

基于 LSTM-GAN 的加油时序数据异常检测

赵颢 李晓 马博 王保全 周喜

(中国科学院新疆理化技术研究所 新疆 乌鲁木齐 830011)

(中国科学院大学 北京 100049)

(新疆理化技术研究所新疆民族语音语言信息处理实验室 新疆 乌鲁木齐 830011)

摘要 加油站时序数据中蕴藏着大量信息,但加油数据庞大复杂且各变量间蕴含隐性关系,目前的异常检测方法经常产生“假异常”或遗漏真实异常,无法有效挖掘加油数据中的异常点。针对这种现象,提出一种基于无监督学习的异常检测方法。通过 PCA 对数据进行降维,提取有效的特征信息,采用基于 LSTM 的 GAN 模型 (LSTM-GAN) 对加油时序数据进行检测,通过生成器和鉴别器共同得到的异常损失定义异常点。通过在加油数据集和公开数据集上的实验证明了该方法的有效性,并且相较于目前的异常检测方法具有一定提升。

关键词 生成式对抗网络 长短期递归神经网络 无监督学习 时序数据 异常检测

中图分类号 TP391.1

文献标志码 A

DOI:10.3969/j.issn.1000-386x.2022.07.003

ANOMALY DETECTION OF REFUELING TIME SERIES DATA BASED ON LSTM-GAN

Zhao Yang Li Xiao Ma Bo Wang Baoquan Zhou Xi

(The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, Xinjiang, China)

(University of Chinese Academy of Sciences, Beijing 100049, China)

(Xinjiang Laboratory of Minority Speech and Language Information Processing, The Xinjiang Technical Institute of Physics and Chemistry, Urumqi 830011, Xinjiang, China)

Abstract There is a lot of information in the refueling time series data. The refueling data is huge and complex, and there are hidden relationships among variables. The current anomaly detection methods often find "false anomalies" or miss real anomalies, which can not effectively mine anomalies in the refueling data. In view of this problem, this paper proposes an anomaly detection method based on unsupervised learning. PCA was used to reduce the dimension of the data and extract the effective feature information. We adopted the GAN model based on LSTM (LSTM-GAN) to detect the refueling time series data. The abnormal point was defined by the abnormal loss obtained by the generator and the discriminator. The experiments were carried out on the refueling data set and the public data set. The results prove the effectiveness of the proposed method and show that the method has a certain improvement compared with the current ones.

Keywords Generative adversarial networks (GAN) Long and short term recurrent neural network Unsupervised learning Time series data Anomaly detection

0 引言

随着物联网技术的普及,种类丰富的信息传感器

已经在人们的生活中随处可见,加油站中的信息采集也由人工记录全面转化为电子传感设备自动录入,使得数据采集更加高效。但由于系统故障错误或人为操作失误等因素,会产生异常加油数据。如何高效、准确

收稿日期:2020-03-01。中国科学院西部之光项目(2019-XBQNXZ-A-004);自治区天山青年计划项目(2018Q032);中科院创新青年促进会资助项目(Y9290802);中科院科技服务网络计划(STS计划)项目(KFJ-STQYD-102)。赵颢,硕士生,主研领域:大数据分析,数据挖掘。李晓,研究员。马博,副研究员。王保全,博士生。周喜,研究员。

地在海量加油数据中检测异常,成为加油站日常运营管理中不可忽视的问题。传统上,基于规则的异常检测方法是检测异常数据的常用解决方案,但由于加油数据体量巨大且拥有多变量的时序特性,无法简单通过基于规则的方法来准确判别异常。机器学习方法也逐步运用在异常检测中并取得了一定成果,但想要在加油站时序数据中检测异常,仍然是一项挑战。因为有监督机器学习方法需要大量真实的正常数据和带有标注的异常数据来训练学习,但实际中加油数据缺乏标记且异常数据较为稀少;而无监督学习大多是将样本数据进行线性变换和投影后嵌入向量空间,通过比较原始数据和预测数据的差异值来检测异常,但实际加油数据虽然数量较多,但个体加油样本普遍较少,且复杂的加油时序序列数据隐含的内在相关性往往是非线性的。

为了解决这一问题,本文针对加油时序数据异常检测提出一种基于生成式对抗网络^[1](Generative Adversarial Network, GAN)的无监督异常检测方法,该方法通过捕捉加油站时序数据之间的非线性相关性来建立模型,并结合模型中的生成器和鉴别器共同来进行异常检测。

