DOI:10.19651/j.cnki.emt.2105846

智能电表数据异常在线检测的无监督学习

王 凯 黄 丹 梁晓伟 陶 琳
(国网安徽省电力有限公司营销服务中心 合肥 230088)

摘 要: 针对智能电表在应用过程中出现的数据异常问题,通过定义为超出各负荷正常使用模式的任何异常用电实例或趋势,设计了基于负载、上下文和环境等不同类型的特征来构建数据驱动模型,评估了基于回归、基于神经网络、基于聚类和基于投影的 4 种不同的无监督学习方法在实际智能电表数据异常检测中的性能。结果表明,不同的异常检测方法对不同类型的异常具有不同的检测能力,其性能取决于用于训练该方法的特征集。因此,对于每种异常检测方法,都要仔细检查不同类型的特征。

关键词: 数据驱动;异常检测;智能电表;特征选择;无监督学习
中图分类号: TP393.06;TP18 **文献标识码:** A **国家标准学科分类代码:** 510.4099

Unsupervised learning for on-line detection of abnormal data in smart meters

Wang Kai Huang Dan Liang Xiaowei Tao Lin
(Marketing Service Center of State Grid Anhui Electric Power Co., Ltd., Hefei 230088, China)

Abstract: Aiming at the problem of abnormal data in the application of smart meters, by defining any abnormal power consumption instance or trend beyond the normal use mode of each load, a data-driven model based on different types of characteristics such as load, context and environment is designed, and four different unsupervised models based on regression, neural network, clustering and projection are evaluated the performance of the learning method in the actual smart meter data anomaly detection is analyzed. The results show that different anomaly detection methods have different detection ability for different types of anomalies, and their performance depends on the feature set used to train the method. Therefore, for each anomaly detection method, different types of features should be carefully examined.

Keywords: data driven;anomaly detection;smart meter;feature selection;unsupervised learning

0 引 言

智能电表的部署是实现配电网信息化集成的基础,提取数据中最有用部分并将其转换为可操作信息已成为研究的重点[1]。智能电表数据流中异常检测对负载预测、网络攻击检测、故障断电检测、窃电检测、需求响应等应用具有重要作用[2]。通常,利用机器学习方法对智能电表数据进行处理。文献[3]提出了具有置信度采样的深度半监督卷积神经网络。文献[4]提出了基于滑动窗口的有监督集成方法。然而,当涉及到异常检测时,必须处理固有的无监督学习问题。文献[5]提出了基于动态回归模型和自适应异常阈值的无监督异常检测方法。文献[6]利用基于低维相似矩阵的无监督聚类算法检测异常用电量。然而,文献[5]和文献[6]只考虑特定类型的异常,并没有考虑特征选择在模型训练中的作用。

本文旨在对智能电表进行数据驱动研究,利用真实智能电表数据流识别负载异常情况。对基于回归、基于神经网络、基于聚类和基于投影的 4 种不同的无监督异常检测学习方法进行了系统的比较研究。针对不同方法研究不同特征,进而得到每种方法的最佳特征组合。

1 特征选择

从广义角度,用电数据的特征可基于负载、上下文和环境 3 类特征。对于特定问题和特定数据驱动方法,在每个类别中确定正确的特征选择至关重要。

1.1 负载特征

基于负载的特征反映了不同时间步长条件下居民家庭的用电量,由具有不同时滞的历史用电量数据中获得[7]。

收稿日期:2021-02-26

考虑基于负载的特性可以表示：

$$\begin{cases} L^t = \{P^t, P_Y^t, P_W^t, P_M^t\} \\ L^w = \{P^{t-24}, P^{t-23}, \cdots, P^{t-1}\} \end{cases} \tag{1}$$

其中，L^t 为时间 t 的一组历史负载数据。在这组数据中，P^t 为时间 t 的用电量，P_Y^t 为前一天在时间 t 的用电量，P_W^t 为过去两周在时间 t 的用电量，P_M^t 为时间 t 的用电量平均值，L^w 为前 24 h 的数据集。

