基于改进 Kalman 的传感器数据加权处理算法

陈艳春[1],达钰鹏[2]
(1. 石家庄铁道大学,河北 石家庄 050043;
2. 河北省人力资源和社会保障厅信息中心,河北 石家庄 050071)

摘 要:物联网(IOT)和大数据的发展,对物联网数据质量和处理速度提出了新的要求,而物联网触感器原始数据由于噪音和虚假异常值的影响,如直接应用于大数据分析会严重影响分析结果的可靠性和有效性,大量传感器的部署也导致了虚假异常值数量的成倍增长;同时,由于物联网终端数量庞大且性能有限,数据价值密度低,使用机器学习方法进行处理性价比不高且不具有通用性,如何高效、可靠、通用地处理传感器数据并进行异常检测成为一个热点问题。该文基于统计学方法,结合 Kalman 滤波和加权融合思想,提出一种加权传感器处理预测算法。通过实验证明该算法对噪音数据处理效果相较于移动平均值,MSE 和 MAE 都得到了提升,性能提升明显,并利用 lightgbm 算法验证数据处理前后预测模型性能的变化,证明通过该算法处理后的数据更易于进行模型训练和预测。

关键词:物联网;传感器;Kalman 滤波;格拉布斯;lightgbm
中图分类号:TP212 **文献标识码**:A **文章编号**:1673-629X(2021)03-0157-06
doi:10.3969/j.issn.1673-629X.2021.03.027

A Sensor Data Weighting Algorithm Based on Improved Kalman

CHEN Yan-chun[1],DA Yu-peng[2]
(1. Shijiazhuang Tiedao University,Shijiazhuang 050043,China;
2. Information Center,Hebei Provincial Department of Human Resources and Social Security,
Shijiazhuang 050071,China)

Abstract:The development of the Internet of Things (IOT) and big data puts forward new requirements for the data quality and processing speed of the IOT,while the original data of the IOT will seriously affect the reliability and effectiveness of the analysis results due to the impact of noise and false outliers,such as the direct application to big data analysis,and the deployment of a large number of sensors also leads to the multiple growth of the number of false outliers. At the same time,due to the large number and limited performance of IOT terminals,low data value density,low cost performance and versatility of using machine learning methods to process sensor data,how to efficiently,reliably and universally process sensor data and carry out anomaly detection has become a hot issue. Based on the statistical method,combined with Kalman filtering and weighted fusion,we propose a weighted sensor processing and prediction algorithm. Compared with the moving average,MSE and MAE have been improved and the performance has been improved obviously. The performance of the prediction model before and after data processing has been verified by lightgbm algorithm,which proves that the data processed by this algorithm is easier for model training and prediction.

Key words:Internet of Things;sensor;Kalman filtering;Grubbs;lightgbm

0 引 言

任何传感器的测量都是带有误差的,误差产生的原因既有设备本身的问题,在数据采集过程中,如受传感器老化,也有转换器以及无线电传输过程中的干扰,使得接收数据中经常会产生异常跳变点,这种偏离的数据点被称为异常值[1]。异常值分为两类,一种是孤立的虚假异常值,这类异常值一般是孤立出现,属于噪音的一种;另一种是真实的异常值,这类异常值一般连续出现,反映观测对象发生异常变化。

在物联网监测中,为了兼顾计算成本和计算量,常常采用 SPC、PCA 等方法对建筑等监测对象进行异常识别[2-3]。此类基于统计的方法可以基于历史数据实

收稿日期:2020-05-06 **修回日期**:2020-09-10
基金项目:河北省重点研发计划项目(19210804D)E201910185;国家铁路局安全技术中心(ZRGY-CCGP-19080115)
作者简介:陈艳春(1974-),女,博士后,教授,博导,研究方向为技术创新与区域经济;达钰鹏(1988-)男,硕士,系统架构设计师,CCF 会员(B6472M),研究方向为数据分析、网络安全。

现对异常的识别,其理论基础是经过实践检验的[4]。但这类算法对输入数据的准确性要求较高,直接使用未经处理的原始数据极易发生误报。同时,因为使用场景的不同,很多传感器的观测误差存在周期性变化,处理算法如果采用静态参数,则无法拟合周期性变化。

