文章编号:1671-4598(2023)10-0233-07 DOI:10.16526/j.cnki.11—4762/tp.2023.10.036 中图分类号:TP393 文献标识码:A

WSN 中基于超椭圆判决边界的异常检测的动态建模

方小明，刘艳梨
(江苏安全技术职业学院 电气工程学院，江苏 徐州 232001)

摘要：为了实现无线传感器网络的动态数据流环境中的异常检测，提出了一种称之为动态数据捕获异常检测的迭代超椭圆判决边界方法；具体实现是建立起异常检测超椭圆模型，每个节点基于到当前时间为止的测量值来调整其超椭圆模型；当边界参数变化较小时，动态数据捕获异常检测算法终止，最终收敛到覆盖正常和异常测量值的超椭圆边界；为了提高模型对监测环境中数据变化的跟踪能力，提出了一种采用遗忘因子并结合滑动窗口的基准估计和有效 N 跟踪的方法来提高模型在非平稳环境中的跟踪能力，从而实现对数据真实流属性的捕捉；仿真实验结果表明，提出的动态建模方法相比于目前先进的静态建模方法，不仅具有更高的准确性和异常检测能力，而且具有更强的数据变化的跟踪能力和检测能力。

关键词：无线传感器网络；异常检测；超椭圆边界；迭代；数据跟踪；准确性

Dynamic Modeling of Anomaly Detection Based on Superellipse Decision Boundary in WSN

FANG Xiaoming, LIU Yanli
(School of Electrical Engineering, Jiangsu College of Safety Technology, Xuzhou 232001, China)

Abstract: In order to realize anomaly detection in dynamic data flow environment of wireless sensor networks, an iterative hyperelliptic boundary decision method called as dynamic data capture anomaly detection is proposed. The concrete implementation is to establish the anomaly detection hyperellipse model, each node adjusts its hyperellipse model based on the measured values up to the current time. When the boundary parameter changes are small, the dynamic data capture anomaly detection algorithm terminates, and finally converges to the hyperellipse boundary covering the normal and abnormal measurements. In order to improve the tracking ability of the model to the data changes in the monitoring environment, a forgetting factor combined with the benchmark estimation of sliding window and effective N tracking method is proposed to improve the tracking ability of the model in non-stationary environment, so as to capture the real data flow attributes. The simulation results show that, compared with the advanced static modeling methods, the proposed dynamic modeling method not only has higher accuracy and anomaly detection ability, but also has stronger tracking and detection ability of data changes.

Keywords: wireless sensor network; anomaly detection; hyperelliptic boundary; iteration; data tracking; accuracy

0 引言

在有线感知基础设施部署过于昂贵或不能实现的环境中，无线传感器网络（WSN，wireless sensor network）为监测和数据收集提供了一个成本高效的平台[1-2]。WSN 由一组节点构成，每个节点都配备一组感知设备。在每个节点上安装不同的感知元件（如温度和湿度传感器），使得 WSN 能够收集大量多维的和相关的样本。WSN 的一个重要挑战是检测由周围环境中感兴趣的事件或节点故障引起的异常测量值。在节点上发现异常测量值，使得我们可以通过减少网络上原始数据的通信，节省无线节点的有限资源。为了检测异常，需要对节点的行为进行建模。

人们提出了各种数据挖掘方法来建立节点的行为模型。在分散式方法中，WSN 中的每个节点都建立一个自身正常行为的局部模型，将局部模型的参数转发到基站或簇头，然后根据局部模型计算全局模型。近年来，人们提出了许多采用这种方法的不同数据建模方法。然而，这些模型大多为静态模型，不能适应环境中的变化。此外，这些模型的准确性依赖于初始训练周期的正确选择。如果初始训练周期不能很好地代表将来的测量值，模型就是失败的。因此，重要的问题是如何持续学习非平稳环境中的行为模型，即如何检测非平稳环境中的异常事件。

异常检测是 WSN 中一个活跃的研究课题。在 WSN 中，异常检测技术已应用于许多方面，包括入侵检测、事件检测和质量保证[3-5]。在这些应用中，有许多因素会影响异常检测的使用，如传感器的移动、环境条件（有利的或不利的）、环境的动态性和能量约束。因此，异常检测技术在实际应用中的一个关键问题是如何将其推广到具有动态变化

收稿日期:2023-04-04； **修回日期**:2023-05-10。
基金项目:国家自然科学基金项目(51975277)。
作者简介:方小明(1982-),男,硕士,讲师。
引用格式:方小明,刘艳梨. WSN 中基于超椭圆判决边界的异常检测的动态建模[J]. 计算机测量与控制,2023,31(10):233-239.

