DOI:10.20033/j.1003-7241.(2025)05-0061-05.

# 基于邻域链的传感器网络数据异常检测方法

晁永兰

(青海交通职业技术学院信息工程学院,青海 西宁 810003)

**摘 要:** 传统对于传感器网络异常数据的检测主要采用核密度估计算法,易受到数据分布不均匀的影响,导致检测准确率较低等问题。为此,提出基于邻域链的传感器网络数据异常检测方法。根据数据采样值的波动范围,计算节点数据的置信度与置信区间,以此明确异常数据来源,利用邻域链算法计算两节点间数据的相似性,判断数据状态,通过对数据列表进行重构,分析数据在网格内部和边缘的分布规律,并将数据邻域距离与阈值比较,以输出数据异常值,由此实现数据异常检测。对比实验结果表明,所提方法能够较为准确地检测出传感器网络中的异常数据。

**关键词:** 邻域链算法;传感器网络;数据异常;检出率

中图分类号:TP311.13 文献标识码:A 文章编号:1003-7241(2025)05-0061-05

# Anomaly Detection Method of Sensor Network Data Based on Neighborhood Chain

**CHAO Yong-lan**

( Qinghai Communications Technical College, Xining 810003 China )

**Abstract:** Traditionally, the detection of abnormal data in sensor networks mainly uses kernel density estimation algorithms, which are susceptible to the impact of uneven data distribution, resulting in low detection accuracy and other issues. Therefore, a neighborhood chain based data anomaly detection method for sensor networks is proposed. According to the fluctuation range of data sampling values, calculate the confidence level and confidence interval of node data to identify the source of abnormal data. Use the neighborhood chain algorithm to calculate the similarity of data between two nodes, judge the data status, reconstruct the data list, analyze the distribution rule of data within and at the edge of the grid, and compare the data neighborhood distance with the threshold value to output data abnormal values, thereby achieving data abnormality detection. Comparative experimental results show that the proposed method can accurately detect abnormal data in sensor networks.

**Keywords:** neighborhood chain algorithm; sensor network; abnormal data; detection rate

## 0 引言

无线传感器网络技术因其获取信息的快速性与数据处理的高效性,被应用于众多领域中,是物联网的重要支撑技术之一。为了能够及时监测到各种可能发生的事件,需要对传感器各个网络节点采集到的多元数据进行解析,以实时准确地检测出异常数据,并及时采取相应的修补措施。

对于传感器网络中异常数据流的检测算法,文献[1]提出使用 BIRCH 聚类算法对原始数据进行聚类,以隔离异常数据。然而,这种方法需要在获取所有数据后进行聚类,并且无法在线检测异常数据,并且需要很长时间来计算每个数据之间的距离;文献[2]提出了一种基于大数据技术的检测算法,将不在模型识别范围内的数据视为异常数据,但该方法不适用于维度较高的数据集合。为了解决上述问题中存在的问题,利用邻域链算法提出了数据异常检测方法,不仅降低了时间复杂度,而且还提高数据异常检测的检测率。

## 1 传感器网络数据异常检测方法设计

### 1.1 异常数据来源验证

在自然环境中,传感器在某一时间的采样值 $r_t(i)$会出现波动情况,且波动范围内的数据维度不断增大[3]。当传感器在任意 $t$ 时刻的读数 $r_t(i)$不在固定范围内或 $r_t(i)=r_{t-1}(i)$时,则认为发生了异常。

为了验证异常数据的来源节点,需要计算数据和限定区间内的相异性 $\gamma$,并计算数据采集时间的间隔差和采样值的平均值。若平均值小于 1,则可认为数据异常是由于采样误差或置信区间的估计偏差造成的[4]。

通过上述分析,可以对传感器网络中的高维数据进行简化处理,使其在网络内部的分布符合标准正态分布

*基金项目:青海省科技成果转化专项项目(2021-0204-GXC-0016)

