基于多模态数据流的无线传感器网络异常检测方法

费 欢[1] 肖 甫[1),2)] 李光辉[3)] 孙力娟[1),2)] 王汝传[1),2)]

[1)](南京邮电大学计算机学院 南京 210003)

[2)](江苏省无线传感网高技术研究重点实验室 南京 210003)

[3)](江南大学物联网工程学院 江苏 无锡 214122)

摘 要 伴随着无线通信技术的不断发展和广泛应用,信息物理融合系统(Cyber-Physical System,CPS)作为物联网领域的最新研究方向成为近年来研究者广泛关注的热点.无线传感器网络(Wireless Sensor Networks,WSN)作为 CPS 系统物理空间的主要感知网络,若有效提高对应感知数据的准确性和可靠性,可及时准确地发现突发事件、监测网络工作状况,因此对传感器网络节点数据流进行异常检测,发现其中的异常数据并确认其来源具有重要意义.该文在无线传感器网络多模态数据流研究的基础之上,提出了一种对传感器异常数据进行检测以及监测节点自身工作状态的方法,该方法不仅应用了无线传感器网络中的时空相关性原理,还更进一步,研究了同一节点中多模态数据流之间的相干性,并以此作为理论基础,利用多维数据和滑动窗口模型对异常数据及其来源进行检测和验证.该文的方法可以分为 3 个步骤:首先,利用滑动窗口中的历史数据对传感器数据流进行异常数据的检测;其次,利用节点的空间相关性对异常的来源进行确认和识别;最后,对由于测量误差导致的异常值进行筛选,使输入 CPS 的数据进一步的精确化.通过实验对比,该文的方法对传感器异常数据的检测率保持在 95%;在不同数据维度的条件下,对四维数据集的检测率比单维数据集提高了 3%.

关键词 无线传感器网络;时空相关性;多模态数据流相干性;异常检测;物联网;信息物理融合系统

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2017.01829

An Anomaly Detection Method of Wireless Sensor Network Based on Multi-Modals Data Stream

FEI Huan[1] XIAO Fu[1),2)] LI Guang-Hui[3)] SUN Li-Juan[1),2)] WANG Ru-Chuan[1),2)]

[1)](*School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing* 210003)

[2)](*Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing* 210003)

[3)](*School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu* 214122)

Abstract With rapid development of the wireless communication technology, cyber physical system (CPS) that is a significant research orientation in internet of things has become a hot topic recently. As the main sensing network of physical space for CPS, improving the data accuracy and reliability of wireless sensor network (WSN) can effectively recognize emergent events and monitoring the situation of networks. Therefore, it is significant to detect abnormal data and identify their source. In order to accomplish this object, a novel anomaly detection and node-self monitoring method of wireless sensor networks based on multi-modals data stream that takes spatial-temporal correlation of different sensor nodes along with association of multi-modals data properties into account is proposed. Thus, abnormal data in streams can be detected while the source of anomaly

收稿日期:2016-05-30;在线出版日期:2016-10-30.本课题得到国家自然科学基金(61472368,61373137,61572260)、江苏省高校自然科学重大项目基金(14KJA520002)、江苏省六大人才高峰项目基金(2013-DZXX-014)资助.**费 欢**,男,1990 年生,博士研究生,主要研究方向为无线传感器网络.E-mail:feihuan1990@163.com.**肖 甫**(通信作者),男,1980 年生,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为物联网与传感网.E-mail:xiaof@njupt.edu.cn.**李光辉**,男,1970 年生,教授,博士生导师,主要研究领域为无线传感器网络、智能无损检测.**孙力娟**,女,1963 年生,教授,博士生导师,主要研究领域为无线传感网.**王汝传**,男,1943 年生,教授,博士生导师,主要研究领域为物联网、网络安全.

could be verified effectively based on multi-dimension data model and sliding window. The method proposed in this paper is divided into three steps: Firstly, abnormal data in sensor stream can be detected by historical relativity of data sets based on sliding window. Then, it is essential to identify the source of abnormal data and verify what makes the anomaly by spatial correlation of sensor nodes. Finally, the anomaly resulted from measurement error should be filtered and data set could be further sanitized into CPS. Experiment results demonstrated that the proposed method can obtain 95% accuracy of the anomaly detection while the accuracy of anomaly detection rate in four-dimensional data stream is 3% higher than that of the single dimension data stream for different dimension data.