1.2　上下文特征

上下文特征并不特定于用电量，但确实对用电量有间接影响。一天中的时间 T_d^t、一周中的某一天 D_w^t、周末与工作日 W_s^t、假日 H^t 和一年中的季节 S^t 都是上下文信息的实例[8]，如式(2)所示。

$$C = \{T_d^t, D_w^t, W_s^t, H^t, S^t\} \tag{2}$$

1.3　环境特征

某些设备(如空调系统)的用电量取决于某些环境特征，如温度。因此，居民家庭总用电量受环境特征的影响[9]。在本文中，环境特征可视为集合 E，由温度($Temp^t$)和湿度(Hum^t)因素组成，如下所示：

$$E = \{Temp^t, Hum^t\} \tag{3}$$

基于负载特征、上下文特征和环境特征可以相互关联，进而影响学习过程的质量。因此，必须研究不同特征组合对每种检测方法的影响，以便针对每个模型定制特征。

2　无监督在线异常检测

异常检测问题本质上是无监督的学习问题。对于智能电表数据，必须探索此类数据流中可能出现的异常类型，以及检测不同异常的潜在应用。为了分析客户的“异常”负载模式，本文没有特定的预先确定的标签。在线无监督学习方法在发现新数据便会更新，因此，在线无监督学习可以学习新模式和趋势变化[10]，例如季节性变化。此外，在线无监督学习方法可以实时实施，以便快速检测异常。在本文中实现了 4 种无监督的在线异常检测方法。

2.1　基于回归负载预测(LPBSVR)的方法

LPBSVR 方法的工作原理是将预测负载与实际负载进行比较。因此，需要建立在预测方法的基础上。该方法采用支持向量回归(SVR)方法进行负载预测。SVR 旨在最小化与支持向量相关的误差，预测模型是基于异常值进行训练[11]。因此，该方法适用于异常模式且作为异常值处理的异常检测。

利用 SVR 得到的回归模型，其残差可以计算为每个时隙中各数据的实际用电量与预测用电量之间的差值。这些残差用概率分布函数(PDF)来表征，可以用来检测异常数据[12]。例如，假设残差的 PDF 是具有均值 μ 和方差 σ 的正态分布，则不在[$\mu-3\sigma, \mu+3\sigma$]范围内的数据可视为异常值，即异常数据和异常值。

2.2　基于神经网络负载预测(LPBNN)的方法

对于单一负载预测，神经网络(NN)具有强大的数据驱动功能[13]。因此，可以用于开发基于负载预测的异常检测方法。除了预测方法外，LPBNN 与 LPBSVR 相似。本文研究了不同的神经网络结构，有 1～7 个隐藏层，每个隐藏层有 3～13 个节点，并且 ReLU 和 Sigmoid 都是激活函数。然后，对于均方误差(MSE)，将最佳结果用于预测精度。

对于智能电表的每个新读数，首先传递给经过训练的神经网络模型来预测用电量。然后，获得残差；如果它超出了所提到的范围，那么标记为异常。对每个新数据进行决策后更新模型：如果新数据标记为正常，则作为新的训练数据更新神经网络模型和残差 PDF；否则，即如果新数据异常，则不用于更新神经网络模型。这个过程准确地实现了基于回归的模型，在这种情况下 SVR 模型进行更新。

2.3　基于聚类的方法

将所有可用数据分为“异常数据”和“正常数据”两组。本文使用孤立森林(IF)方法[14]作为基于聚类的异常检测方法。IF 的基本思想是隔离实例，而不计算测量值之间任何类型的距离。这有助于提高计算时间和在线检测。在异常情况下，IF 使用了两个主要特征：1)异常数据非常罕见；2)异常数据的某些特征与正常数据的特征有很大的不同。由于具有明显特征值的实例在早期分区过程中进行划分，同时，在包含异常的不同部分中，较少的异常会生成较少的分区，从而导致树中的路径越短。因此，聚类使用二叉树聚类时，对隔离具有敏感性，异常数据倾向于隔离在更接近二叉树根的位置。