的在线数据流中。

文献［6］提出了一类支持向量机（SVM，support vector machine）模型来发现 WSN 数据中的异常现象。这种方法主要假设所有的训练数据都可以在传感器上获得，并且训练以批处理的方式进行。尽管这些方法可以为正常数据提供良好的决策边界，但它们对于每个传感器有很高的计算开销；文献［7］提出了一种基于长短期记忆网络自编码（LSTM-Autoencoder）的网络流量异常检测方法，将真实网络流量从数据包和会话流级别两方面提取数据特征，采用离散小波变换（DWT，discrete wavelet transform）分解原始特征向量得到更高维特征，用已训练的 LSTM-Autoencoder 模型对训练数据进行重构，通过分析重构误差分布确定检测阈值。该方法的主要缺点首先是训练对数据中的噪声敏感，其次很难理解是什么触发了报告的异常；文献［8－9］把超椭圆边界用来建模系统的正常行为与批处理训练。这种方法允许训练数据中存在噪声，并向用户报告个别异常。然而，其超椭圆边界是在一个训练周期上计算的，而且要求节点在训练期间将测量值保存在存储器中，在训练结束时所有的测量值以批处理方式处理。尽管这些方法在计算上是高效的，但它们不能适应环境中的变化，是一种静态模型。作为比较，本文将这种方法称为静态数据捕获异常检测（SDCAD，static data capture anomaly detection）；文献［10］提出了一种基于四分之一超球 SVM 算法的异常数据检测方法，利用从传感器节点中收集到的原始数据建立支持向量机预测模型，并结合粒子群算法找出最佳参数，然后利用最佳参数对原本的模型进行优化；文献［11］提出了一种新的时间－空间－属性单类超球面支持向量机来建模 WSN 中的异常事件检测问题，并提出了在线和部分在线离群点检测算法。但部分在线离群点算法在训练和更新时需要大量的计算；文献［12］提出了一种累积和（CS，cumulative sum）算法来检测网络异常。尽管基于 CS 的异常检测算法计算效率高，但基于其阈值的检测机制通常不能准确地建模正常行为；文献［13］提出了数据流自回归模型的迭代估计，并采用 CS 作为在线异常检测；对于多维数据中的异常检测是著名的批（子群）处理技术，它采用马氏距离[9,14-15]进行异常检测；文献［16］提出了一种基于改进压缩感知（CS，compressed sensing）重构算法和智能优化 GM（1，1）的 WSN 异常检测方法。首先通过建立双层异质 WSN 异常检测模型，并采用压缩感知技术对上层观测节点收集到的下层检测节点温度测量数据进行处理，同时结合温度数据稀疏度未知特点，构造有效的稀疏矩阵和测量矩阵，并重新定义测量矩阵正交变换预处理策略，使得 CS 观测字典满足约束等距条件；其次，重新定义离散蜘蛛编码方式，蜘蛛种群不断协同进化，以获得稀疏结果中非零元素的位置信息，利用最小二乘法得到非零元素的幅度信息，实现对未知数量检测节点数据的精确重构，在此基础上采用蜘蛛种群迭代进化得到优化后 GM（1，1）的参数序列，通过检测参数序列的相关阈值来判定节点是否发生异常；文献［17］提出了一种基于传感器网络时间序列数据的检测方法，方法利用传感器采集的 K 个正常数据的中位数建立枢轴量，构造置信区间，并提出了一种计算数据区间差异度的方法来判断发生异常的来源。实验结果表明，该方法对传感器网络的异常数据检测率保持在 98％以上，误报率保持在 0.5％以下，具有一定的实用性；文献［18］提出一种基于平衡迭代规约层次聚类（BIRCH，balanced iterative reducing and clustering using hierarchies）的 WSN 流量异常检测方案。该方案在扩充流量特征维度的基础上，利用 BIRCH 算法对流量特征进行聚类，并通过设计动态簇阈值和邻居簇序号优化 BIRCH 聚类过程来提高算法的聚类质量和性能鲁棒性。进一步设计了基于拐点的综合判决机制，结合预测，聚类结果对流量进行异常检测，以保证方案的检测准确性；为了提高无线传感网络的鲁棒性，针对目前的网络漏洞检测方法无法计算出相邻节点的相对位置信息，存在无线传感器网络漏洞检测误差大的问题，文献［19］提出了先利用覆盖漏洞发现算法组建传感器极点坐标，获取最相近节点间位置信息，计算出任意节点被其最相近节点覆盖的边缘弧信息序列，然后得到对应传感器节点间需要增加的新传感器数量，从而实现无位置信息的无线传感器网络漏洞检测方法；文献［20］针对 WSN 中传感器自身安全性低、检测区域恶劣及资源受限造成节点采集数据异常的问题，提出了一种基于图信号处理的 WSN 异常节点检测算法。算法首先依据传感器位置特征建立一近邻图信号模型，然后基于图信号在低通滤波前后的平滑度之比构建统计检验量，最后通过统计检验量与判决门限实现异常节点存在性的判断。通过在公开的气温数据集与 PM2.5 数据集上的仿真验证结果表明，与基于图频域异常检测算法相比，在单个节点异常情况相同条件下，所提出的算法检测率提升了 7 个百分点。在多个节点异常情况相同条件下，其检测率均达到 98％，并且在网络节点异常偏离值较小时仍具有较高的检测率。

为了实现 WSN 中动态数据流环境的异常检测，本文提出了一种迭代方法来建立超椭圆判决边界，其中每个节点基于到当前时间为止的测量值来调整其超椭圆模型，本文将提出的这种方法称为动态数据捕获异常检测（DDCAD，dynamic data capture anomaly detection）。当边界参数变化较小时，DDCAD 算法终止；同时，还提出了一种遗忘因子方法来提高模型在非平稳环境中的跟踪能力；仿真实验结果表明，提出的方法通过适应环境中的变化，在非平稳环境中比现有的批处理方法能够获得更高的准确性，更适合于实际应用。

1 动态数据流环境下的迭代超椭圆边界算法

首先给出描述异常检测超椭圆模型所需的定义。令 $X_k=\{x_1, x_2, \cdots, x_k\}$ 为一个 WSN 中的一个节点在时刻 $\{t_1, t_2, \cdots, t_k\}$ 的前 k 个样本，其中每个样本是 R^d 中的一个 $d\times 1$ 向量。向量中的每个元素表示由节点测量的感兴

趣的属性，如温度和相对湿度。$\boldsymbol{X}_k$ 的样本均值 $\boldsymbol{m}_k$ 和样本协方差 $\boldsymbol{S}_k$ 计算如下：

$$\boldsymbol{m}_k = \frac{1}{k}\sum_{j=1}^{k}\boldsymbol{x}_j \tag{1}$$

$$\boldsymbol{S}_k = \frac{1}{k-1}\sum_{j=1}^{k}(\boldsymbol{x}_j-\boldsymbol{m}_k)(\boldsymbol{x}_j-\boldsymbol{m}_k)^T \tag{2}$$

以具有协方差矩阵 $\boldsymbol{S}_k$ 的、以 $\boldsymbol{m}_k$ 为中心的有效半径 t 的超椭圆定义为：

$$e_k(\boldsymbol{m}_k,\boldsymbol{S}_k^{-1};t) = \{\boldsymbol{x}\in R^d \mid (\boldsymbol{x}-\boldsymbol{m}_k)^T\boldsymbol{S}_k^{-1}(\boldsymbol{x}-\boldsymbol{m}_k)\leqslant t^2\} \tag{3}$$