1 相关工作

目前关于无监督学习的异常检测算法主要可以分为四类:(1)基于线性模型的方法;(2)基于距离的方法;(3)基于概率和密度估计的方法;(4)基于统计机器学习的方法。

基于线性模型的无监督异常检测方法中,一种主流的方法是主成分分析(Principal Component Analysis, PCA)^[2],它对数据进行降维后嵌入到低维空间,通过比较单个数据样本跟整体数据样本的偏离程度来挖掘异常;另一种方法是偏最小二乘法(Partial Least Squares, PLS)^[3],它不仅以克服共线性问题,还去除了对回归无益的噪声影响,因此被广泛应用于模型建立和异常检测。然而,它们只对具有高度相关性的数据有效^[4],加油数据中个体和个体之间的加油行为却往往是相互独立的,因此并不适用于加油时序数据。

基于距离的异常检测方法中^[5],K近邻(K-Nearest-Neighbor, KNN)算法是比较简单有效的一种,它通过参考特征空间中距离该样本最近的 k 个“邻居”的类别来对样本进行分类,从而达到检测异常点的目的^[6]。此外还有基于密度、基于聚类的一些衍生方法,例如局部异常因子(Local Outlier Factor, LOF)算法^[7],它会给

数据中的每个点计算离群因子,通过判断这个离群因子是否接近于1来判定该点是否离群。以及对前者进行改进的基于聚类的局部离群因子(Cluster-Based Local Outlier Factor, CBLOF)^[8]方法,它使用一个预定义的异常评分函数来识别异常。虽然这些基于距离的方法在某些情况下是有效的,但在加油时序数据中,数据间的距离难以界定,且正常数据可能没有足够的样本邻居,因此基于距离的异常检测方法不适用于加油时序数据。

基于概率和密度估计的方法中,有基于角度的异常检测(Angle-Based Outlier Detection, ABOD)算法^[9],还有特征装袋(Feature Bagging, FB)算法^[10],它们都是通过考虑变量之间的相关性来处理数据。但是这些方法不能考虑时序数据中时间步长上的相关性,因而并不适合用来处理加油时序数据。

还有基于统计机器学习的方法,通过神经网络具备的强特征提取能力和信息记忆能力来捕捉正常数据中的异常点,提出很多有效的模型,并应用在了各个领域。比较有代表性的有变分自编码器(Variational Auto-Encoder, VAE)^[11]和循环神经网络(Recurrent Neural Network, RNN)^[12]等,它们都是利用正常数据训练神经网络,然后通过检测实际数据的重构误差来进行异常检测。

综上,通过神经网络来进行机器学习,捕获数据特征间隐藏的相关性,建立符合实际需求的模型,能够更好地开展对加油时序数据的异常检测工作。

本文继承了基于无监督学习的异常检测算法思想,并受到GAN模型博弈原理^[13]的启发后,提出一种新的异常检测方法。主要工作如下:(1)提出一种基于GAN的无监督异常检测方法,用于检测加油时序数据中的异常;(2)模型结构采用了广泛应用在图像开发中的GAN框架,将长短期循环神经网络(LSTM)嵌入GAN框架作为生成器和鉴别器^[14]来分析加油时序数据,捕捉其时间相关性;(3)综合鉴别器的判别结果和生成器的数据重构残差来对每一个测试样本进行异常评分。

2 异常检测模型

2.1 GAN模型

近年来提出的生成式对抗网络(GAN)框架能够通过博弈训练生成模型,它以对抗的方式同时训练生成器和鉴别器:生成器(generator, G)的思想是将一个噪声包装成一个逼真的样本,判别器(discriminator, D)

则需要判断送入的样本是否真实,即通过对抗训练共同进步,判别器 D 对样本的判别能力不断上升,生成器 G 生成样本能力也不断上升^[15]。目前 GAN 主要应用在一些生成图像的工作中,并取得了一定成果,例如在医学领域可以通过 GAN 模型来对肿瘤 CT 图像进行语义分割与检测^[16];或是在影像方面进行视频修复^[17],但是目前在时间序列数据中采用 GAN 框架的工作还比较少,且加油时序数据还要求同时处理多个具有潜在关系的变量。