2.4　基于投影的方法

通过使用投影向量将输入空间投影到子空间进行降维。主成分分析(PCA)和轻量级在线异常检测器(LODA)[15]是两种常用的基于投影的异常检测方法。由于 LODA 是基于特征向量的随机稀疏投影集合，对随机稀疏投影的计算效率很高。因此，本文选择 LODA 作为基于投影的异常检测方法。

训练数据相对于投影向量 $\boldsymbol{\omega}_i$ 的投影值得到了一维集合，用于获得每个集合的直方图。因此，从训练数据中得到了 k 个一维直方图。LODA 输出可以定义为样本数据的负对数似然：

$$f(\boldsymbol{x}) = -\frac{1}{k}\sum_{i=1}^{k}\log\hat{p}_i(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\omega}_i) \tag{4}$$

其中，$f(\boldsymbol{x})$ 值越高，表明样本异常的概率越低。$\hat{p}_i$ 为向量 $\boldsymbol{\omega}_i$ 和输入 $\boldsymbol{x}$ 的投影值各自的概率。对于每个输入数据的在线异常检测，计算每个 $\boldsymbol{\omega}_i$ 中的投影值，并通过其直方图找到各自的 $\hat{p}_i$。因此，可以根据训练的直方图找到输入数据的 LODA 输出。通过将该值与某个阈值进行比较，将输入数据标记为正常或异常数据。如果输入数据识别为异常，则使用它来更新直方图。

通过以下优化问题计算每个直方图的 LODA 中的组距数[16]：

$$F = \underset{w,b}{\text{maximize}} \sum_{i=1}^{b} n_i \log \frac{bn_i}{N} - [b - 1 + (\log b)^{2.5}] \quad (5)$$

$$N = \sum_{i=1}^{b} n_i \quad (6)$$

其中，N 为样本总数；n_i 为第 i 个组距中的样本数量；b 是组距的总数。对每个直方图分别求解该优化问题。作为 LODA 中的另一个重要参数，稀疏投影向量的数量计算为：

$$\hat{\sigma}_k = \frac{1}{N} \sum_{i=1}^{N} |f_{k+1}(x_i) - f_k(x_i)| \quad (7)$$

其中，k 为直方图的数量，该参数的最佳值可以通过 $\min_k(\hat{\sigma}_k/\hat{\sigma}_1)$ 的公式确定。

3　案例研究

本文的测试案例是基于安徽省合肥市政务区的智能电表数据，所收集的数据是 5 户居民家庭连续 92 天的用电数据，采样频率为 15 min/条。经过预处理和清洗后，数据分为 70%的训练数据和 30%的测试数据。异常检测本质上是无监督的学习过程。因此，本文选取的数据没有确定的异常标签，还需要获得一个基准来定义异常用电量。

3.1　定义异常用电量

异常用电量可以有不同特征和不同持续时间。为了获取不同长度的异常，本文对数据使用不同大小的移动窗口，并将最近窗口中的数据与以前窗口中的数据进行比较。通过检查对数据库中的各种实验数据，发现近 3 周的用电量数据可以有效地反映历史数据的趋势。因此，通过选取每 3 周的数据平均值，可以构建一个模型来捕捉数据随时间的变化趋势。通过比较，本文得到了一组任意窗口大小的残差数据，然后用均值 μ_p 和标准差 σ_p 拟合残差的正态分布函数，这些残差是每个窗口内所有单独时间段残差的总和。

对于任何偏离 $3\sigma_p$ 的残差，本文将其视为异常趋势或异常数据。对于在多个移动窗口中确定的点，通过分析与每个窗口大小相关的残差曲线来得到最大连续窗口，在检测的点中选取单峰曲线。因此，所有检测到异常中的峰值是检测到异常的最大窗口的大小。