式中，$(\boldsymbol{x}-\boldsymbol{m}_k)^T\boldsymbol{S}_k^{-1}$（$\boldsymbol{x}-\boldsymbol{m}_k$）为 $\boldsymbol{x}$ 到 $\boldsymbol{m}_k$ 的马氏距离，$\boldsymbol{S}_k^{-1}$ 为 e_k 的特征矩阵。

超椭圆 e_k 的边界定义为：

$$\delta_{e_k}(\boldsymbol{m}_k,\boldsymbol{S}_k^{-1};t) = \{\boldsymbol{x}\in R^d \mid (\boldsymbol{x}-\boldsymbol{m}_k)^T\boldsymbol{S}_k^{-1}(\boldsymbol{x}-\boldsymbol{m}_k) = t^2\} \tag{4}$$

采用 $t^2=(\chi_d^2)_p^{-1}$（即具有 d 自由度的卡方统计量的倒数）和 $p=0.98$，在数据具有正态分布的假设下，会得到一个覆盖至少 98% 的数据的超椭圆边界，故全文都采用 t^2 这个值。

定义 1：将关于 e_k 的单点一阶异常定义为在其外面的任意数据向量 $\boldsymbol{x}\in R^d$，即对于 e_k 来说：

$$\boldsymbol{x}\text{ 为异常}\Leftrightarrow(\boldsymbol{x}-m_k)^T\boldsymbol{S}_k^{-1}(\boldsymbol{x}-\boldsymbol{m}_k) > t^2 \tag{5}$$

已知节点在 t_k 的样本，要处理节点上的下一个样本。在 t_{k+1}，我们记录测量向量 $\boldsymbol{x}_{k+1}\in R^d$。首先，用式（5）来测试 $\boldsymbol{x}_{k+1}$，然后用它来增大 e_k。如果 $\boldsymbol{x}_{k+1}\notin e_k$，就声明它是一个异常，并将它发送给基站进行进一步处理。特征矩阵迭代更新公式为：

$$\boldsymbol{m}_{k+1} = \boldsymbol{m}_k + \frac{1}{k+1}(\boldsymbol{x}_{k+1}-\boldsymbol{m}_k) \tag{6}$$

$$\boldsymbol{S}_{k+1}^{-1} = \frac{k\boldsymbol{S}_k^{-1}}{k-1}\left[\boldsymbol{I}-\frac{(\boldsymbol{x}_{k+1}-\boldsymbol{m}_k)(\boldsymbol{x}_{k+1}-\boldsymbol{m}_k)^T\boldsymbol{S}_k^{-1}}{\frac{k^2-1}{k}+(\boldsymbol{x}_{k+1}-\boldsymbol{m}_k)^T\boldsymbol{S}_k^{-1}(\boldsymbol{x}_{k+1}-\boldsymbol{m}_k)}\right] \tag{7}$$

我们采用 $\boldsymbol{S}^{-1}=I$（其中 I 是单位阵），而不采用从前几个样本获得的估计值来初始化迭代方法，因为前几个样本通常会产生一个奇异的样本协方差矩阵。

我们用正常和异常测量值来增大 e_k。假设大部分数据都是正常的，因此可以抵消用异常测量值进行更新的任何不希望的影响。然而，也可以设计更复杂的方法，以不同的方式处理异常。这时应考虑异常是否是环境中的正常变化（漂移）。这类分析需要额外的输入来确定异常的类型。

令 $\boldsymbol{X}_n=\{x_1, x_2, \cdots, x_n\}$ 为一个节点上的观测序列。DDCAD 椭圆 $\{e_k(\boldsymbol{m}_k, \boldsymbol{S}_k^{-1}; t) \mid 1\leqslant k\leqslant n\}$ 是明确的。如果 $(\boldsymbol{m}, \boldsymbol{S})$ 为样本均值和 $cov(\boldsymbol{X}_n)$，则 $e_{ns}(\boldsymbol{m}, \boldsymbol{S}^{-1}; t)$ 就是批处理静态（SDCAD）方法[9]所采用的椭圆，其中下标 s 表示静态。当采用节点上的每个输入来更新 DDCAD 椭圆时，$e_n(\boldsymbol{m}_n, \boldsymbol{S}_n^{-1}; t)\cong e_{ns}(\boldsymbol{m}, \boldsymbol{S}^{-1}; t)$。也就是说，序列 $\{e_k\}$ 应当在非常接近于静态椭圆 e_{ns} 终止。

图 1 为包含了当 $k\to n$ 时 $\{e_k\}$ 行为的图形。图 1（a）所示的数据是 $n=818$ 对（温度，湿度）散点的集合，这些散点来自后面仿真实验中的 DS1；图 1（c）所示为序列 $\{e_k: 1\leqslant k\leqslant 818\}$ 中的几个 DDCAD 椭圆。虚线椭圆为序列中的最终椭圆，可见，$\{e_k\}\to e_{818}$ 的收敛非常明显；图 1（b）是在 $k\to n=818$ 时 $\boldsymbol{S}_k^{-1}$ 的两个特征值。$S_{818}^{-1}\cong S_{818s}^{-1}$ 的特征值分别为 $\{\alpha_{M,818}=11.88, \alpha_{M,818s}=12.09\}$ 和 $\{\alpha_{m,818}=0.39, \alpha_{m,818s}=0.39\}$。图 1（b）表明，$\boldsymbol{S}_k^{-1}$ 的较小特征值在 $k=75$ 达到其终值，而 $\boldsymbol{S}_k^{-1}$ 的较大特征值与 $\alpha_{M,818s}$ 的偏差非常大，在 $k=327$ 达到最大值，$|\alpha_{M,818s}-\alpha_{M,818}|=189.41$。图 1（a）中垂直线左侧的点对应于前 320 个样本，这在 DDCAD 中导致最初非常窄的椭圆，如图 1（b）中较大的特征值的高值所示。图 1（b）表明，序列 $\{e_k\}$ 在大约 $k=600$ 时开始逼近 e_{818s}，大约占输入样本的 75%。