为了处理加油的时间序列数据,将 GAN 的生成器和鉴别器构造为两个长短期递归神经网络(LSTM)。首先将加油数据通过 PCA 进行降维后输入至 GAN 模型,在添加来自潜在空间的随机噪声后将序列输入至生成器中,生成器根据输入生成序列样本传递给鉴别器,鉴别器将尝试区分生成的序列样本和实际的训练样本,并根据鉴别结果更新自身和生成器的参数,这样可以训练鉴别器尽可能地真序列和假序列分配正确的标签。在经过足够多轮的迭代后,生成器将被训练得能生成足够逼真的样本,并且鉴别器也能够以较高精度区分真(正常)数据和假(异常)数据。在加油时序数据异常检测工作中,通过计算实际测试样本和生成测试样本之间的差异来得到数据的重构损失以及鉴别器对实际测试样本的判别损失,综合这两部分来检测数据中的潜在异常。本文方法的总体流程如图 1 所示。

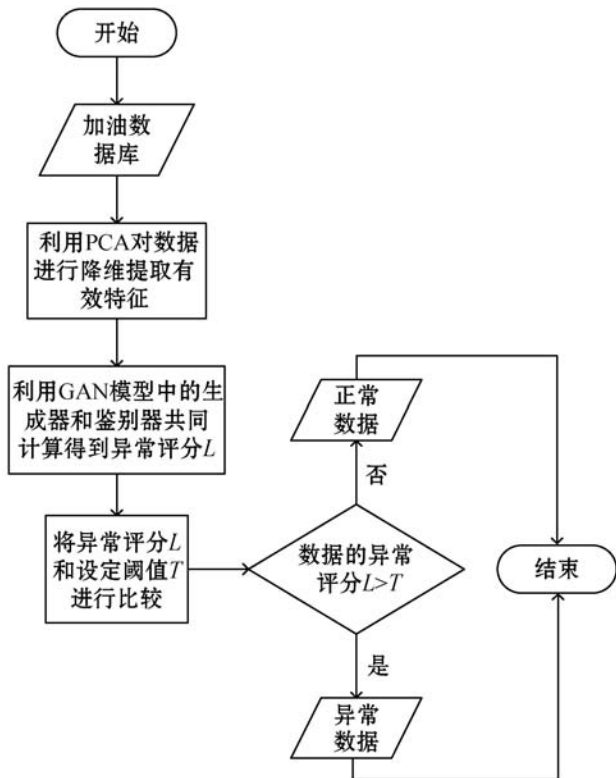


图 1 GAN 模型流程

2.2 LSTM 神经网络

长短期记忆循环神经网络(LSTM)模型由 Hochreiter 等^[18]在 1997 年提出,是对循环神经网络(RNN)的一种改进。RNN 的隐藏层只有一个状态,它对短时间内的输入较为敏感,而 LSTM 添加了一个内部记忆单元 c 来保存长期状态,并同时增加 3 个门用于控制^[19],分别是遗忘门 f_t 、输入门 i_t 和输出门 o_t 。遗忘门 f_t 决定上一层隐藏层输出被遗忘的程度大小:

$$f_t = \delta(W_f \times x_t + U_f \times h_{t-1} + V_f \times C_{t-1} + b_f) \quad (1)$$

式中: C 表示 LSTM 中的细胞状态。

输入门 i_t 控制输入和当前计算的状态更新到记忆单元的程度大小:

$$i_t = \delta(W_i \times x_t + U_i h_{t-1} + V_i C_{t-1} + b_i) \quad (2)$$

输出门 o_t 控制输入和当前输出取决于当前记忆单元的程度大小:

$$o_t = \delta(W_o \times x_t + U_o \times h_{t-1} + V_o \times C_t + b_o) \quad (3)$$

通过这三个控制门,记忆单元 c 即可有效记忆历史输入值,从而增加了模型的特征包含能力和记忆能力,其在 t 时刻的状态即可表示为:

$$c_t = f_t \times c_{t-1} + i_t \times \tanh(W_c \times x_t + U_c \times h_{t-1} + b_c) \quad (4)$$