为了解决实时监控中由于虚假异常值出现产生的误报问题,常用的手段有两个[5-6]。一是在同一位置部署多个同类型传感器同时采样,利用传感器误差近似正态分布的特点,利用统计学方法中的拉伊达法或格拉布斯法剔除虚假异常点后计算平均值,可以实现对噪音和孤立虚假异常点的较好过滤,当传感器数量较多时使用拉伊达法,较少时使用格拉布斯法。这类基于统计理论的方法的优势是计算方法简便,通用性强,鲁棒性好,在系统边缘节点便可部署使用,天然适合分布式部署,但缺点是需要在单个位置部署大量传感器,系统硬件部署和维护成本较高。二是利用各种滤波手段对单个传感器数据进行实时修正。例如小波分析、Kalman 滤波等,但这类滤波手段也存在着一些不足。以 Kalman 滤波算法为例,Kalman 滤波是一种利用线性系统状态方程,通过系统输入输出观测数据,对系统状态进行最优估计的算法,本质是利用两个正态分布的融合仍是正态分布这一特性迭代最优估计值。Kalman 滤波虽然性能优越且得到大量实践证明其正确性,存在的不足是:由于无法确定测量过程中的系统噪声和量测噪声特性,只是试验中给定了 Q 和 R 的噪声参数,而 Kalman 滤波是基于精确数学模型递推的过程。随着测量时刻的不断增加,根据递推公式求得的 $P(k \mid k)$ 会逐渐趋于 0 或者某一稳态的常数,但是最优观测值 $X(k \mid k)$ 与实际数据的差距越来越大,这种情况下,Kalman 滤波器的预测和估计的功能逐渐丧失。而且由于滤波的实现平台在计算式计算误差的不断累计传递也可能会使滤波器出现发散的现象[7]。同时,这类滤波算法参数调试复杂,而随着传感器设备老化,设备误差会变化,相应参数也需要做出调整,需要长期跟踪滤波效果,适时调整算法参数,增加了日常维护人员的工作负担[8]。

基于以上,该文提出一种结合以上两种手段的传感器实时数据处理算法,通过高频率采样,将单传感器单位时间内多次采样值看作为多传感器的一次采样值,用统计方法剔除虚假异常值后的观测值,再利用 Kalman 滤波处理求得最优估计值,较好地解决了虚假异常值产生的误报问题。

1 算法描述

该文提出的算法,根据实践中观测值短期内存在变化自相关性,而长期内正态分布的特点[9],将单位时间段内传感器监测数据看作一个线性定常系统。假设在单位时间段内所观测的值是恒定的,可将单传感器多次采样值看作为多传感器的一次采样值,则该时间段内传感器的观测误差和噪音都是近似正态分布的,那么基于该值计算的移动平均值也是近似正态分布的,这样的情况下是可以通过 Kalman 进行数据融合的。

首先,利用统计方法剔除虚假异常值影响,再用移动平均值降低随机噪音的影响。滑动窗口方式获取数据,设采样传感器数为 N,当 $N>100$ 时,采用拉伊达法即 3 西格玛法,计算观测值 X_i 与平均值 X 残差的绝对值 V_i,当 V_i 大于 3 倍标准差 std 即($\mid X_i - X \mid \geqslant 3 *$ std)时,认为该观测值为异常值,剔除所有异常值后计算剩余值的平均值 gbmean。当 $N \leqslant 100$ 时,采用格拉布斯法,根据 n 和显著性水平 a 计算 $g(n,a)$,将观测值从小到大排序,计算最大值和最小值各自与平均值的残差的绝对值 V_i,当 $V_i > g(n,a) *$ std,判断其为异常值剔除,并重新计算剩余观测值的均值和 std,重复以上步骤直至没有出现异常值,而后计算剩余值的平均值 gbmean[10-11]。格拉布斯法相较于拉伊达法更严谨准确,但由于需要排序和反复计算均值、标准差,当 N 较大时计算量较大,而当 $N>100$ 时候,其结果和拉伊达法接近,为简化计算可使用拉伊达法剔除异常值[12]。

然后,利用 Kalman 滤波计算最优估计值。以 gbmean 作为经验值,移动平均值 f 作为观测值,二值单位时间内各自标准差作为其观测误差,而后根据上一次滤波时 f 值和 gbmean 值各自与 t 值的绝对误差进行加权融合出新的误差 std_1 和 std_2,这样的误差结合传感器自身最近误差以及与最优估计值误差,当出现孤立异常值时更相信 gbmean 值,当出现连续异常值时更相信 f 值。