现在简单讨论一下极限情况。在式（6）和（7）中，当 $k\to\infty$ 时，$\|\boldsymbol{m}_{k+1}-\boldsymbol{m}_k\|\to 0$，同样地，$\|\boldsymbol{S}_{k+1}^{-1}-\boldsymbol{S}_k^{-1}\|\to 0$。式（7）中分母中的因子 $(k^2-1)/k$ 在 $k\to\infty$ 时减慢了收敛速度。由于我们处理的观测集总是有限的，故收敛速度将比极限更重要。

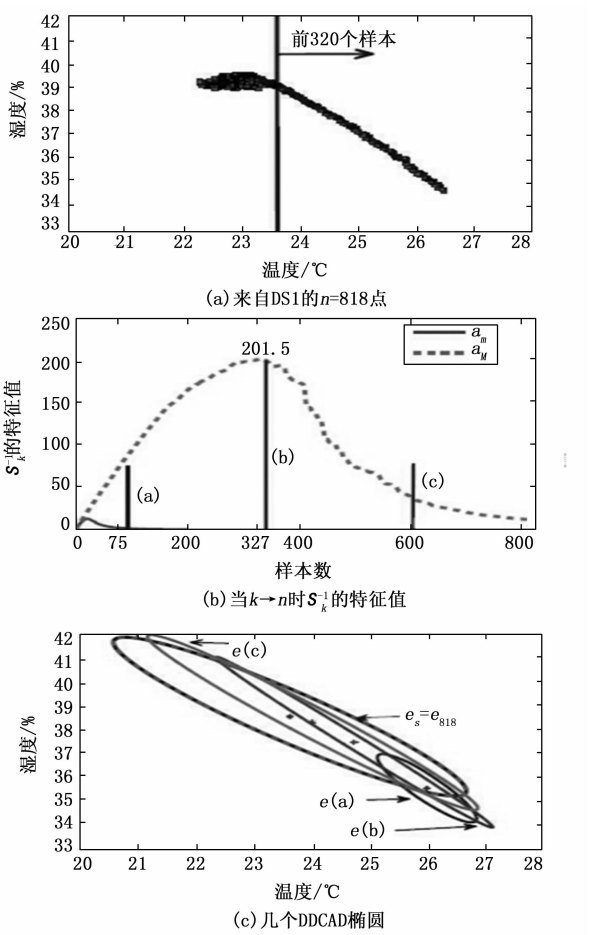

图 1　DDCAD 序列 e_k 收敛到其最终状态 $e_{818}=e_s$

2　采用遗忘因子的跟踪能力

为了使 DDCAD 算法能够跟踪监测环境中的数据变化，我们为旧的测量值引入遗忘因子。通过引入遗忘因子 $0<\lambda$

$<l$ 定义 k 个样本周期的加权样本协方差，给出来自于之前 j 个样本的测量值的权值 λ^j。因此，式（8）所示的指数移动平均可以用来更新 $k>2$ 的样本均值：

$$\boldsymbol{m}_{k+1,\lambda}=\lambda\boldsymbol{m}_{k\lambda}+(1-\lambda)x_{k+1} \tag{8}$$

对于 k 个样本，采用指数遗忘因子 λ 的加权样本协方差为：

$$\boldsymbol{S}_{k\lambda}=\frac{1}{k-1}\sum_{j=1}^{k}(\boldsymbol{x}_j-\boldsymbol{m}_{k\lambda})(\boldsymbol{x}_j-\boldsymbol{m}_{k\lambda})^T\lambda^{k-j} \tag{9}$$

首先要找到考虑遗忘因子的迭代协方差矩阵更新公式，然后得出特征矩阵的迭代更新公式。通过整理式（9），可以基于上一步的协方差矩阵加上一个更新值，写出 $\boldsymbol{k}+1$ 时刻的协方差矩阵的更新公式。式（10）为协方差矩阵的一步更新：

$$\boldsymbol{S}_{k+1,\lambda}=\frac{\lambda(k-1)}{k}\boldsymbol{S}_{k\lambda}+\frac{1}{k}\sum_{j=1}^{k}(\boldsymbol{x}_{k+1}-\boldsymbol{m}_{k+1,\lambda})(\boldsymbol{x}_{k+1}-\boldsymbol{m}_{k+1,\lambda})^T \tag{10}$$

将式（10）中的 $\boldsymbol{m}_{k+1}$ 替换为式（8）中 $\boldsymbol{m}_{k+1}$ 可得：

$$\boldsymbol{S}_{k+1,\lambda}=\frac{\lambda(k-1)}{k}\boldsymbol{S}_{k\lambda}+\frac{\lambda^2}{k}\sum_{j=1}^{k}(\boldsymbol{x}_{k+1}-\boldsymbol{m}_{k\lambda})(\boldsymbol{x}_{k+1}-\boldsymbol{m}_{k\lambda})^T \tag{11}$$

为了计算特征矩阵的更新公式，我们用矩阵逆引理式（12）来求两个矩阵的和的逆。假设 $\boldsymbol{E}$ 是可逆的且 $\boldsymbol{B}$ 是一个方阵。注意，在本文中，$\boldsymbol{E}$ 是一个数，$\boldsymbol{C}$ 和 $\boldsymbol{D}$ 是向量。将这个引理应用到式（11）中，经过重新整理，得到式（13）：

$$(\boldsymbol{B}+\boldsymbol{CED})^{-1}=\boldsymbol{B}^{-1}-\boldsymbol{B}^{-1}\boldsymbol{C}(\boldsymbol{E}^{-1}+\boldsymbol{DB}^{-1}\boldsymbol{C})^{-1}\boldsymbol{DB}^{-1} \tag{12}$$

$$\boldsymbol{S}_{k+1,\lambda}^{-1}=\frac{k\boldsymbol{S}_{k\lambda}^{-1}}{\lambda(k-1)}\times\left[\boldsymbol{I}-\frac{(\boldsymbol{x}_{k+1}-\boldsymbol{m}_{k\lambda})(\boldsymbol{x}_{k+1}-\boldsymbol{m}_{k\lambda})^T\boldsymbol{S}_{k\lambda}^{-1}}{\frac{k-1}{\lambda}+(\boldsymbol{x}_{k+1}-\boldsymbol{m}_{k\lambda})^T\boldsymbol{S}_{k\lambda}^{-1}(\boldsymbol{x}_{k+1}-\boldsymbol{m}_{k\lambda})}\right] \tag{13}$$