所以 LSTM 单元在 t 时间点的所有输出为:

$$h_t = o_t \times \tanh(c_t) \quad (5)$$

式中: W, U, V, b 分别是系数组成的矩阵和偏置的向量; δ 是激活函数 Sigmoid。

2.3 基于 LSTM-GAN 的异常检测

可以将异常检测问题定义为:给定一个训练数据集 $x \subseteq D_{\text{train}}^{n \times l}$,即训练集共有 n 条数据,每条数据有 l 个特征值,以及一个测试数据集 $y \subseteq D_{\text{test}}^{m \times l}$,即训练集共有 m 条数据,每条数据有 l 个特征值,异常检测就是为测试集中的每一条数据分配是否异常的标签(0 代表正常,1 代表异常)。这里设定训练集中的每一条数据都是正常的。

通过数据预处理以后,将训练集 $D_{\text{train}} = (x_i, i = 1, 2, \dots, n)$ 以及对其添加随机噪声后生成的样本 $Z = (z_i, i = 1, 2, \dots, n)$ 输入 GAN 模型,对抗训练可以看成是采用生成器与判别器之间的一个极小极大化策略^[20],目标函数如下:

$$\min \max L(D, G) = \mathcal{E}_{x \sim p_{\text{data}}(D)} [\log D(x)] + \mathcal{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (6)$$

式中: \mathcal{E} 指代取期望; $D(x)$ 为判别映射函数; $G(z)$ 为生成映射函数; P_{data} 为真实数据分布; P_z 为原始噪声的分布。

经过足够多轮的迭代训练后模型达到拟合,训练好的鉴别器(D)和生成器(G)即可用于检测测试集中的异常。模型训练过程如图 2 所示。

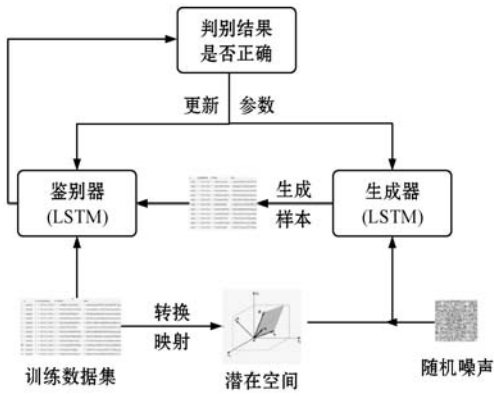


图2 GAN模型训练过程

这里通过加权计算鉴别器(D)的鉴别损失 L_D 和生成器(G)的重构残差 L_G 两部分得到异常损失 L ,通过比较 L 的交叉熵 H 和预定值 T 来判断是否为异常点。并将每一条数据标记如下:

$$y_i = \begin{cases} 1 & H(L_i, 1) > T \\ 0 & \text{其他} \end{cases} \quad i = 1, 2, \dots, m, y \subseteq D_{\text{test}} \quad (7)$$

式中: L_i 代表测试集中第 i 条数据 y_i 的异常损失。

2.4 异常损失

在GAN模型训练达到拟合后,将训练好的鉴别器和生成器共同用来检测异常,如图3所示的测试过程中,将两部分都应用于异常评判中能够更好地来识别异常,所以基于GAN的异常检测由两部分构成。

(1) 基于鉴别器的异常检测:训练后的鉴别器能够以较高灵敏度将“假数据”(即异常数据)与真实数据区分开来,可以直接用来进行异常检测。

(2) 基于生成器的异常检测:训练后的生成器能够生成足够逼真的数据样本,它实际上是一种从潜在空间到真实数据的映射: $G(Z):Z \rightarrow X$ 。如果潜在空间中的输入十分接近,则生成器可输出相似的样本。因此,如果能够在测试集 D_{test} 的潜在空间中找到对应的映射,那么测试集 D_{test} 和通过生成器重构的测试样本 $G(Z)$ 之间的相似性可以在一定程度上反映测试集是否符合生成器所反映的数据分布。这样可以利用测试集和生成样本之间的重构残差 L_G 来识别测试数据中的异常点。