两个不同窗口大小获得的 1 天内两个连续异常构建的异常基准，如图 1 所示。图 1 中描述了 1 天内的用电量、过去 3 周内同一天的平均值以及两个检测窗口大小的残差。大小为 3 的窗口检测整个时间段 44～57，但由于大小为 13 的窗口是一个较大的检测器，并且在检测跨度中有一个峰值(而不是大小为 3 的窗口中的多个峰值)，因此本文选择后一个窗口作为异常周期。

3.2　比较 4 种方法

本文将所提出的 4 种方法都应用于现有的测试数据。在所有特征都被利用的基本情况下，基于负载的特征、上下文特征和环境特征等 3 类用电数据都应用于模型以检测异常。4 种方法对部分测试数据的性能，如图 2 所示。基准

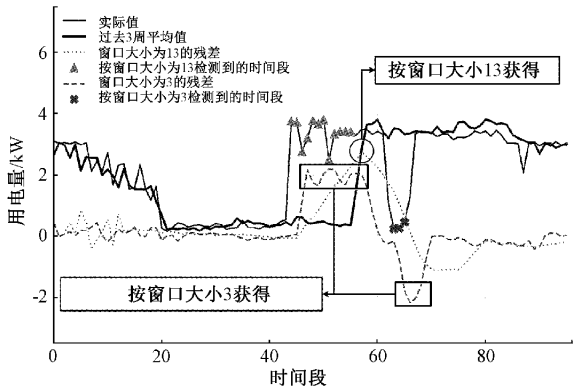

图 1　基于统计模型的异常基准

异常用三角形标注在数据曲线上，每种方法检测到的异常时间段在曲线下方用不同的形状表示。

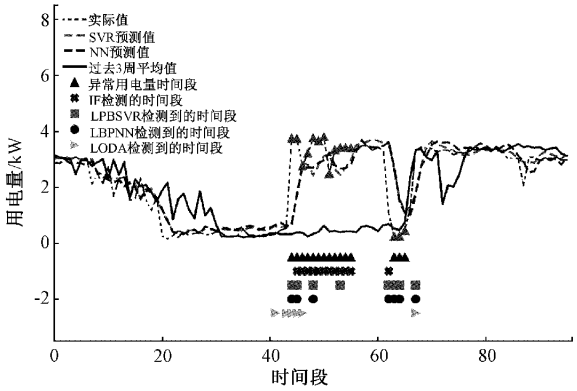

图 2　不同的异常检测方法与基准进行比较

本文使用 Matthews 相关系数(MCC)[17]对 4 种方法进行比较：

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + TF)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

其中，TP、TN、FP、FN 分别为真阳性(正确识别)、真阴性(正确拒绝)、假阳性(错误识别)和假阴性(错误拒绝)。MCC 可用作二值分类中的准确性度量，其本质包括异常检测的特殊情况。MCC 得分介于－1(最差性能)和 1(最佳性能)之间。

基于以上结果，IF 的性能最好，MCC 为 0.81，而 LPBSVR 和 LPBNN 的 MCC 分别为 0.54、0.49 和 0.47。在基本情况下，使用所有给出最佳预测结果的特征，即最低的均方误差。然而，尽管 LPBSVR 和 LPBNN 具有良好的预测性能，但在异常检测方面的性能较差。这是由于将用电量与其预测值进行了比较，而不是与之前的用电量趋势进行比较，在异常的基准点上，以前的用电量趋势有所不同。

IF 和 LODA 考虑所有特征来检测异常而不是进行预测。如果检测到多个异常点甚至超过基准点，则这些点在

某些特征上不同于通常的趋势，如湿度、温度。

3.3　特征选择与灵敏度分析

为了研究不同特性对每种方法性能的影响。不同特征选择对每种方法精度模拟结果，如表 1 所示。

表 1　不同特征选择对每种方法精度模拟结果

特征组合	IF	LODA	LPBSVR	LPBNN
全部特征	0.813 2	0.544 7	0.493 0	0.475 1
L^t	0.791 56	0.731 3	0.927 6	0.728 8
L^t+L^w	0.766 6	0.731 3	0.320 6	0.227 3
L^t+C	0.792 4	0.346 7	0.927 6	0.766 2
L^t+C+L^w	0.878 3	0.797 6	0.431 02	0.520 4
L^t+C+E	0.919 6	0.434 27	0.885 2	0.687 4