把用式（8）和式（13）对 e_k 的更新序列称为遗忘因子 DDCAD（FFDDCAD，forgetting factor DDCAD）。

遗忘因子 λ 应当接近于 1。在估计理论文献中，λ 的建议范围是［0.9 0.99］。本文采用 $\lambda=0.99$，并建议 FFDDCAD 的 λ 范围为［0.99 0.999］。与先前的测量值相比，这时增加了当前测量值的重要性，但对于非常大的 k，迭代算法变得不稳定。图 2 为对于实际数据集 DS2 的这种影响。可见，在样本 5 000 后，较大的特征值是非常不稳定的；当 $k\to\infty$ 时，式（13）括号中的值趋近于 I，则特征矩阵更新趋近于 $\boldsymbol{S}_{k\lambda}^{-1}/\lambda$。因此，当 k 增大时，新测量值的影响变小，且 $\boldsymbol{S}_{k\lambda}^{-1}/\lambda$ 控制特征矩阵的变化。为了解决这个问题，通过引入一种基于滑动窗口的基准方法和一种称为有效 N 跟踪方法的近似方法来限制 k 的增长。

2.1 采用滑动窗口的基准估计

为了限制 FFDDCAD 更新公式中 k 的增长，可以在大小为 w 的滑动窗口上采用 FFDDCAD。为了提供比较基准，从窗口开始重新计算总体估计，以便在每次输入后得到准确的 FFDDCAD 椭圆。对于在线算法来说，尽管这种方法的计算效率不高，但它提供了采用主动测量值的超椭圆边界的精确值（即在滑动窗口中的测量值），并用作基准，作为比较在计算中限制大 k 效应所提出的方法。

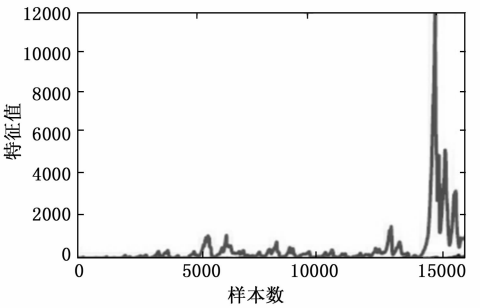

图 2 采用 FFDDCAD 在每次更新后特征矩阵的特征值

2.2 有效 N 跟踪方法

在这种方法中，为了解决跟踪 k 较大的问题，当 $k\geqslant n_{eff}$ 时，我们简单地用不变的 n_{eff} 来代替式（13）中的 k。其思想是在 $k\geqslant n_{eff}$ 后，分配给数据样本的权重趋于 0，即 $\lambda^k\cong 0$，因此相应的样本几乎被完全遗忘。在本文中，取 $n_{eff}=3\tau$，其中 $\tau=1/(1-\lambda)$ 为具有指数遗忘因子 λ 的迭代算法的记忆范围。基准方法和有效 $\boldsymbol{N}$ 跟踪方法的示意如图 3 所示。方框所示为在椭圆边界计算中所考虑的样本。在有效 N 跟踪方法中，旧样本的权重按指数下降。

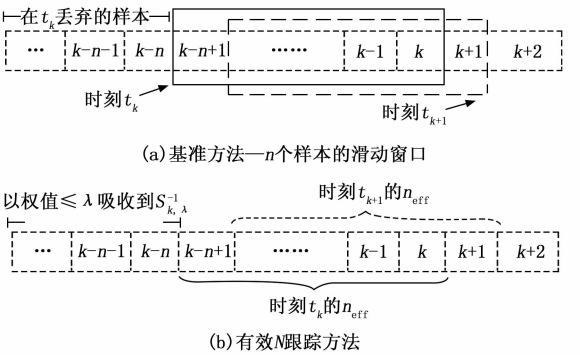

图 3 在样本 k 和 $k+1$ 的基准方法和有效 N 方法的示意图

2.3 计算复杂度讨论

在计算复杂度方面，SDCAD、DDCAD 和 FFDDCAD 都需要对数据进行一次遍历，所以它们的计算复杂度都随 n 线性增长，有渐近复杂度 $O(nd^2)$；迭代方法（DDCAD 和 FFDDCAD）以在线处方式处理数据，具有恒定的存储复杂度，而 SDCAD 方法的存储需求随 n 线性增长；采用有效 N 跟踪的 FFDDCAD 准确性和效率使得其适合于在线流数据分析，特别是在 WSN 中。

3 仿真实验

本节首先给出在评价不同方法时采用的数据集，然后比较提出的采用有效 N 方法和基准方法的 FFDDCAD，并比较了两种 FFDDCAD 方法在合成数据集上的检测率和误报率。在合成数据集中，将［−10 10］上的均匀噪声随机加入到 1%的样本中，并将这些样本标记为异常，而其他剩余的样本视为正常。另一种比较方法是基于与提出的基准方法的偏差而引入的，这种方法不需要标记数据集，因此

允许采用实际的数据集进行比较。接下来，我们比较了 FFDDCAD 相比于 SDCAD 在异常检测上的效果。最后，我们比较了本文提出的采用有效 N 方法的 FFDDCAD 与和文献［13］中提出的变化检测技术。