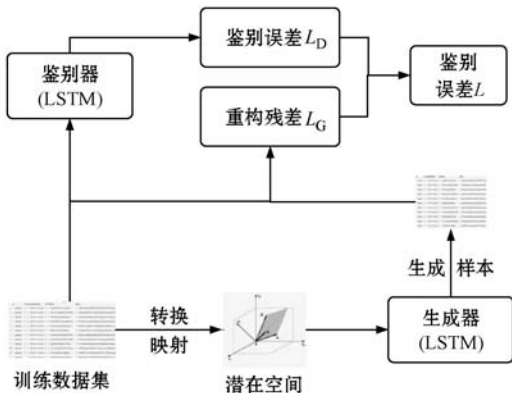


图3 GAN模型测试过程

为了找到与测试样本相对应的最佳 Z_i ,首先从测试集中随机抽取一个集合 Z_1 作为样本,通过潜在空间映射输入生成器,得到重构样本 $G(Z_1)$ 。然后用 D_{test} 和 $G(Z_1)$ 定义的误差函数得到的梯度来更新映射到潜在空间中的样本。为了便于计算,这里通过计算协方差来比较数据样本之间的相似性^[21]。

$$\min \text{Err}(D_{\text{test}}, G(Z_i)) = 1 - \text{Sim}(D_{\text{test}}, G(Z_1)) \quad (8)$$
式中: $\text{Err}(D_{\text{test}}, G(Z_i))$ 为重构样本损失函数; $G(Z_1)$ 为生成映射函数; $\text{Sim}(D_{\text{test}}, G(Z_1))$ 为相似性函数。

经过迭代多轮更新后,误差会达到足够小,此时将样本 Z_i 记录为测试样本在潜在空间中的对应映射。所以得到测试样本的重构残差:

$$L_G(D_{\text{test}}) = \sum_{i=1}^m |y_i - G(Z_i^z)| \quad y_i \subseteq D_{\text{test}} \quad (9)$$

再结合鉴别器的判别损失,可以得到异常检测的损失 L 为:

$$L = \lambda L_G + (1 - \lambda) L_D \quad (10)$$

式中: L_G 为生成器的重构残差; L_D 为鉴别器的判别损失; λ 为影响因子。

3 实验

3.1 实验配置

为了验证本文方法在时序数据异常检测上的准确性和有效性,本文在两个数据集上进行了实验,第一个是在中国某省份各加油站采集的加油数据集,另一个是公开时序数据集UCR^[22]。

实验机系统为Windows 7 64位操作系统,CPU为Intel(R) Core(TM) i7-3770HQ CPU @ 3.40 GHz,内存8 GB,Python版本为3.5,使用的数据库为MongoDB 3.0。

加油数据集采用了某省份所有加油站两个月的加油记录作为原始数据,对其进行数据清洗、数据融合和特征处理等操作后得到真实可靠的实验数据,将其作为实验中的正常数据;异常数据选取历史检测工作中通过预警记录、人工筛查等方式确认的部分异常数据。最后将正常数据的80%作为实验的训练集,另外20%正常数据和异常数据共同作为测试集。

3.2 实验数据集

3.2.1 加油数据集

加油数据集采用了某省份所有加油站两个月的加油记录作为原始数据,对其进行数据清洗、数据融合、特征处理等操作后得到真实可靠的实验数据,将其作

为实验中的正常数据;异常数据选取历史检测工作中通过预警记录、人工筛查等方式确认的部分异常数据。最后将正常数据的 80% 作为实验的训练集,另外 20% 正常数据和异常数据共同作为测试集。

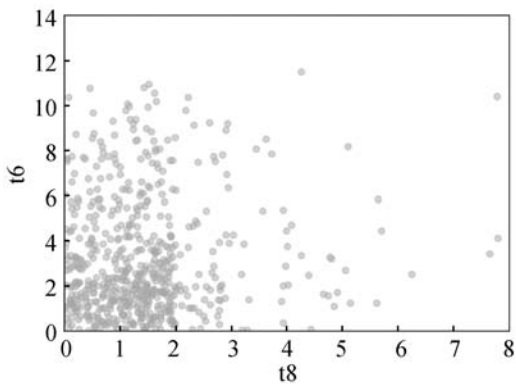
3.2.2 UCR 数据集

UCR 数据集是一个用于时间序列分类预测的公开数据集,其中包含 80 多个用于时间序列分类的数据集,本文从中选取了 Wafer 数据集和 Ford 数据集作为对比实验数据集。Wafer 数据集共有 7 174 条记录,含有类别标签和 152 个特征;Ford 数据集共有 4 921 条记录,含有类别标签和 500 个特征。两个数据集都按照 7:3 的比例划分训练集和测试集。