所有方法均已使用不同的阈值进行了检查，并得到每种方法的最佳 MCC 结果。此外，不同特征组合下预测方法的 MSE 结果，如表 2 所示。

表 2　不同特征组合下预测方法的 MSE

特征组合	LPBSVR	LPBNN
全部特征	0.274 1	0.301 6
L^t	0.651 6	0.637 5
L^t+L^w	0.272 3	0.389 3
L^t+C	0.750 1	0.677 0
L^t+C+L^w	1.337 6	0.377 8
L^t+C+E	0.783 9	0.654 7

在基于预测的方法中，虽然使用所有特征可以提高预测能力，但并不一定提高异常检测能力。基于 L^t+C+L^w 特征与基准测试的比较，如图 3 所示。

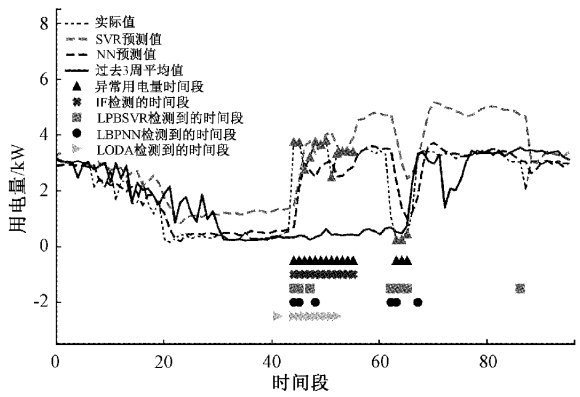

图 3　基于 L^t+C+L^w 特征与基准测试的比较

图 3 中使用了特征 L^t，C 和 L^w。尽管 MSE 较高，但 LPBSVR 和 LPBNN 的 MCC 较低。在基于预测的方法中，需要一个接近先前用电量趋势的预测，而不必接近实际用电量。对于基于预测的方法，模拟先前用电量趋势的特征更适合于异常检测。基于 L^t 特征与基准测试的比较，如图 4 所示。

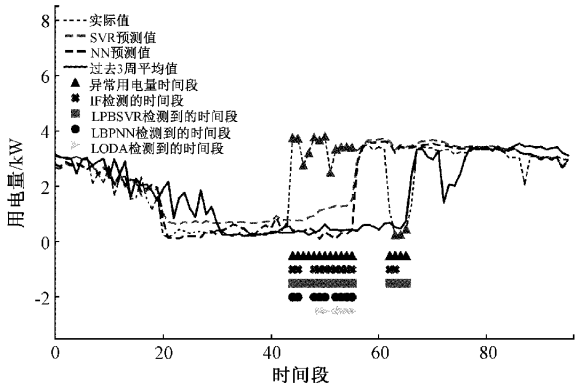

图 4　基于 L^t 特征与基准测试的比较

基于预测的方法适当地模拟先前用电量趋势作为实际负载的预期值。相反，为了更准确的预测，应该使用更好的描述性和互补性的特征。在大多数情况下，神经网络在预测用电量方面比 SVR 更有效。然而，除了同时使用 L^t，C 和 L^w 情况外，LPBSVR 在大多数情况下检测异常的性能都优于 LPBNN。这是由于 SVR 是基于异常值进行训练，进而有助于异常检测。同时，IF 的性能对 L^w（即窗口特征）不敏感。这是由于这些相互关联且密切相关的特征具有内聚性。因此，IF 无法在训练模型中对这些特征进行适当的区分。相反，如果将窗口特征替换为环境特征，则可以实现 IF 的最佳性能。这表明 IF 中的树节点对环境特征非常敏感。