实时计算 Kalman 增益,根据一元卡尔曼滤波增益系数计算公式(1):

$$K = \frac{\text{std}_2^2}{\text{std}_1^2 + \text{std}_2^2} \tag{1}$$

计算出最优估计值 t。文中 f 值和 gbmean 值的误差并非如在经典 Kalman 滤波中是估计值,而是基于单位时间段值滑动数据计算而来的,避免估计中错误累积导致的离散问题。在计算出估计值后,结合历史 t 值,基于 SPC 法,根据该时间段 t 值的采样次数,利用格拉布斯法查表或拉伊达法确定控制限,进行异常值判断。

之所以选择 f 值和 gbmean 值进行融合,一是为了解决缺失值问题,由于各种因素,传感器数据出现缺失值是十分常见的,移动均值可以比较好地解决常见线

性系统的缺失值填充问题。二是为了更好地识别连续出现的异常值。如果直接使用带有噪音的原始观测值,f 值虽然可以改善但仍无法完全剔除虚假异常值影响,很容易出现误报;而仅使用 gbmean 值,由于会将第一个出现的真实异常值认为是虚假异常值,需连续出现的异常值影响总体方差后才能体现异常情况,故对于真实发生的异常反应具有滞后性。该文使用了 Kalman 法将 f 值和 gbmean 值进行融合,当异常值连续出现时能够较快反映异常情况,在解决孤立虚假异常值的基础上,实现对异常情况的较快反应。

2　算法流程

以图 1 为例,说明此算法的计算流程。

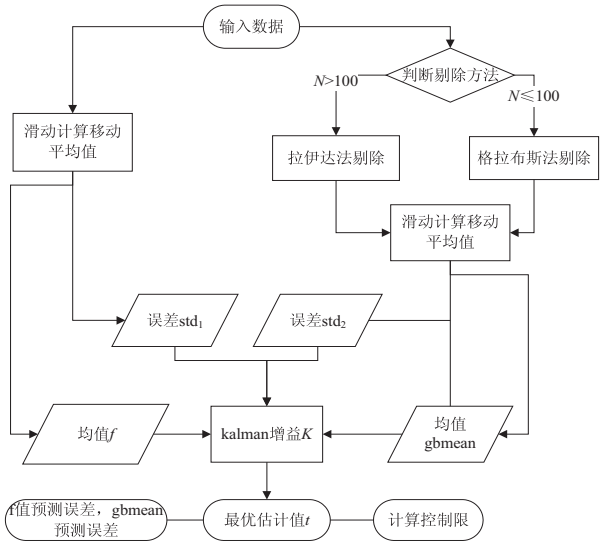

图 1　算法流程

第一步:获取 Z 时间内的所有采样值,先计算 Z 时间内采样数据的移动平均值 f,以其标准差为其观测误差 x_1,上一次预测绝对误差为 x_2。根据式(2)~式(4)[13]:

$$w_1 = \frac{x_2}{x_1 + x_2} \tag{2}$$

$$w_2 = \frac{x_1}{x_1 + x_2} \tag{3}$$

$$\mathrm{std}_1 = w_1 * x_1 + w_2 * x_2 \tag{4}$$

计算其误差 std_1。而后,如采样次数 $N \leqslant 100$ 时,根据设定的显著性水平 a,利用格拉布斯法剔除异常值后计算移动平均值 gbmean;当 $N>100$ 时,利用拉伊达法剔除异常值后计算移动平均值 gbmean,以其标准差为其观测误差 x_3,上一次预测绝对误差为 x_4,计算其误差 std_2。

第二步:移动窗口取数据进行滚动计算,实时更新 f 值和 gbmean 值及其各自误差,根据式(1)计算 K 值,进而根据 Kalman 校正式(5):

$$t = \mathrm{gbmean} + K * (f - \mathrm{gbmean}) \tag{5}$$

求出最优观测值 t,并计算 Z 时间段内各采样点 t 值的标准差 std_3。

第三步:计算控制限,当 $N>100$ 时,V_i 控制限为 $3 * \mathrm{std}_3$;当 $N \leqslant 100$ 时,需根据格拉布斯法则计算 $g(n,a)$,V_i 控制限为 $g(n,a) * \mathrm{std}_3$。超出控制限的为异常值。