3.1　数据集

采用 3 个数据集来评价本文提出的异常检测迭代模型，并将其与现有的静态方法进行比较。第一个数据集（称为 DS1）由某院校物联网研究实验室的 54 个传感器收集的测量数据构成；第二个数据集（称为 DS2）是从部署在某市城市道路之间的 23 个交通传感器收集的数据；第三个数据集（称为 DS3）是由部署在某市小镇的森林中的 16 个传感站收集的数据。图 4 为 3 个数据集的散点图。

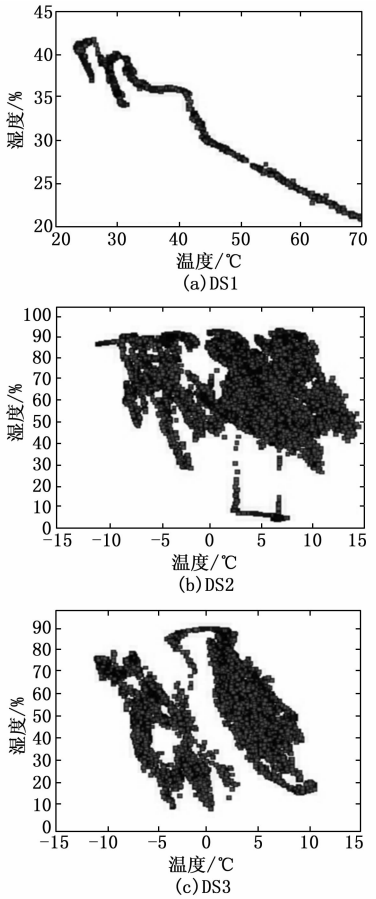

图 4　用于评价的数据集的散点图

通过考虑具有不同正态分布 $N(\sum_1, \mu_1)$ 和 $N(\sum_2, \mu_2)$ 的 M_1 和 M_2 两种模式的两个合成数据集 SDS1 和 SDS2 如图 5 所示。模式 M_1 和 M_2 的参数值如表 1 所示。M_1 为初始模式，M_2 为最终模式。M_1 的变换如下。

首先，从 M_1 中抽取 500 个样本 $\{k=1\cdots500\}$。当协方差矩阵中的每个个体值和平均值在 10 个相等的步骤中变化时，采样将继续进行。在第一个步骤后，从新的正态分布中取 200 个样本 $\{k=501, \cdots, 700\}$。在每个新的步骤后，将 200 多个样本添加到数据集中。最后一步结束于模式 M_2。在第一个数据集 SDS1 中，步骤比第二个数据集 SDS2 要小得多。通过这种方式，可以测试步骤大小如何影响跟踪方法。在图 5 中，$t_2=(\chi_d^2)_{0.98}^{-1}$的椭圆位于 M_1 和 M_2 处。点表示数据样本，星号表示每个正态分布 1%的样本，它们受到［−10 10］的均匀噪声的干扰，这些样本标记了真实的异常，而其余的样本标记为正常。这种标记用于计算这些数据集的检测率和误报率。

表 1　用于生成合成数据集的两个正态分布的参数

SDS1	SDS2
$M_1\ \sum_1=\begin{pmatrix}0.679\,7 & 0.166\,9\\ 0.166\,9 & 0.789\,1\end{pmatrix}$	$\sum_1=\begin{pmatrix}10.024\,6 & 1.279\,0\\ 1.279\,0 & 2.163\,0\end{pmatrix}$
$\mu_1=(20,20)$	$\mu_1=(45,42)$
$M_2\ \sum_2=\begin{pmatrix}0.708\,9 & 0.157\,5\\ 0.157\,5 & 0.847\,2\end{pmatrix}$	$\sum_2=\begin{pmatrix}7.690\,9 & 0.664\,6\\ 0.664\,6 & 2.162\,4\end{pmatrix}$
$\mu_2=(5,5)$	$\mu_2=(5,5)$

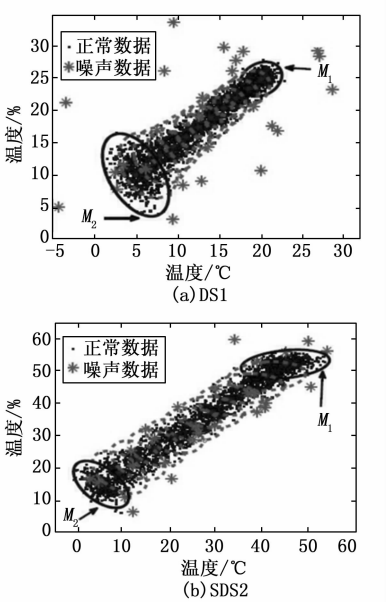

图 5　用于评价的合成数据集的散点图

3.2　DDCAD 的收敛性

运行第 2 节中提出的 DDCAD 方法，并将其与计算整个数据集的协方差矩阵和均值的批处理 SDCAD 方法[9]进行比较。采用焦距（两个椭圆之间的距离的量度）来检查 DDCAD 的最终椭圆边界与 SDCAD 的距离有多近。焦距考虑了两个椭圆的形状和位置，结果如图 6 所示。图 6 中点构成的虚线椭圆为 DDCAD 得到的最终椭圆，实线构成的椭圆为采用 SDCAD 方法得到的最终椭圆；可以看到，DDCAD 算法和 SDCAD 算法的最终结果非常相似，两个最终椭圆之间的焦距即 FD（e_n，e_{ns}）非常小，对于 DS1 为 0.001 6，DS2 为 0.001 4，DS3 为 0.002 4。这些小的距离并没有对最终的边界产生视觉上的明显影响。

3.3　跟踪能力的比较

为了比较提出的跟踪方法，我们首先用合成数据集来比较所提出的异常检测模型的准确性。对于基准方法，考虑 300 个样本的窗口大小。同样，n_{eff}设置为 300 个样本。表 2 所示为两种跟踪方法的检测率和误报率，其中 *DR* 表示