Keywords wireless sensor networks; spatial-temporal correlation; association of multi-modals data; anomaly detection; Internet of Things; Cyber-Physical System

1 引 言

信息物理融合系统(Cyber-Physical System, CPS)综合了计算机、通信和控制等多个学科,作为一种具备局部操作、全局控制特点的新型网络系统引起学术界和工业界的广泛关注.作为连接虚拟网络世界与真实物理世界的桥梁,CPS 系统是一种面向物-物相连的新型网络应用系统,相较于以往的单一网络,CPS 系统对应了不同接入技术到不同用户终端的通信,集成了现有典型网络与应用,包括因特网、无线传感器网络、移动通信网等[1-3],可实现多模态数据的感知并对数据进行分析、决策以及执行相应的指令.

CPS 系统的基本架构可以分为物理层(Physical Layer)和虚拟层(Virtual Layer)两个部分.物理层主要由多种类型的传感器及控制器组成,负责物理环境中的数据采集和作业控制.然后,将传感器采集到的多种不同类型的数据经过处理和融合通过 CPS 系统的输入端进入虚拟层,由实时决策系统进行分析,将相应的指令返回给物理层的控制器执行相应的动作.如在常见的大棚监测系统中,多种类型的传感器可以实时采集多模态数据并将它们发送给服务器,服务器通过分析这些数据判断农田所处的状态,如果发现光照太强,则控制遮光板减少采光量;若发现湿度过低,则控制喷灌装置进行补水;如发现二氧化碳浓度过高,也可开启排风系统加强空气流通.与此同时,服务器还可以将这些数据及操作记录下来,及时通知管理人员,使其在第一时间内掌握相关的动态.除了环境监测之外,CPS 系统还被广泛地应用于医疗健康、安全监管、智慧交通等领域[4].

尽管应用的领域越来越广,但 CPS 系统仍然处于发展初期,其面临着诸多挑战.(1) CPS 系统的输入数据通常为多源异构数据流,在决策系统进行分析之前需要进行融合处理,避免发生冲突;(2) 由于 CPS 决策系统具有实时性,需要分析实时数据从而发出指令,这就要求数据处理的方法要具有较高的效率,传统离线、全局式的方法很难适用,需要更加灵活高效的在线分布式方法;(3) CPS 系统的数据趋于动态性和不稳定性,需要通过数据挖掘来筛选其有效信息,过滤掉无用的或错误的数据.除传统时空相关性,还需要进一步挖掘多模态数据之间的相干性并以此提高数据的可靠性,这对于 CPS 系统做出正确的决策至关重要[5].

作为一种信息感知和数据获取的重要手段,无线传感器网络(Wireless Sensor Network, WSN)在 CPS 系统中扮演着重要的角色,它可以不间断地感知海量数据,建立数据库,为数据分析提供基础.传统的 WSN 主要用于单一类型数据的采集、存储和传输,而 CPS 系统需要无线传感器网络同时提供多种来源不同的数据并对这些数据进行融合和分析,这就对输入数据的可靠性和准确性提出了更高要求.但由于分布区域的不确定性以及传感器节点资源有限,易受到外界因素的干扰和破坏或者外部环境突发事件的影响,无线传感器网络采集到的数据很有可能与实际特征产生明显偏差,此时采集到的数据就出现了异常,这类数据被称为异常数据[6].产生异常数据的原因主要包括:(1) 传感器节点分布的区域内发生了特定的事件(例如发生森林火灾时传感器的温度读数会明显升高);(2) 传感器由于自身的软硬件故障或者能量耗尽导致节点无法正常工作;

(3)由于外界因素的影响使节点采集到的数据较正常数据发生了偏差(例如处于阴影区域的传感器节点的光照强度数值会明显低于直接暴露在阳光下的节点读数),称这种偏差为测量误差[7-9].

无线传感器节点采集到的异常数据对于 CPS 系统的决策往往是有意义的.源于特定事件的异常数据往往反映外部环境确实发生了显著变化,需要及时预警并采取紧急处理措施;而源于传感器节点故障或能量耗尽而产生的异常数据反映出传感器网络健康状况存在问题,需要进行维护.存在测量误差的数据由于无法代表实际环境特征,往往会对系统准确掌握外部环境的变化规律产生影响.为了使 CPS 系统做出准确判断,需要对无线传感器网络采集的数据进行检测以便及时发现其中异常数据并对其来源进行分析和辨别.对无线传感器网络中的异常数据进行检测通常需要解决两个问题:(1)如何快速准确地检测出异常数据并对产生异常数据的原因进行分析和验证;(2)如何设计高效的分布式检测算法,尽可能减少节点间的消息通信,降低传感器节点的能量消耗.