整个算法中需要设定的参数只有 N 以及使用格拉布斯法时的显著性水平 a,也只需存在 N 个 f 值、N 个 gbmean 值和 N 个 t 值共 $3N$ 个值,所需存储量和计算量较小,调整参数难度较低,可在物联网终端或边缘服务器进行数据处理,实现分布式部署。

3　仿真数据验证

3.1　数据准备

该文所采用的数据源于石家庄站铁路站房 2013 年 7 月至 2014 年 7 月数据中的 M16BL-1 传感器数据 data2,在 anaconda3 环境下进行数据分析。为了验证算法的滤波性能,在已有数据基础上,将采样密度调整至每分钟一次,结合原始值利用 pandas 中线性插值法填充调整后产生的缺失值,生成数据 525 583 条,填充后数据折线趋势图如图 2 所示。

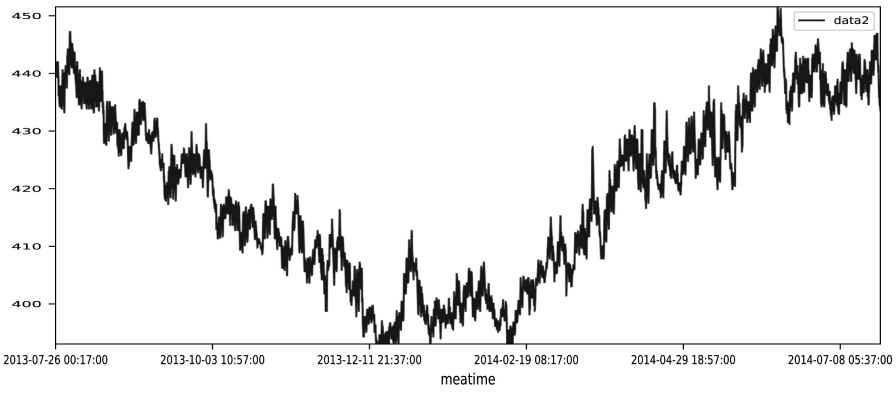

图 2　data2 趋势图

而后先利用 numpy. randon. normal 函数生成均值为 0,标准差为 8 的正态分布随机噪声,再利用 numpy.

randon. randint 函数生成 1 000 个[50,100]和 1 000 个[-100,-50]的随机异常值点,这些异常值随机分布,有孤立有连续,data2 值加上噪音和异常值点的数据为 etl2 值,也就是要处理的数据值,折线趋势图如图 3 所示。

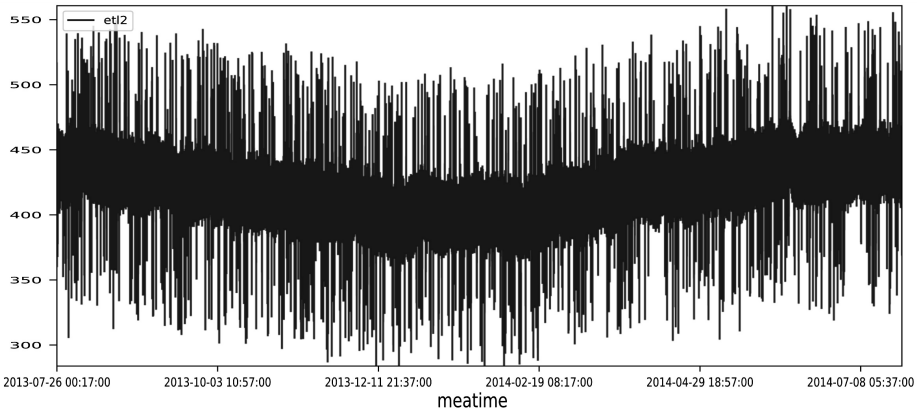

图 3 etl2 趋势图

数据表格式如表 1 所示。

表 1 处理前数据

meatime	data2	etl2
2013-7-26 0:17	434.4	437.929 8
2013-7-26 0:18	434.38	431.733
2013-7-26 0:19	434.36	453.806 2
2013-7-26 0:20	434.34	432.323 3
2013-7-26 0:21	434.32	435.196 9
2013-7-26 0:22	434.3	446.959 8