收稿日期:2023-12-26

规律，公式如下：

$$p=P\left(\frac{P_T-r_t(i)}{\sigma_t}\right)=1-\Phi(\delta) \tag{1}$$

式中，$\Phi(\delta)$表示数据的标准正态分布，其与数据输出值$P$之间成线性关系，即输出值越大，数据分布越密集；输出值越小，数据分布越稀疏。为简化运算流程，这里取$\delta=2$；$p$表示数据随机变量；$P$表示节点输出值集合；$P_T$表示不同采样时刻采集到的数据点构成的集合；$\sigma_t$表示采集误差。

如果整个传感器网络中共存在$k$个正常数据，则在网络部署前和部署后，所有数据构成的集合将会存储在传感器的中心节点中。因此，可根据网络中正常数据的个数$k$与给定的置信度$\alpha$估计数样本的区间$[L_1,L_2]$，若采样值$r_T(i)$不满足区间计算，则将当前节点视为异常数据来源[5]，公式如下：

$$L_1+\alpha\le r_T(i)\le L_2+p \tag{2}$$

式中，$L_1$、$L_2$分别表示数据置信区间的最大值与最小值；$r_T(i)$表示$T$时刻的采样值；$\alpha$表示置信度。若采样值$r_T(i)$在该区间内，则通过调整滑动窗口来更新集合中的数据元素。

此外，当传感器节点发生故障时，在数据分布较为稀疏的条件下，传感器的任意一个节点与其相邻的多个节点之间也具有强关联性[6]，如图1所示。

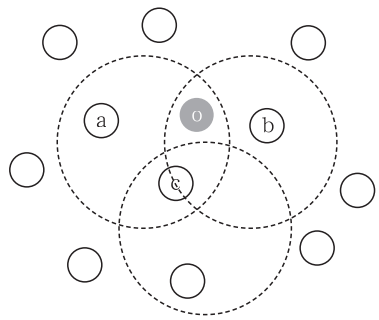



**图1 传感器网络不同节点对数据异常来源的验证**

分析图1可知，如果o处存在异常数据，则应在其节点和附近节点a、b和c检测到该事件(例如温度显著升高)。因此，传感器节点采集数据流的时空关联性为数据异常检测提供了数据依据[7]。

当传感器某一节点出现数据异常时，当前采集到的数据流中可能存在数据缺失或损坏现象，所以，为避免这一情况发生，不同位置节点将会在故障出现前将历史数据备份到邻居节点，并在所有完好节点间传播[8]。假设存在$N$个分布在二维空间中的无线传感器网络节点，该范围内的异常事件用结构函数$s(t,x,y)$表示，其中，$t$表示异常数据集$s$合出现的时间，$(x,y)$表示异常源的坐标。那么，节点采样时的异常数据集合$s$的观测值可以表示为：

$$X_i(n)=S_i(n)+N_i(n)\bullet\gamma(T) \tag{3}$$

式中，$X_i(n)$表示异常数据集合的观测值；$S_i(n)$表示异常数据集合的实测值集合；$N_i(n)$表示数据采集时混入的高斯噪声；$\gamma(T)$表示$T$时刻的数据的区间差异度。

由于数据采集时外界噪声的存在，所以数据流的观测值具有较大程度的失真，因此需要对其进行修复处理[9]，公式如下：

$$D=E\left(\overline{S}-X_i(n)\right)^2 \tag{4}$$

式中，$D$表示均方差；$E$表示取余函数；$\overline{S}$表示节点对事件$s$的观测值集合。

当观测值的均方差满足下式时，即可认为与该节点直接连接的相邻节点出现异常数据：

$$\mathrm{mod}(D)=\mathrm{mod}(|W|) \tag{5}$$

式中，mod(.)表示换阶函数；$W$表示数据长度。

通过将传感器节点的输出数据进行随机变量化，使其符合标准正态分布，并计算数据置信度与总样本区间，对数据范围进行划分，利用观测值的失真程度与数据长度比较，确定异常数据来源，为接下来估计两节点间数据的相似性奠定基础。

## 1.2 基于邻域链的两节点间数据相似性计算

在验证异常数据来源之后，为准确检测出异常数据，引入邻域链算法计算网络节点间数据的相似度。首先需要建立一条连接起点和终点的邻域链，且链中的每一个数据点均为其后一个点的近邻数据点[10]，邻域链如图2所示。