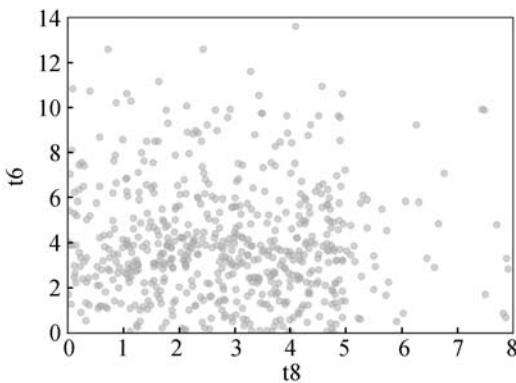
3.3 实验结果及分析

3.3.1 加油数据样本生成

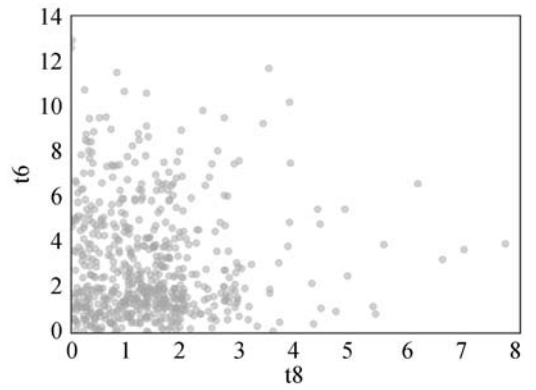
在 GAN 模型的博弈对抗过程中,为了使鉴别器能够敏锐地分辨真假样本,生成器会根据输入尽可能生成逼真的数据样本。为了验证 LSTM-GAN 模型是否可以捕捉到时序数据样本的特征分布,将加油数据集训练过程中初期(第 10 轮)和后期(第 80 轮)生成的数据样本进行可视化,这里选择了 t6 和 t8 两个特征,结果如图 4 所示,其中:图 4(a)为原始数据的特征分布;图 4(b)为训练初期(第 10 轮)生成的样本;图 4(c)为训练后期(第 80 轮)生成的样本。



(a) 真实样本分布



(b) 初期生成样本(第 10 轮)



(c) 后期生成样本(第 80 轮)

图 4 实验结果

可以看出,在训练初期生成器的稳定性较差,产生的数据样本有很强的随机性,但随着迭代轮数的增加,在训练达到第 80 轮时,生成器产生的样本已经基本符合真实样本的数据分布,说明通过多轮训练 LSTM-GAN 模型可以有效学习到加油时序数据样本特征的分

3.3.2 最佳子序列长度分析对比

在实验中,加油数据被处理成序列输入模型中进行训练,而子序列长度会影响到模型训练时间以及预测结果的好坏,从而影响异常检测的结果。选择合适的子序列长度不仅可以提高模型的效率,还能提升模型的检测效果。这里通过尝试一些子序列长度来确定最佳值。具体结果如表 1 所示。

表 1 不同子序列长度结果对比

指标	子序列长度					
	10	30	50	70	90	110
Precision/%	56.70	57.35	76.09	75.59	46.20	55.89
Recall/%	74.25	81.59	81.05	81.01	87.52	87.32
F1	0.61	0.65	0.76	0.75	0.60	0.66

可以看出子序列长度较小时,GAN 模型在加油数据集上的预测效果一般,精确率低于 60%;随着子序列长度提升至 50,预测效果相较于之前较好,精确率和 F1 值都有较大提升;而当子序列长度达到 100 左右时,虽然模型召回率能提升至接近 90%,但精确率开始大幅下滑,整体预测效果较差。所以在模型训练时子序列长度选定为 50,此时模型异常检测的效果最好。

3.3.3 PCA 降维分析对比

如前所述,为了减少模型的计算量,提高模型效率,在输入模型前对加油数据使用 PCA 进行降维处理,剔除次要变量。为了能够尽可能保留原始样本信息,这里计算了 PCA 各保留特征数的方差率,结果如