当然,为了更加直观地体现算法处理效果,增加的噪音和异常值较为明显,实际异常值不会这么多。

3.2 算法实现

在算法实现上,为了快速实现,采用 python 语言进行开发,在 anaconda3 环境下,使用了 pandas、outliters、math 包。其中格拉布斯法是 outliters 包的 smirnov_grubbs 函数实现。移动平均值使用 pandas. dataframe. rolling 函数计算,由于采用数据的短期内自相关而长期内正态分布的特点,根据参考文献[10]的统计结果,选取 15 分钟作为 Z 时间段长短,设置滑动时间窗口为最近 15 分钟的采样数据,移动平均值为 f15,即 N 值为 15,数据以 dataframe 格式存储为 df1,显著性水平为 0.99,初始预测误差为 15,gbmean 值和 t 值计算代码如下:

```
Import pandas aspd
Import matplotlib. pyplot as plt
From sklean. metrics import mean_squared_error
From sklean. metrics import mean_absolute_error
Import numpy as np
Form outliers import smirnov_grubbs as grubbs
Import math
df1 = pd. read_csv( " 含噪音数据. csv" )
df1[ ´gbmean´] = None
df1[ ´t´] = None
df1[ ´vi1´] = None
df1. loc[ 14 , ´vi1´] = 15
df1. loc[ 14 , ´vi2´] = 15
for i in range( 15 , len( df1) ) :
    s = grubbs. test( df1. loc[ i-14 : i , ´etl2´] , 0. 99 )
df1. loc[ i , ´gbmean´] = s. mean( )
    std1 = df1. loc[ i-14 : i , ´f15´]. std( )
x1 = df1. loc[ i-1 , ´vi1´]
x2 = df1. loc[ i-1 , ´vi2´]
    w1 = cou( std1 , x1)
    w2 = 1-w1
    std2 = s. std( )
    w3 = cou( std2 , x2)
    w4 = 1-w3
    std1 = std1 * w1+w2 * x1
    std2 = std2 * w3+w4 * x2
stdx2 = math. pow( std2 , 2) /( math. pow( std2 , 2) + math. pow
( std1 , 2) )
f1. loc[ i , ´t´] = df1. loc[ i , ´gbmean´] +stdx2 * ( df1. loc[ i , ´f15
´] -df1. loc[ i , ´gbmean´] )
df1. loc[ i , ´vi2´] = abs( df1. loc[ i , ´t´] -df1. loc[ i , ´gbmean´] )
df1. loc[ i , ´vi1´] = abs( df1. loc[ i , ´t´] -df1. loc[ i , ´f15´] )
```

计算出 gbmean 值、t 值后,去除少量由于移动均值计算产生的缺失值,数据如表 2 所示。

表 2 处理后数据

meatime	data2	etl2	f15	gbmean	t	vi1	vi2
2013/7/26 0:32	434.10	428.50	437.51	431.66	434.65	2.86	2.99
2013/7/26 0:33	434.08	443.29	438.28	431.93	435.15	3.13	3.22

续表 2

meatime	data2	etl2	f15	gbmean	t	vi1	vi2
2013/7/26 0:34	434.06	448.92	437.95	431.93	434.89	3.07	2.95
2013/7/26 0:35	434.04	421.95	437.26	431.42	434.61	2.65	3.19
2013/7/26 0:36	434.02	439.18	437.53	432.70	436.23	1.29	3.53