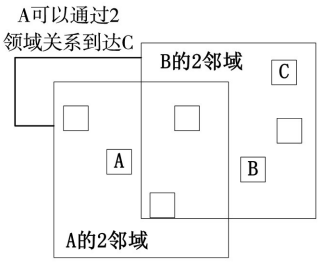



**图2 建立点A到点C的邻域链**

若存在一个正整数$q$，则在点$A$和点$C$之间一定具有如下关系：

$$R(A,C)=\min q \tag{6}$$

式中，$R$表示数据集合；$\min q$表示建立的从$A$到$C$邻域链的最小邻域数。

为消除节点间数据(假设为$A$和$C$)的冗余度，采用邻域可达性($NRC(A,C)$)与邻域密度($NRS(A,C)$)综合计算数据相似性[11]。

邻域可达性($NRC(A,C)$)定义为：以传感器网络中的任意两点$A$和$C$为起讫点，在两者之间建立一个使邻域值最小的函数，即：

$$\left(NRC\left(A,C\right)\right)=f\max\left(R\left(A,C\right),R\left(C,A\right)\right)\bullet\mathrm{mod}\left(D\right) \tag{7}$$

式中，$f(.)$表示指数函数；$\mathrm{mod}(D)$表示异常事件来源。

邻域密度($NRS(A,C)$)定义为：以传感器网络中的任意两点$A$和$C$为起讫点，在两者之间建立一个使邻域值最大的函数，即：

$$\left(NRS\left(A,C\right)\right)=\max\left\{Q\left(A,C\right),Q\left(C,A\right)\right\} \tag{8}$$

式中，$Q(A,C)$表示从$A$到$C$的邻域链中，邻域中心点到边缘的最短距离。

在实际传感器网络中，如果数据序列的滑动窗口向量为固定值$u$时，则需要利用邻域可达性和邻域密度分别计算数据流$Y_i=(Y_{1i},Y_{2i},\cdots,Y_{pi})$的协方差矩阵：

$$V_{11}=\frac{Y_i}{u}\sum\left[\left(NRS\left(A,C\right)\right)-\left(NRC\left(A,C\right)\right)\right] \tag{9}$$

由于数据采集时间$t$的不一致性，传感器节点采集到的数据规模容量也在不断扩大[12]。通常情况下，在采集时间$t$的前一刻和后一刻的数据与历史数据具有较大相关性。假定数据采集时间的间隔为$\triangle t$，则在$t-\triangle t$到$t+\triangle t$时段内，采集的数据序列可表示为：

$$F\left(t\right)=\left(V_{11}\left(t-\Delta t\right),V_{11}\left(t\right),V_{11}\left(t+\Delta t\right)\right) \tag{10}$$

两节点间邻域链移动后的数据序列$F(t)$可表示为式(11)：

$$F_1\left(t\right)=\sqrt{\sum\left[F\left(t\right)-D\left(F\left(t\right)\right)\right]} \tag{11}$$

式中，$D(F(t))$表示数据序列的信息熵。

对于数据序列的信息熵，其距离可表示为：

$$d\left(F_1\left(t\right),F_2\left(t\right)\right)=\left|F_1\left(t\right)-F_2\left(t\right)\right| \tag{12}$$

式中，$F_2(t)$表示标准化处理后的数据序列。

根据数据序列信息熵的距离，结合邻域密度与可达性，得到数据相似性$COM(A,C)$的计算公式为：

$$COM\left(A,C\right)=\frac{1}{d\left(F_1\left(t\right),F_2\left(t\right)\right)} \tag{13}$$

由式(13)可以看出，在数据相似性度量下，如果两个节点间的数据存在紧密相连关系，则认为它们之间是相近的；如果两者之间存在密度隔离，则认为它们之间的距离较远。对于相近的两节点间数据进行状态判断，分析当前节点输出值是否为异常，由此实现异常数据点的检测。

### 1.3 实现传感器网络数据异常检测

由于数据在传感器网络内部和边缘的分布没有特定规律，若某一节点的数据频繁出现异常情况或者异常数据分散在正常数据节点周边，则采用传统方法无法有效区分开正常数据与异常数据。为解决此问题，本文充分考虑传感器节点数据的分布规则，结合数据快速排序算法，通过将网络划分为若干个单元格，并计算各单元格及其邻域中的数据点数量$N(C)$和$N_D(C)$，并将其填充在相应的空白单元格中，根据预设的距离阈值$\beta$对数据列表进行重构，列表的特征用于确定无线传感器网络数据的异常情况。