本文第 2 节将详细介绍相关工作及研究成果;第 3 节将介绍无线传感器网络异常检测基本原理及相关定义;第 4 节将详细介绍以无线传感器网络多模态数据流为基础的异常检测和识别方法;第 5 节为对应实验结果及分析;最后是论文小结.

2　相关工作

为提高无线传感器网络数据的准确性,对采集到的数据进行异常检测显得尤为重要,其中传统事件检测方法主要用于区分无线传感器网络节点是否发生了事件或错误.根据对异常数据的不同定义,传统的无线传感器网络异常数据检测方法主要分为以聚类为基础的、以统计为基础的、以距离为基础的和以人工智能为基础的[10-14].

Krishnamachari 等人[15]提出了一种基于特殊事件在时间和空间上具有一定相关性但传感器故障则相对独立的假设,利用各传感器节点进行数据交换来统计传感器发生事件的概率,提出一种贝叶斯故障识别算法,该方法是一种基于聚类的方法,然而该方法只能区分事件和错误,对于错误的类型及原因很难进行有效地识别和区分.潘渊洋等人[16]提出了一种基于数据集密度的异常数据检测方法,该方法依据传感器节点的历史数据进行分析并对未来数据给出预测值,将实测值与预测值进行比较从而发现其中的异常数据,实现了传感器数据的实时检测,提高了算法的效率;但是该算法在处理高维度数据情况下对应的时间复杂度明显过高. Samparthi 等人[17]通过数学统计模型,利用核密度函数对传感器数据流进行异常检测,从而发现数据集中的异常数据,这是一种典型的基于统计的方法;但是该方法需要事先建立一个完整的数据集,缺乏实时性.张建平等人[18]提出了一种基于 Hadoop 的异常传感数据时间序列检测方法,针对传统 DTW 算法计算复杂度过高的问题,引入显著特征智能匹配的约束计算方法,通过对非直线路径进行局部限制,在保证较高检测准确率的同时有效降低了算法的时间复杂度和空间复杂度. Lee 等人[19]提出一种基于假设的数学统计方法,将对数据集的整体性和一致性具有正增益作用的数据视作正常数据,具有负增益作用的数据视为异常数据;但是该方法因为采用的是集中式处理,同样面临实时性问题. Lee 等人[20]提出一种通过比较传感器实测值与其数学期望之间的差异来实时计算传感器节点发生错误的概率,并依据时空相关性区分异常的类型;但是该方法并没有针对高维度的数据集给出对应的解决方案.毕冉等人[21]提出了一种无线传感器网络中能量高效的 Top-k 查询算法,该算法根据数据的异常程度对数据集进行筛选和排序,反馈给用户 k 个最大或最小的数据值并对异常进行定位.胡石等人[22]对 Top-k 算法进行了改进,根据无线传感器网络节点的数据分布规律,构造与之相适应的数据网格,将多维数据归一化处理后置入相应的网络单元,接着将重构后的列表与距离阈值 σ 进行比较发现其中的异常数据. Subramaniam 等人[23]解决了核密度估计方法中单一阈值无法满足多维数据对象异常检测的问题,并利用核密度估计对数据分布模型进行更新维护. Palpanas 等人[24]提出了一种利用核密度估计对传感器数据流中的异常数据进行在线检测的方法.该方法无需了解数据分布的先验知识,而是通过比较滑动窗口中的当前数据与先验模型之间的差异是否超出阈值来判断其是否为异常数据.任倩倩等人[25]提出了一种面向容错的事件区域检测估计算法,该算法可有效估计异常事件的发生范围和检测边界,同时具备一定的容错能力.张书奎等人[26]提出了一种容错检测算法,通过分布式地构建融合树,各节点向对应的最近树根节点发送所采集的数据,从而实现单个/多个事件的鲁棒容错检测.曹冬磊等人[27]提出了一种

具有容错功能的算法来检测 WSN 中事件发生的区域和边界,其将事件视为随机过程,通过比较传感器采样值与其统计特征间的差异程度进行事件区域的估计,同时具有容错能力,可以适应网络拓扑结构的变化.