利用 matplotlib 包进行数据可视化,趋势图如图 4　和图 5 所示。

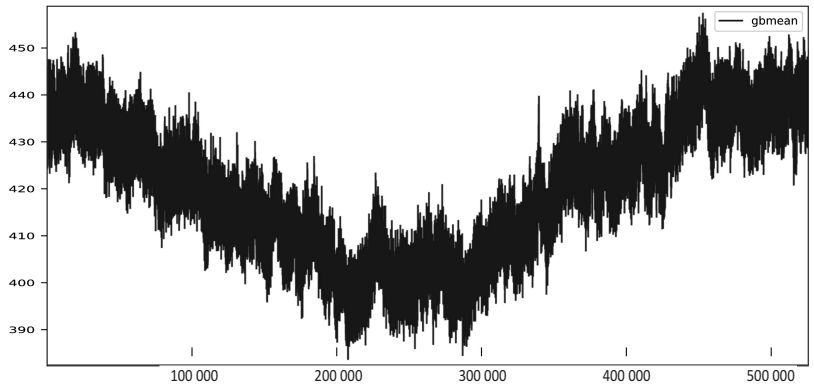

图 4　gbmean 趋势图

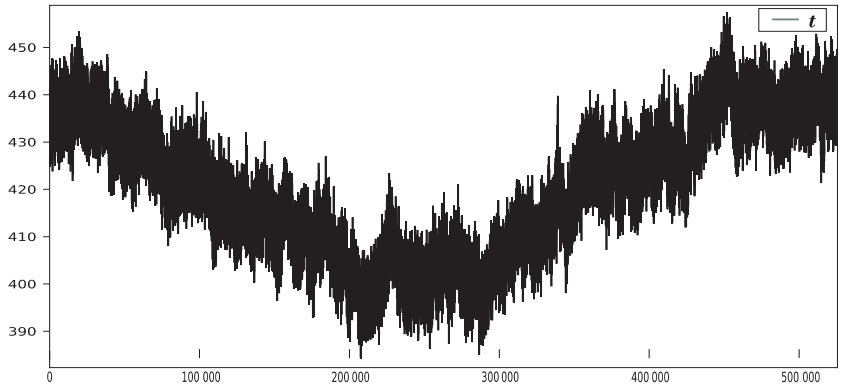

图 5　t 趋势图

计算出 t 值,每次计算最近 15 个 t 值的均值 x_t 和方差 std_t,当 $N=15$,显著性水平为 0.99 时,最优估计值的残差 $g(n,a)=g(15,9)=3.292$,当 $|t-x_t|>3.292*\mathrm{std}_t$ 时,判断发生异常。

3.3　结果验证

首先,利用 matplotlib 包进行数据可视化,比较将 f15、gbmean、t 和真值 data2 值放在一张图里进行对比,定性比较处理效果,由于数据集较大,图 6 是整体滤波效果。

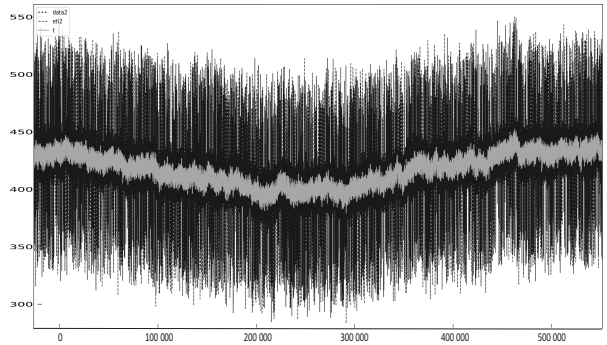

图 6　滤波效果图

然后用 sklean. metrics 包进行定量比较。计算 etl2、gbmean、t 与 data2 值的均方误差(MSE)和平均绝对误差(MAE),MSE 是最常用的回归损失函数,计算方法是求预测值与真实值之间距离的平方和,MAE 是目标值和预测值之差的绝对值之和的平均值。MSE 会赋予异常点更大的权重,也就是说对预测误差给予更大的权重,用 MSE 可以比较方法对异常值处理的好坏,而 MAE 值体现了误差,显示了方法处理的性能。源数据 etl2 的 MSE 约为 119.45,f15 值的 MSE 约为 7.99,gbmean 值的 MSE 约为 8.00,t 值的 MSE 约为 6.98,t 值的 MSE 相较于 gbmean 下降了 12.73%;处理前数据 etl2 的 MAE 约为 7.04,f15 值的 MAE 约为 2.11,gbmean 值的 MAE 约为 2.23,t 值的 MAE 约为 2.03,t 值的 MAE 相较于 gbmean 下降了 8.72%。

同时计算各采样点 f15 值、t 值和 gbmean 值相较于真值 data2 绝对残差的标准差,f15 值残差标准差约为 1.88,t 值残差标准差约为 1.68,gbmean 值残差标准差约为 1.74,t 值残差标准差相较于 gbmean 值残差标准差降低了 3.18%,算法效果较为稳定。综上可