设置$N(C)$为单元格$C$中的数据点数量，设置$N_D(C)$为该单元格$C$附近的$D$邻域中数据点的数量。每个非空单元的数据列表是包含$N(C)$和$N_D(C)$的排序列表，则异常数据检测的具体实现步骤如下：

(1) 归一化数据样本集合$\Gamma$的元素；

(2) 利用数据相似度构造单元网格$C_{COM}$及数据列表$C_{COM,obj}$，初始化$C_{COM,obj}$，$N(C)=0$，$N_D(C)=0$；

(3) 在每个单元格中，利用邻域可达性计算公式计算样本数据与标准数据集之间的最短距离$R_D$，并将计算结果与数据相似性、数据点数量同时填入相似性列表中；

(4) 判断$R_D$是否大于设定阈值$\vartheta$，若是，则进入步骤(5)，否则，返回步骤(3)；

(5) 将重构列表中的$C_{COM}$进行降序排序，以更新数据列表$C_{COM,obj}$，根据邻域中的数据点$N_D(C)$的动态感知向量判断数据的状态，若$R_D>\beta$，则该单元格内的数据为异常数据；否则，为正常数据；

(6) 输出$j$个异常值。

至此，完成基于邻域链算法的传感器网络数据异常检测。

## 2 实验论证

为测试本文算法的性能，利用公开数据集与人工合成数据集对原始数据展开仿真实验，以验证所提方法的有效性。

### 2.1 实验准备

本文使用的实验数据来自无线传感器网络原型系统，通过人工合成获得了近80 000条数据记录。传感器网络的范围约为120 m× 80 m，共有200个节点。通过节点之间的多跳广播，感测到的温度、光和其他信息被周期性地发送回Sink节点。

**表1 实验参数信息**

| 参数 | 数值 | 参数 | 数值 |
|---|---|---|---|
| 数据总个数 | 79 985 | 初始区间个数 | 100 |
| 数据最大值 | 58 489 | 读数距离 | 20 |
| 数据最小值 | 1 254 | 非叶子节点个数 | 20 |
| 传感器在事件区域的测量值 | [28,30] | 传感器在正常区域的测量值 | (100,10) |
| 传感器的错误测量值 | [30,100] | 通信半径 | $\sqrt{2}$ |

所有实验均在MATLAB R2010a平台上进行。在模拟实验中，使用四维数据，手动设置故障节点、事件节点和测量误差节点的数据集，使故障节点的数据长期保持不变；在一定范围内的所有事件节点的数据与相同采样

时间的正常节点的数据显著不同，并且变化过程趋于一致。测量误差节点的数据保持不变，但在同一采样时间与正常节点的数据存在显著偏差。实验参数信息如表1所示，其中每个参数值乘以100。

由于数据流的统计特征主要取决于数据滑动窗口的长度，因此，通过研究类似实验得知，当窗口长度增加到120左右时，数据的均方差基本趋于稳定。故实验中的数据滑动窗口的长度取120。

实验使用ZigBee协议栈来实现节点之间的通信，路由器节点负责收集传感器数据并定期将数据发送给协调器节点，路由器设备还将自动转发网络中其他节点的数据。协调器节点接收路由器报告的数据，通过串行端口将其上传到中央基站(PC)，并在PC上完成检测操作。

## 2.2 实验说明

将数据采集周期设定为25 s，并每隔5 s从实验数据中随机抽取1 000条数据构成数据序列，按照3:1的比例将其划分为训练数据与实验数据，在实验数据集中添加带有高斯噪声的异常数据值，标准化所有数据，将其限定在[0,1]区间内，从[0,1]区间的规律分布中随机生成需要引入的异常数据。

对于文中提出的邻域链算法，其相关参数设置如下：$k$邻域数为10；在相似性度量下的最近局部密度为0.01；数据点的运算复杂度$v$=0.002；邻域可达性代价与邻域可达性跨度均取0.5。

