance	ShanghaiTech Campus	317,398 frames	N/A	5.38%	Video	[86]
Video surveillance	UCF-Crime	1,900 videos (13.8m frames)	N/A	13 crimes	Video	[145]
System log analysis	HDFS Log [164]	11.2m	N/A	2.90%	Sequence	[40]
System log analysis	OpenStack log	1.3m	N/A	7.00%	Sequence	[40]

五、异常检测的模型表现对比

各类综述论文下的模型表现，因为所用数据集，参数或后处理不一致，导致表现对比可能存在差异，这里仅供参考。



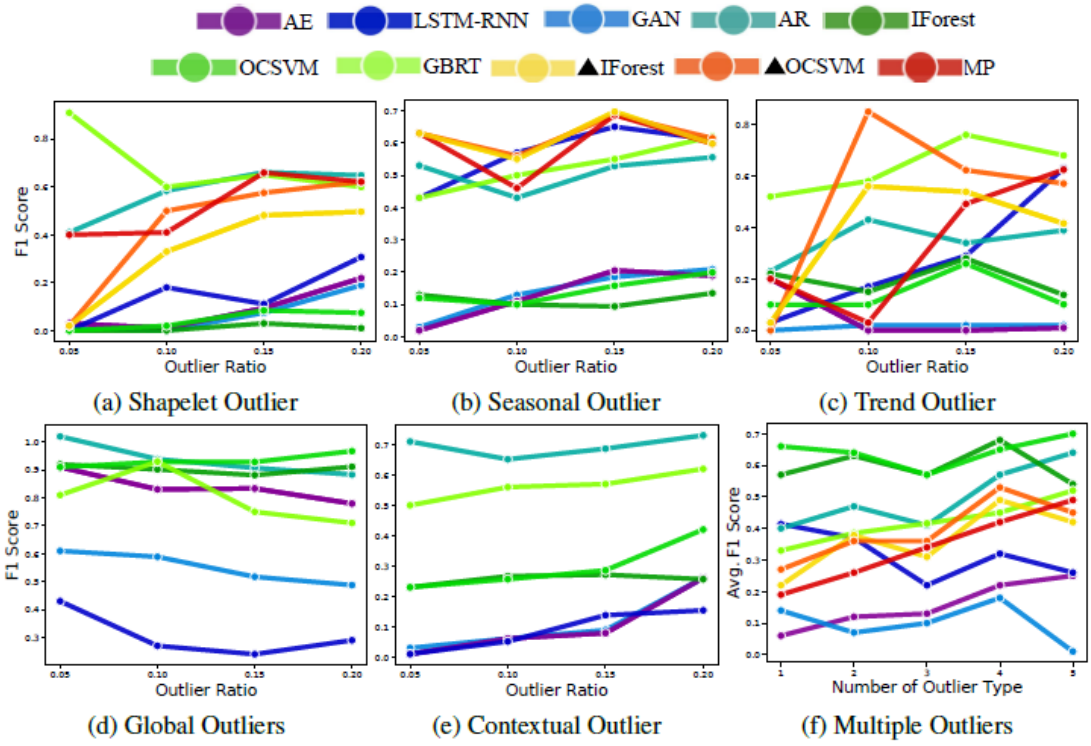


图10很明显打脸了一众深度模型，我们不可否认实际业务场景中，深度模型性能的不稳定，传统模型确实更好用，但在学术圈里，深度模型还是有它研究价值在，而有些深度异常检测论文

的F1分数比图10高，除了参数和数据问题，也可能是像Anomaly Transformer代码Issue中很多人提到的类似“detection adjustment”后处理优化的结果。所以这块仁者见仁智者见智吧。

六、结论和未来方向

综述[4]给出了未来异常检测的结论和发展方向：

- **把异常度量目标加入到表征学习中**：表征学习时，一个关键问题是它们的目标函数是通用的，但没有专门针对异常检测进行优化。在前面有提到依赖于异常度量的特征学习，它便是通过施加来自传统异常度量的约束，来帮助解决这个问题；
- **探索少标记样本的利用**：探索利用这些小标记数据来学习更强大的检测模型和更深层次架构；
- **大规模无监督/自监督表示学习**：首先在无监督/自监督模式下从大规模未标记数据中学习可迁移的预训练表示模型，然后在半监督模式下微调异常检测模型；
- **复杂异常的深度检测**：对条件/组异常的深度模型的探索明显较少。另外多模态异常检测是一个很大程度上尚未探索的研究领域；
- **可解释和可操作的深度异常检测**：具有提供异常解释的内在能力的深度模型很重要，能减轻对人类用户的任何潜在偏见/风险以及实现决策行动；
- **新颖的应用和设置**：例如分布外 (OOD) 检测、curiosity learning等。

个人来看，在【三、异常检测的模型分类】里谈论的模型中，我们可以相互借鉴，比如Anomaly Transformer便采取了依赖异常度量的特征表征学习，同时还借鉴了端对端异常分数学习中的先验驱动模型，引入了先验关联。当我们带着各类模型优缺点的基础知识去阅读新论文时，也能引发思考，比如Anomaly Transformer的先验关联采用高斯分布是否普遍有效？若窗口内存在离散异常尖峰（即多峰异常），那单峰先验关联和多峰序列关联是否便存在一定关联差异，那么检测效果是不是会有负面影响？多头Anomaly Attention是否可以缓解这个问题？这有些跑题了，但希望本篇文章大家能带来些反思和启发，鼓励阅读综述原文，深入了解其中的思想。

参考资料

[1] Python Outlier Detection (PyOD) - 赵越，Github: https://github.com/yzhao062/Pyod。
[2] Anomaly Detection Learning Resources - 赵越，Github: https://github.com/yzhao062/anomaly-detection-resources。
[3] Braei, M., & Wagner, S. (2020). Anomaly detection in univariate time-series: A survey on the state-of-the-art. *arXiv preprint arXiv:2004.00433*.
[4] Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, *54*(2), 1-38.

[5] Lai, K. H., Zha, D., Xu, J., Zhao, Y., Wang, G., & Hu, X. (2021, June). Revisiting time series outlier detection: Definitions and benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
[6] 异常检测综述：Deep Learning for Anomaly Detection: A Review - zero - 知乎：https://zhuanlan.zhihu.com/p/419161328

Scan the QR code to add me on WeChat

Ai

现居 | 广东深圳
现岗位 | 机器学习算法工程师
兴趣 | 数据挖掘、机器学习、深度学习
微信 | ahf1996
公众号 | zaicode
微信社群 | **加入宅码社群**
Github | https://github.com/AlvinAi96
博客 | https://www.cnblogs.com/alvinai
知乎 | https://www.zhihu.com/people/aibyai

宅码
print("Hello, world!")
45篇原创内容

公众号

喜欢此内容的人还喜欢

【知出乎争】tsfresh使用小结
宅码