图 5 所示。

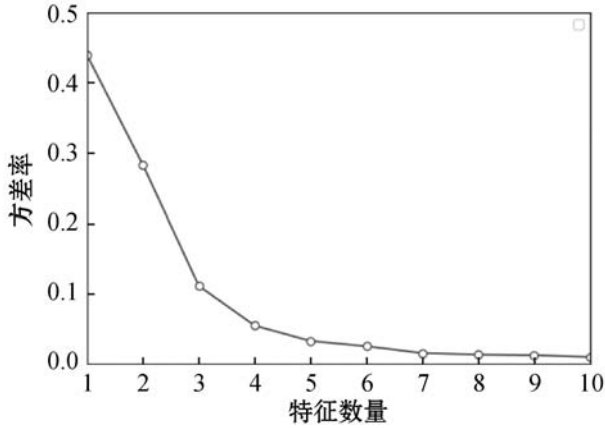


图 5 PCA 方差率

可以看出当通过 PCA 保留 5 个特征时, 占有所有特征方差的百分比已超过 90%, 意味着基本保留了所有信息, 所以确定 PCA 保留特征数为 5。

3.3.4 加油数据集实验

在加油数据集实验中, 通过与已经验证在异常检测领域上有效的 K 近邻 (KNN)、装袋法 (FB)、自动编码器 (AutoEncoder) 和孤立森林 (IsolationForest)^[23] 进行比较评估, 来验证本文算法的性能。为了考量模型检测异常的整体效果, 实验结果使用精确率 (Precision)、召回率 (Recall) 和 F1 值来综合评估 LSTM-GAN 模型的异常检测性能:

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F_1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \quad (13)$$

式中: TP 是实际检测出的异常样本数量 (检测值:1; 真实值:1); FP 是检测为异常但实际非异常的样本数量 (检测值:1; 真实值:0); FN 是未被检测出的异常样本数量 (检测值:0; 真实值:1); TN 是实际检测为正常的样本数量 (检测值:0; 真实值:0); $Precision$ 简写为 Pre ; $Recall$ 简写为 Rec 。

实验结果如表 2 所示。

表 2 加油数据集实验结果对比

模型	Precision/%	Recall/%	F1
KNN	62.79	37.85	0.47
FeatureBagging	83.30	32.07	0.46
AutoEncoder	63.13	72.52	0.68
IForest	92.80	44.79	0.60
LSTM-GAN	76.09	81.05	0.76

实验结果表明相比于传统的异常检测算法, 本文提出的 LSTM-GAN 模型在召回率和 F1 值上都具有较好的效果, 虽然精确率相对于 FeatureBagging 和 IForest 较低, 但是后两者的召回率较差, 无法同时兼顾精确率和召回率。而 LSTM-GAN 模型的精确率和召回率都在 75% 以上, 且从 F1 值上可以看出 LSTM-GAN 模型整体预测效果更好, 说明 LSTM-GAN 模型在加油时序数据上有着较好的检测结果。

3.3.5 UCR 数据集实验

Wafer 数据集上的结果如表 3 所示。

表 3 Wafer 数据集实验结果对比

模型	Precision/%	Recall/%	F1
KNN	71.31	77.32	0.74
FeatureBagging	73.61	77.56	0.76
AutoEncoder	96.53	11.79	0.21
IForest	70.11	92.68	0.80
LSTM-GAN	98.35	78.89	0.86

Ford 数据集上的结果如表 4 所示。

表 4 Ford 数据集实验结果对比

模型	Precision/%	Recall/%	F1
KNN	70.59	60.01	0.65
FeatureBagging	69.57	56.57	0.62
AutoEncoder	86.61	24.35	0.38
IForest	71.33	53.23	0.61
LSTM-GAN	89.92	62.02	0.72

结果表明, 在 UCR 的两个数据集上 LSTM-GAN 模型相比于其他三种方法的整体预测效果较好。在 Wafer 数据集上 IsolationForest 和 LSTM-GAN 相较于其他两种方法效果更好, 且 LSTM-GAN 的 F1 值更高。在 Ford 数据集上由于数据的维数较高且样本数量较少, 整体预测结果都不是很理想, 但 LSTM-GAN 模型还是能够有近 90% 的精确率和超过 60% 的召回率, 说明 LSTM-GAN 模型在时序数据上相比于其他方法有较好的预测结果。