## 2.3 数据异常检测结果

### 2.3.1 传感器异常信号的检测

本次实验数据来自节点上的传感器所采集的数据，单轴采样频率为200 Hz，ADC转换采样用8位低功耗微处理器ATmega128L的ADC0口，10位采样精度。共采集数据2 175个数据。每个数据占2个字节。图3为原始的传感器含故障信号的恢复波形图。

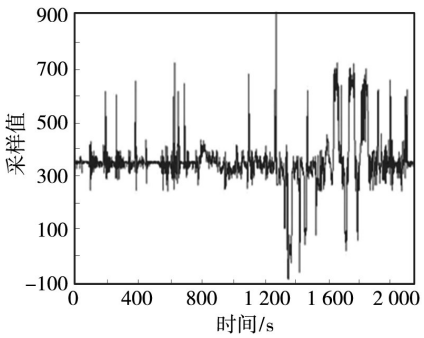



图3 传感器原始数据恢复

通过本文方法得到的异常检测数据如图4所示。

为体现本文方法的故障信号检测性能，将文献[1]BIRCH聚类算法(方法1)、文献[2]大数据技术(方法2)作为文中方法的对比。其他2种方法的异常信号检测结果如图5所示。

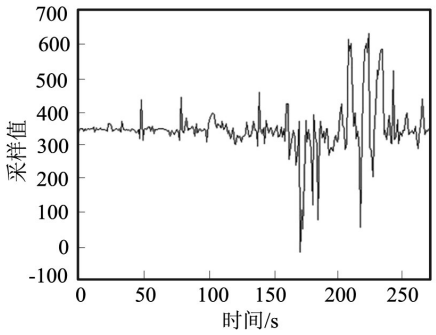
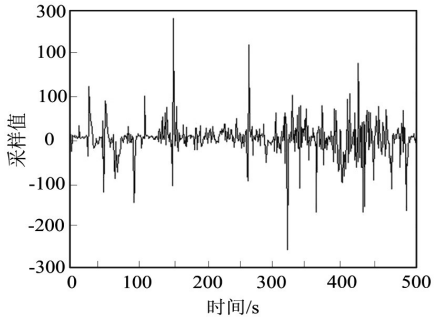
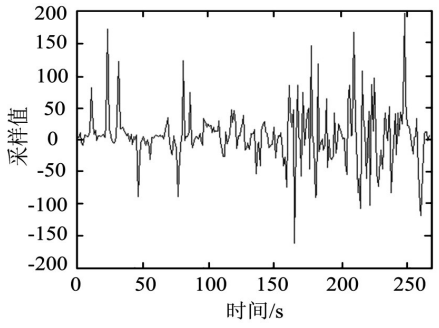



图4 本文方法下得到的异常检测数据




(a) 文献[1]BIRCH聚类算法




(b) 文献[2]大数据技术

图5 对比方法检测结果

通过图4和图5的对比可以看出，本文方法得到的检测结果和预设的故障信号最为类似。虽然对比方法在一些节点上也能够检测出故障，但从整体区域看，很多地方出现了错误检测的地方，本文方法在阈值突破的区域基本上都能和原始故障信号区域重合，优势较为明显。