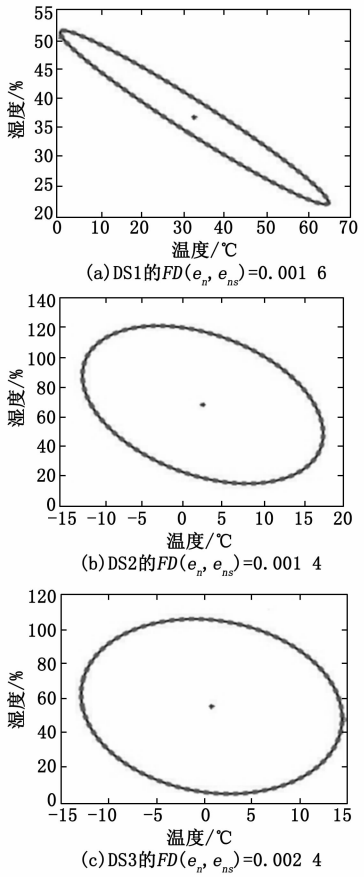

图 6 采用 DDCAD 和 SDCAD 得到的最终椭圆边界与相应的焦距

检测率，*FA* 表示误报率。可见，有效 N 跟踪方法具有与基准方法相当的准确性。这说明有效 N 跟踪方法是基准方法的良好近似，n_{eff}可以代替跟踪迭代公式中的 k 来当解决 k 变大时的不稳定性问题。

表 2 不同跟踪方法在合成数据集上的比较

数据集	基准方法		有效 *N* 跟踪方法	
	DR	*FA*	*DR*	*FA*
SDS1	96	2.4	96	3.1
SDS2	84	2.6	85	3.3

3.4 异常检测能力的比较

我们对两个合成数据集 SDS1 和 SDS2 比较采用有效 N 跟踪方法的 FFDDCAD 和文献［9］提出的 SDCAD 方法，得到的检测率和误报率如表 3 所示。可见，在代表非平稳环境的这两个数据集中，采用有效 N 跟踪方法的 FFDDCAD 比批处理的 SDCAD 方法有更高的准确性。这是因为用于批处理学习的数据不是来自单个分布，所以正态性假设很弱，从而导致模型无法检测异常。

表 3 异常检测能力的比较 %

数据集	SDCAD		FFDDCAD	
	DR	*FA*	*DR*	*FA*
SDS1	55	2.1	96	3.1
SDS2	29	1.6	85	3.3

3.5 数据流中的变化检测

本节比较了本文提出的 FFDDCAD 方法与文献［13］的方法用于数据流的在线异常检测。在数据流分析中，通常采用动态预测模型和残差分析（如 CS）来检测数据流中的变化或异常。为便于比较，我们不直接采用文献［13］的方法，而是采用递归最小二乘（RLS，recursive least squares）迭代建立以湿度作为输入（激励信号）的温度预测的自回归各态历经（ARX，autoregressive eXogenous）模型，阶数为 $n_p=4$，并对其残差应用 CS 来发现数据流的变化。FFDDCAD 的定义是发现单点异常，并且可以很容易地修改来检测变化点。当 FFDDCAD 模型在数据流中发现 n_a 个连续的单点异常时，它可以发出变化信号。

由于在实际的数据集中缺乏基本的真实性，这使得很难解释变化的点。因此，这里我们仅采用 DS1 和两个合成数据集来比较两种方法的结果。ARX/RLS 和 FFDDCAD 在初始状态时都视为是不准确的，因此，延迟采用这两个模型对初始化后的前 $n_d=50$ 样本进行异常检测。注意，在每个变化点之后，模型重置回其初始状态。

图 7 为 ARX/RLS 方法和 FFDDCAD 方法对于数据流变化检测的结果，加号表示变化点；可见，FFDDCAD 方法和 ARX/RLS 方法对于 DS1 的性能是相当的，但采用 FFDDCAD 方法可以检测到更多的变化点，表明 FFDDCAD 方法优于 ARX/RLS 方法；而对于 SDS1，ARX/RLS 方法不能发现模式之间的变化点，而 FFDDCAD 方法可以检测到 5 个变化点；在 SDS2 中，FFDDCAD 方法可以检测所有模式变化，而 ARX/RLS 方法仅检测到一个模式变化；总

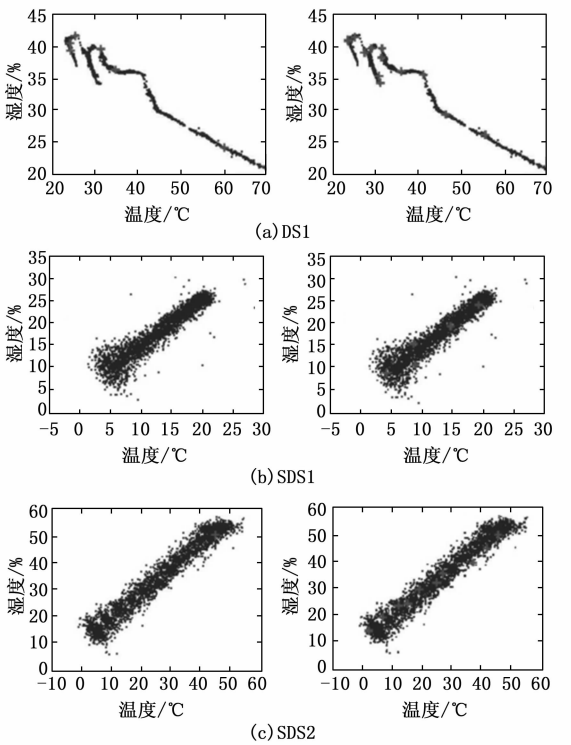

图 7 ARX/RLS（左）与 FFDDCAD（右）对于数据流分析和变化点检测的比较

之，FFDDCAD 方法对于数据流变化的检测始终优于 ARX/RLS 方法。

4　结束语

本文针对 WSN 中的异常检测提出了一种迭代模型，其迭代性使得它更适合于流数据分析；此外，在模型中引入遗忘因子，使其适合于非平稳环境；评价表明，提出的方法可以密切跟踪环境中的变化，在非平稳环境中能获得比批处理方法更好的准确性。同时在数据流的异常检测中，本文提出的采用遗忘因子的 FFDDCAD 可以更好地检测环境中的变化，计算复杂度也比目前先进的方法更低。

参考文献：

[1] 陈志聪，李清华，吴丽君，等．面向桥梁结构健康监测的 MP-MR-MC 无线传感器网络数据收集方法［P］．CN108769944A. 20181106.