然而上述方法很少专门针对传感器节点的多模态数据给出特定的解决方案,都忽略了多模态数据间的相干性.事实上,为了全面详细地对复杂的环境特征进行准确描述和再现,综合多种不同类型的数据进行评估是十分必要的.在这种情况下,如果采用现有的单一模态方法处理多模态数据集则需要进行重复计算,则该方法将成为一个串行算法,其时间复杂度会随着数据集规模的增大成倍增加,这将耗费传感器节点的有限资源,降低节点的使用寿命;其次,此类算法也忽视了不同模态的数据集会影响算法的性能和检测结果.以部署在野外环境的传感器节点为例,采集到不同环境数据的波动幅度、波动频率、采样中值、均值、方差等统计学特征均存在明显的差异,不考虑这种客观存在的差异性而采用同样的方法进行无差别对待无疑会降低其可靠性和鲁棒性.此外,当节点的部署区域内发生事件时的观察特征显然不是孤立的,比如发生火灾时不仅温度会大幅度上升,同时湿度也会下降,短时间内二氧化碳的浓度也会上升,氧气浓度会明显下降;降雨时湿度会迅速上升,温度则同步下降.因此,为了对环境中的不同事件进行准确的识别,需要分析无线传感器网络在环境监测过程中采集到的多模态数据流之间的相干性.

本文在现有工作的基础上,针对以往研究成果存在的不足,进一步研究了节点多模态数据之间的相干性,从数据集的特征入手,提出了一种以多模态数据流的统计特征分析为基础的无线传感器网络异常检测方法来实现数据集中异常值的检测和识别,并以此为基础监测分布区域内的特殊事件、节点故障和测量误差.

3 WSN 异常数据检测基本原理

一般而言,当传感器节点感知的数据和实际特征相比出现连续若干次偏差时,可以认为该节点出现了异常.然而,导致无线传感器网络节点产生异常的原因有很多,由区域内特殊事件引起的称为事件节点;由自身故障或外部攻击导致无法正常采集数据的称为故障节点;由采样过程中的外界因素影响而使数据与实际值产生误差的称为误差节点.

3.1 WSN 数据流的时间及空间相关性

部署在监测区域内的节点收集到的外界数据是随采样时刻 t 变化的一簇变量.通常情况下,采样时刻 t_i 处的数据是否异常只与 t_{i+1} 时刻的数据值有关,而与之后时刻的值无关.这一现象说明异常数据具有历史相关性,并且采样周期的长短对这种相关性有着较大影响.例如,对同一传感器节点而言,每一天的温度数据均遵循相同的规律缓慢波动,其存在着时间上的相关性,如果相邻两个采样时刻的间隔足够小,那么就有可能获得在最小精度内的相同数据.此外,传感节点采集的数据通常还具有空间相关性.通常情况下无线传感器节点是以密集分布的形式覆盖监测区域,在密集分布的条件下,任意传感器节点与其距离较近的邻居节点采集的数据也具有一定的相关性.如图 1 所示,如果在 o 处出现了火源,那么在其附近的节点 a,b 和 c 应该均能监测到该事件(比如温度值明显上升).因此,传感器数据流的时间及空间的相关性为实现异常检测提供了理论依据.

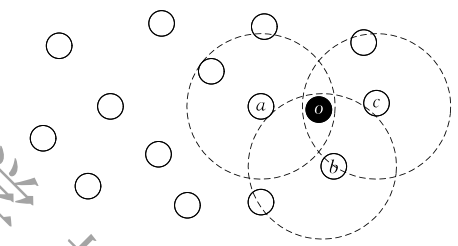

图 1 不同位置节点对同一事件源的监测

假设二维空间内分布有 N 个无线传感器网络节点,则该区域内的事件可以用一个以 $s(t,x,y)$ 为函数的时空模型进行描述,其中 t 表示物理事件 s 发生的时刻,(x,y) 表示事件源的位置,事件 s 的实际值以集合 S 表示,节点对事件 s 的观测值以集合 $\hat{S}$ 表示.那么某一节点 n_i 在采样时刻 n 对于事件 s 的观测值可以表示为

$$X_i[n] = S_i[n] + N_i[n] \tag{1}$$

$$Y_i[n] = f_i(X_i[n]) \tag{2}$$

$$\hat{S} = g(Y_1[n_1],\cdots,Y_1[n_j],\cdots,Y_N[n_1],\cdots, Y_N[n_j]) \tag{3}$$

其中:$X_i[n]$表示节点的输出值集合;$S_i[n]$表示事件 s 的实际值集合;$N_i[n]$表示数据采集过程中的外界干扰.节点将 $X_i[n]$通过无线信道经过编码,转换为 $Y_i[n]$传输给汇聚节点(sink)进行处理,得到观测值集合 $\hat{S}$,函数 g 为 sink 对于数据采用的处理函数.对于传感器节点观测值的失真程度可以用均方差 D 来表示[28]

$$D = E[(\hat{S} - S)^2] \tag{4}$$

3.2　多模态数据流模型及其相干性

如上文所述，CPS 系统在通常情况下是一个异源多模态数据输入的系统，因此除了单一数据的时间及空间相关性还需要考虑多种不同类型数据间的相干性来提高 CPS 系统的可靠程度. 在讨论多模态数据流的相干性之前，首先需建立一个多维数据流模型.