### 2.3.2 异常检测率对比

为了对本文提出的方法进行定性分析，引入节点数据异常检测率来衡量本文方法的可靠性，其计算公式如下：

$$ETPR(G)=\frac{G\cap E_v}{E_v} \tag{14}$$

式中，$G$表示原始数据中真正异常数据；$E_v$表示检测方法检测出的异常数据。

以上述实验数据集为初始输入，利用本文方法对该数据集进行异常数据检测，结果如图6所示。

通过分析图6可知，在不同数据异常比例条件下，对

于原始数据中不同维度的异常数据，本文方法得到的检测准确率均较高，在80%以上，这说明，本文提出的数据异常检测方法具有较高的检测准确率，可以获得较优的检测效果。

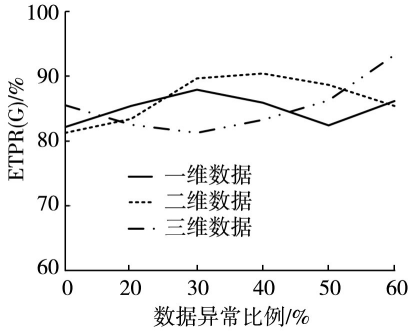



图6 数据异常检测结果

## 2.4 检测误检率对比实验分析

为进一步体现本文方法的综合检测性能，将文献[1]BIRCH聚类算法(方法1)、文献[2]大数据技术(方法2)作为文中方法的对比，基于实验数据集，比较不同方法的误检率。误检率指的是检测方法将正常数据误判为异常数据的比例，可以评价算法的检测准确性。对比结果如图7所示。

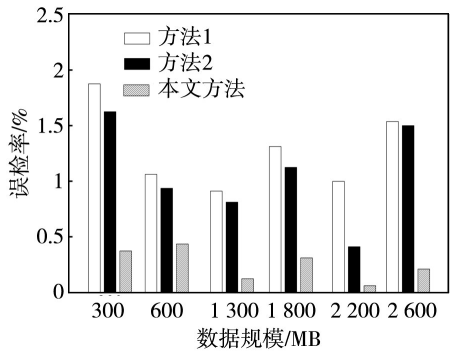



图7 不同数据异常检测方法的误检率结果对比

如图7所示，通过利用三种检测算法对原始数据集进行数据异常检测，结果表明，本文算法对于不同数据规模，具有更低的误检率。方法1通过采用时间序列数据对错误节点进行识别，由于忽略了时间复杂度的影响而出现误判情况；方法2没有考虑到传感器节点采集的数据的分布规律，因此该算法易将边缘的正常数据误报为异常数据。由此可以说明，本文提出的基于邻域链的传感器网络数据异常检测方法具有更高的检测准确率，检测性能较好。

## 3 结束语

本文提出了一种基于邻域链的传感器网络数据异常检测方法。该方法通过验证异常数据来源与计算网络两节点间的数据相似度，结合邻域阈值输出异常数据值。实验结果表明，提出的方法对于异常数据的检测准确率较高，可为传感器网络数据的处理与应用提供重要参考。

**参考文献**：

[1] 赵娇.基于BIRCH聚类算法的高维传感器数据异常检测[J].传感技术学报，2022，35(12)：1686-1690.

[2] 马海昕.基于大数据技术的网络异常检测方法[J].电子元器件与信息技术，2022，6(7)：56-59.

[3] 杨明润，郭星锋，黄元峰，等.基于数据压缩的WSN水质异常数据检测算法[J].电视技术，2022，46(5)：204-207.

[4] 周显春.ARMA融合CNN-LSTM的传感器流数据异常检测方案[J].国外电子测量技术，2022，41(4)：55-61.

[5] 王礼霞，郤清清.基于高阶马尔可夫链的无线传感器网络异常节点检测[J].黑龙江工业学院学报(综合版)，2021，21(8)：93-97.

[6] 蔡兴旭，刘以安，肖颖.基于改进LSTM的桥梁传感器异常数据的检测方法[J].计算技术与自动化，2021，40(2)：8-11，20.

[7] 李晨，王布宏，田继伟等.基于LSTM-OCSVM的无人机传感器数据异常检测[J].小型微型计算机系统，2021，42(4)：700-705.

[8] 张振军.基于数据筛选的无人机测绘数据异常检测[J].西华大学学报(自然科学版)，2022，41(4)：66-71.

[9] 张敏，方健，王红斌，等.配电机房的传感器异常数据检测分析[J].集成电路应用，2021，38(1)：164-165.

[10] 张鸿雁.基于聚类分析的网络数据流异常检测方法[J].长江信息通信，2022，35(12)：54-56.

[11] 郝美薇，张驰，贺小刚，等.基于电力大数据的用户用电量多维度自适应分析系统[J].自动化技术与应用，2023，42(3)：179-183.

[12] 窦琛琛，袁昊，金鑫.基于电力大数据的电网企业认责溯源系统设计[J].自动化技术与应用，2023，42(8)：20-23，33.

作者简介：晁永兰(1980-)，女，本科，副教授，研究方向：计算机应用技术。