4 结 语

本文鉴于目前在加油时序数据上的无监督异常检测方法存在一定缺陷, 提出一种将 LSTM 嵌入 GAN 模型并通过生成器和鉴别器共同检测数据异常的方法。

法。该方法首先通过 PCA 对数据进行降维处理,提取有效特征后输入 GAN 模型进行训练,待模型拟合后将测试数据输入,通过训练好的生成器和鉴别器共同得到异常损失,与设定阈值进行比较来对时序数据进行异常检测。在加油数据集和 UCR 中两个数据集上的实验结果表明该方法对于现有算法有了一定的改进与提升。但是 LSTM-GAN 模型中的 LSTM 神经网络在处理较长时间序列时需要更多的时间且 GAN 模型的博弈训练过程自由性较高,还不是足够稳定,后续的研究过程中将探索是否能够结合其他神经网络提高模型效率并且能使训练过程更加稳定的异常检测方法。

参 考 文 献

- [1] Ledig C, Theis L, Huszar F, et al. Photo-realistic single image super-resolution using a generative adversarial network [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- [2] 曾建华. 一种基于核 PCA 的网络流量异常检测算法[J]. 计算机应用与软件, 2018, 35(3): 140-144.
- [3] Marko S, Christian R, Joe F H. Partial least squares structural equation modeling[M]. Springer International Publishing, 2017.
- [4] 陈斌, 陈松灿, 潘志松, 等. 异常检测综述[J]. 山东大学学报(工学版), 2009, 39(6): 13-23.
- [5] Huang T, Zhu Y, Wu Y, et al. Anomaly detection and identification scheme for VM live migration in cloud infrastructure[J]. Future Generation Computer Systems, 2016, 56: 736-745.
- [6] Ding X, Zhang Y, Chen L, et al. Distributed k-Nearest neighbor queries in metric spaces: Second international joint conference, APWeb-WAIM 2018, Macau, China, July 23-25, 2018, Proceedings, Part I [M]//Web and Big Data, 2018.
- [7] 冯立伟, 张成, 李元, 等. 基于统计衡量和局部近邻标准化的局部离群因子故障检测方法[J]. 计算机应用, 2018, 38(4): 965-970.
- [8] Ali S, Wang G, Cottrell R L, et al. Detecting anomalies from end-to-end internet performance measurements (PINGER) using cluster based local outlier factor [C]//2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), 2017.
- [9] Chen L, Gao S, Cao X. Research on real-time outlier detection over big data streams[J]. International Journal of Computers & Applications, 2017(8): 1-9.
- [10] Tran C T, Zhang M, Andreae P, et al. Bagging and feature selection for classification with incomplete data [C]//European Conference on the Applications of Evolutionary Computation, 2017.
- [11] Osada G, Omote K, Nishide T. Network intrusion detection based on semi-supervised variational auto-encoder [C]//European Symposium on Research in Computer Security, 2017.
- [12] Shipmon D T, Gurevitch J M, Piselli P M, et al. Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data [EB]. arXiv:1708.03665, 2017.
- [13] Konstantinos Bousmalis K, Silberman N, Dohan D, et al. Unsupervised pixel-level domain adaptation with generative adversarial networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- [14] Li D, Chen D, Jin B, et al. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks [C]//International Conference on Artificial Neural Networks, 2019.
- [15] Dongyang Zhang D, Shao J, Hu G. Sharp and real image super-resolution using generative adversarial network [C]//International Conference on Neural Information Processing, 2017.
- [16] 昌杰. 基于深度神经网络的肿瘤图像分析与处理[D]. 合肥: 中国科学技术大学, 2019.
- [17] 于冰, 丁友东, 谢志峰, 等. 基于时空生成对抗网络的视频修复[J]. 计算机辅助设计与图形学学报, 2020, 32(5): 769-779.
- [18] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
- [19] Karim F, Majumdar S, Darabi H, et al. LSTM fully convolutional networks for time series classification [J]. IEEE Access, 2017, 6: 1662-1669.
- [20] Yang Y, Li Y, Zhang W, et al. Generative-adversarial-network-based wireless channel modeling: Challenges and opportunities [J]. IEEE Communications Magazine, 2019, 57(3): 22-27.
- [21] Dotson J, Howell J R, Brazell J D, et al. A probit model with structured covariance for similarity effects and source of volume calculations [J]. Journal of Marketing Research, 2018, 55(1): 35-47.
- [22] DauChen Y, Keogh E, Hu B, et al. The UCR time series archive [EB/OL]. [2019-11-08]. https://www.cs.ucr.edu/~eamonn/time_series_data.
- [23] Liu F T, Ting K M, Zhou Z H. Isolation forest [C]//2008 Eighth IEEE International Conference on Data Mining, 2008.