[2] 吴　洁．无线传感器网络中移动数据收集优化算法研究［D］．南京：南京邮电大学，2020.

[3] 郑文添．无线传感器网络中的蠕虫检测与异常数据检测方案研究［D］．南京：南京邮电大学，2020.

[4] 施　珮．基于无线传感器网络的水质数据流异常检测与预测方法［D］．无锡：江南大学，2020.

[5] 付俊松，刘　云．基于信誉系统及数据噪声点检测技术的无线传感器网络节点安全模型［J］．清华大学学报（自然科学版），2017，57（1）：24－27.

[6] TRINH V V，TRAN K P，HUONG T T. Data driven hyper-parameter optimization of one-class support vector machines for anomaly detection in wireless sensor networks［C］//2017 International Conference on Advanced Technologies for Communications（ATC），Quy Nhon，Vietnam，2017：6－10.

[7] 孙　旭，刘明峰，程　辉，等．结合二次特征提取和 LSTM-Autoencoder 的网络流量异常检测方法［J］．北京交通大学学报，2020，44（2）：17－26.

[8] GROSKLOS G，THEILER J. Ellipsoids for anomaly detection in remote sensing imagery［C］//Algorithms and Technologies for Multispectral，Hyperspectral，and Ultraspectral Imagery XXI，Baltimore，Maryland，United States，2015：947－959.

[9] LYU L，JIN J，RAJASEGARAR S，et al. Fog-empowered anomaly detection in IOT using hyperellipsoidal clustering［J］. IEEE Internet of Things Journal，2017，4（5）：1174－1184.

[10] 华志颖，吴　蒙，杨立君．基于四分之一超球 SVM 的 WSN 异常检测［J］．南京邮电大学学报（自然科学版），2019，39（4）：47－54.

[11] 李　力．无线传感网中一种基于支持向量机的异常事件检测方案［J］．计算机应用与软件，2015（2）：272－277.

[12] KOWALóW，PATAN M. Distributed design of sensor network for abnormal state detection in distributed parameter systems［C］//Trends in Advanced Intelligent Control，Optimization and Automation，Krakow，Poland，2017：621－630.

[13] SHAHID N，NAQVI I H，QAISAR S B. Characteristics and classification of outlier detection techniques for wireless sensor networks in harsh environments：a survey［J］. Artificial Intelligence Review，2016，43（2）：193－228.

[14] 张静恬，伍　赛，陈　刚，等．基于多维数据集的异常子群发现技术［J］．计算机学报，2019，42（8）：1671－1685.

[15] SAFAEI M，ISMAIL A S，CHIZARI H，et al. Standalone noise and anomaly detection in wireless sensor networks：A novel time-series and adaptive Bayesian network based approach［J］. Software Practice and Experience，2020，50（4）：428－446.

[16] 刘洲洲，李士宁．采用压缩感知和 GM（1，1）的无线传感器网络异常检测方法［J］．西安交通大学学报，2017，51（2）：40－46.

[17] 彭能松，张维纬，张育钊，等．基于时间序列数据的无线传感器网络的异常检测方法［J］．传感技术学报，2018，31（4）：595－601.

[18] 郁　滨，熊　俊．基于平衡迭代规约层次聚类的无线传感器网络流量异常检测方案［J］．电子与信息学报，2022，44（1）：305－313.

[19] 陈志皑．无位置信息的无线传感器网络漏洞检测方法［J］．网络安全技术与应用，2018（5）：53－54.

[20] 卢光跃，周　亮，吕少卿，等．基于图信号处理的无线传感器网络异常节点检测算法［J］．计算机应用，2020，40（3）：783－787.

（上接第 232 页）

[8] 牛淑芬，陈俐霞，李文婷，等．基于区块链的电子病历数据共享方案［J］．自动化学报，2022，48（8）：2028－2038.

[9] 张　剑，夏　启，赵雅萍．基于区块链技术的电子病历数据存储系统研究［J］．中国医疗设备，2021，36（7）：106－109.

[10] 朱　海，金　瑜．DS-PBFT：一种基于距离的面向区块链的共识算法［J］．小型微型计算机系统，2022，43（3）：506－513.

[11] 李启南，薛志浩，张学军．改进 Fast-HotStuff 区块链共识算法［J］．计算机工程，2021，47（8）：14－21.

[12] 李静元，王　佳，张　珂．基于区块链和环签名的电子病历共享系统设计［J］．现代电子技术，2022，45（22）：116－120.

[13] 陈友荣，陈　浩，韩　蒙，等．基于信用等级划分的医疗数据安全共识算法［J］．电子与信息学报，2022，44（1）：279－287.

[14] 霍颖瑜，钟　勇．基于区块链的无线路由监测数据存储优化研究［J］．四川兵工学报，2021，42（10）：144－150.

[15] 胡荣磊，丁安邦，于秉琪．一种基于门限签名的区块链共识算法［J］．计算机应用研究，2022，39（12）：3555－3561.

[16] 许德俊，冯东雷，晏雪鸣，等．基于区块链的电子病历自主管理［J］．中国数字医学，2021，16（7）：18－23.

[17] 冯了了，丁　滟，刘坤林，等．区块链 BFT 共识算法研究进展［J］．计算机科学，2022，49（4）：329－339.

[18] 朱　海，金　瑜．DS-PBFT：一种基于距离的面向区块链的共识算法［J］．小型微型计算机系统，2022，43（3）：506－513.

[19] 王建国，李术君．基于区块链技术的电子处方共享流转模型［J］．中国数字医学，2021，16（10）：52－55.

[20] 沈　瑞，李玲娟．一种基于积分制的改进实用拜占庭容错算法［J］．计算机技术与发展，2021，31（6）：59－64.