知,该算法性能和可靠性提升明显。

3.4 对预测效果提升的验证

为了进一步证明数据处理对预测结果的影响,该文以工业界常用的 lightgbm 算法为例,对处理前后的数据进行模型训练和预测。LigthGBM 是决策树预测模型(GBDT)的一种,由微软提供,它和 XGBoost 一样是对 GBDT 的高效实现,原理上它和 GBDT 及 XGBoost 类似,都采用损失函数的负梯度作为当前决策树的残差近似值,去拟合新的决策树。基于 Histogram 的决策树算法,更低的内存占用和更快的处理速度,基于 OpenMP 多线程加速和基于 OpenCL 的异构加速,可以利用多种硬件加速学习过程,应用范围更广[14]。

首先,利用上文的处理算法对数据进行处理;然后,对于周期性较强的时间序列问题,可以通过将时间序列问题转化为回归问题进行预测,这里将时间这一特征进行特征工程处理,分解出表 3 的 10 个特征。

表 3 训练特征

名称	意义	备注
year	年	特征 1
month	月	特征 2
day	日	特征 3
hour	小时	特征 4
minute	分钟	特征 5
day_of_week	周几	特征 6
quarter	季度	特征 7
day_of_year	当年第几天	特征 8
day_of_month	当月第几天	特征 9
week_of_year	当年第几周	特征 10

然后,分别使用原始数据 etl2 和处理后数据 xmean1 进行 lightgbm 模型训练,这里由于已经是回归问题,故采用随机抽取的方式取 75% 的数据为训练集,25% 的数据为测试集,以 MSE 值作为训练依据,以 ETL2 值训练的模型 MSE 值为 118.447,以 xmean1 值训练的模型 MSE 值为 7.762 15,数据处理后模型预测的 MSE 值下降了 93.44%,性能提升明显。

4 结束语

该算法结合了现有物联网数据处理中的两种处理方法,将多传感器数据测量方法应用于单传感器数据处理,兼顾了成本和准确性,可以解决实际工作中物联网大量传感器高采样次数,传感器误差存在周期性变化的场景,实现动态剔除异常值、实时滤波和实时计算控制限。相较于传统 Kalman 滤波法,调整参数少,维护成本低;相较于神经网络等机器学习方法,计算量和存储数据量小,适合实时数据处理和分布式进行部署。

参考文献:

[1] 费业泰. 误差理论与数据处理[M]. 合肥:合肥工业大学出版社,2005.

[2] KULLAA J. Distinguishing between sensor fault, structual damage, and environmental or operational effects in structural health monitoring[J]. Mechanical Systems and Signal Processing, 2011, 25(8): 2976-2989.

[3] SHEWHART W A. Economic control of quality manufactured product[M]. New York: Van Nostrand Company, 1931.

[4] SOHN H, CZARNECKI J, FARRAR C. Structural health monitoring using statistical process control[J]. Journal of Structural Engineering, 2000, 126(11): 1356-1363.

[5] 沙定国. 测量不确定度与测量误差(续三)[J]. 光学技术, 1996(1): 43-46.

[6] 卢元磊,何佳洲,安 瑾,等. 几种野值剔除准则在目标预测中的应用研究[J]. 指挥控制与仿真,2011,33(4):98-102.

[7] 刘泰营. 基于物联网的梁式桥结构长期健康监测系统研究[D]. 北京:北京理工大学,2016.

[8] 殷建军,余忠华,李兴林,等. 基于 Kalman 滤波的过程调节与质量监控方法[J]. 浙江大学学报:工学版,2008,42(8):1419-1422.

[9] 黄祖光,申兆武,刘 军,等. 基于 MSPC 的铁路站房健康监测系统设计[J]. 铁道工程学报,2016,33(7):83-87.

[10] 朱赵辉,孙建会,王万顺,等. 基于格拉布斯准则的小波阈值去噪算法研究[J]. 西北水电,2011(S1):45-48.

[11] 钟继贵. 误差理论与数据处理[M]. 北京:水利电力出版社,1993.

[12] 魏治文,程 琳,来记桃,等. 几种异常值判别准则在安全监测数据处理中的应用[J]. 大坝与安全,2009(1):67-69.

[13] 杨军佳,赵瑞峰,王世军,等. 一种多传感器数据加权融合方法:CN108985373A[P]. 2018-12-11.

[14] KE G, MENG Q, FINLEY T W, et al. LightGBM: a highly efficient gradient boosting decision tree[C]//Neural information processing systems. Long Beach, California, USA: Curran Associates Inc., 2017: 3149-3157.