仿真结果表明，基于预测的方法通常对于幅度非常高或非常低的异常更为准确。相反，其他两种方法（特别是 IF）对于不具有高峰值用电量的异常趋势更为准确。

由于 LODA 依赖于随机投影，因此必须考虑许多重复来测试其性能。以 L^t，C 和 E 为特征的不同居民和时间段的结果表明，LODA 倾向于检测真正接近于 0 的用电量。即 LODA 对非常低的用电量时间段非常敏感。然而，与使用 L^t 或具有 L^t 特征的其他方法相比，LODA 仍具有较佳的性能。

4　结　　论

对于智能电表的异常数据检测，本文检验了 4 种在线无监督机器学习异常检测方法的性能。当考虑所有可用特征时，基于聚类的方法（例如 IF）比基于预测和基于投影的方法具有更好的性能。当预测模型准确模拟先前用电量趋势而不是跟踪即将到来的实时用电量时，基于预测的方法可获得最佳性能。仿真结果表明，基于预测的方法通常对幅度非常高或非常低的异常更为准确。基于投影的方法（例如 LODA）在智能电表数据异常检测方面表现不佳。然而，当只使用可用特征的某个子集时，LODA 可以通过更好的特征选择来表现出更好的性能。在未来的研究工作

中,当涉及到智能电表数据的异常检测时,最好为每个方法单独定制特征。

参考文献

[1] 罗红波,郭洁.基于智能电表数据的变压器负载估计[J].信息技术,2020,44(11):117-120.

[2] 张春萌,关艳.使用智能电表实现工业设备的大数据识别[J].微型电脑应用,2020,36(11):114-117.

[3] 薛斌,张向东,段立,等.基于泛在电力物联网的普适性智能电表状态实时评估方法[J].电力大数据,2019,22(11):38-43.

[4] 刘紫熠,刘卿,王崇,等.基于智能电表运行故障数据的纵向分析模型[J].计算机科学,2019,46(1):436-438,456.

[5] 杨国燕,关靓.基于回归分析的智能电表可靠性分析方法[J].黑龙江大学自然科学学报,2019,36(4):498-504.

[6] 翟峰,杨挺,曹永峰,等.基于区块链与K-means算法的智能电表密钥管理方法[J].电力自动化设备,2020,40(8):38-46.

[7] 孙毅,王永生,顾博川,等.具备边缘计算能力的多芯智能电表设计[J].电子测量技术,2019,42(23):194-199.

[8] 郭亮,董勋,高宏力,等.无标签数据下基于特征知识迁移的机械设备智能故障诊断[J].仪器仪表学报,2019,40(8):58-64.

[9] 陈思伟,高翠云,胡翀.基于相似度分段及重采样的自适应波形数据压缩[J].电子测量与仪器学报,2019,33(4):178-185.

[10] 颜贝,张建林.基于生成对抗网络的图像翻译现状研究[J].国外电子测量技术,2019,38(6):130-134.

[11] 袁玉萍,安增龙.基于ramp损失函数的原空间支持向量机[J].华东师范大学学报(自然科学版),2016(2):20-29.

[12] SHU T X. Sensory data prediction using spatiotemporal correlation and LSTM recurrent neural network [J]. Instrumentation,2019,6(3):10-17.

[13] 崔胜胜,孙剑锋,马斌,等.智能电表数据和监督学习检测非技术损失的研究[J].工业仪表与自动化装置,2020(1):122-126.

[14] 李新鹏,高欣,阎博,等.基于孤立森林算法的电力调度流数据异常检测方法[J].电网技术,2019,43(4):1447-1456.

[15] 倪家明,陈博,董阳,等.一种基于时段特征的匹配算法在智能电表用电预测中的应用研究[J].计算机应用与软件,2020,37(3):82-88.

[16] 孔祥玉,马玉莹,李野,等.基于限定记忆递推最小二乘算法的智能电表运行误差远程估计[J].中国电机工程学报,2020,40(7):2143-2151,2394.

[17] 粘冬晓,杜庆治,龙华,等.基于数据间相关性的异常检测方法[J].数据通信,2018(6):44-47.

作者简介

王凯,硕士,工程师,主要研究方向为电能测量技术。

E-mail:iwd2246@163.com