假设在某一目标区域内均匀部署有 N 个同时配置了可采集 M 种不同模态数据的传感器（如温度、湿度、光照强度、CO_2 浓度等）节点进行数据采集，无线传感器网络中部署的各节点可以通过时间同步机制等来保证数据采集和信息传输的同步性.

在某一采样时刻 t，任一传感器节点采集到的 M 种不同模态的数据可被视作一个 M 维空间的数据点集合 $X=(r_1, r_2, \cdots, r_M)$，而某一采样周期内采集到的数据则构成一个矩阵

$$\{r_j(t_i)\} = \begin{bmatrix} r_1(t_1) & r_1(t_2) & \cdots & r_1(t_N) \\ r_2(t_1) & r_2(t_2) & \cdots & r_2(t_N) \\ \cdots & \cdots & \cdots & \cdots \\ r_M(t_1) & r_M(t_2) & \cdots & r_M(t_N) \end{bmatrix}$$

其中 $t_1, t_2, \cdots, t_N$ 为对应采样时刻[29].

考虑到传感器的数据流在时间序列上具有无限延展性，而传感器节点自身的存储空间有限，本文采用滑动窗口模型对数据流进行处理，其具体定义见定义 1.

定义 1.　滑动窗口模型（Sliding Window Model）是将传感器数据流截取一段长度为 $|W|$ 的窗口并将此窗口等分为 m 个小块，分别为 $Block_1, Block_2, \cdots, Block_m$，每个小块的长度为 n. 当下一采样时刻 t_{next} 的数据进入滑动窗口时，上一采样时刻 t_{for} 的数据则将被替换

$$\text{mod}(t_{\text{next}}, |W|) = \text{mod}(t_{\text{for}}, |W|) \tag{5}$$

其中，mod(a, b)表示取余函数. 同一传感器节点上的若干个传感器内部均使用滑动窗口处理数据流，如图 2 所示.

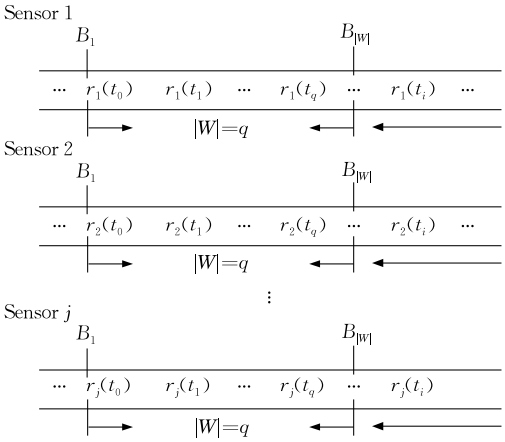

图 2　多维数据图流的滑动窗口模型

如果假设节点 S_n 将前 p 个采样时刻的数据 $\{r_j(t_1), r_j(t_2), \cdots, r_j(t_p)\}$ 载入滑动窗口中，$\delta^2 = \frac{1}{p}\sum_{i=1}^{p}\prod_{j=1}^{m}(r_j(t_i)-\bar{r}_j)^2$ 为该组数据对应的计算方差，其中 $\bar{r}_j$ 表示滑动窗口中所对应传感器采集到的第 j 维数据的平均值. 当新的数据 $r_j(t_{p+1})$ 进入滑动窗口时，窗口向后滑动，同时将窗口中的数据更新为 $\{r_j(t_2), r_j(t_3), \cdots, r_j(t_{p+1})\}$，对应的采样数据方差的计算公式可表示为 $\delta^2 = \frac{1}{p}\sum_{i=2}^{p+1}\prod_{j=1}^{m}(r_j(t_i)-\bar{r}_j)^2$，后续采样时刻的数据依次类推.

建立多维数据流的模型之后，下面对多模态数据流之间的相干性进行分析，首先介绍相干性系数.

定义 2.　相干性系数[30]：

$$\rho_{ik} = \frac{C(X_i, Y_k)}{\sqrt{C(X_i, Y_i)}\sqrt{C(X_k, Y_k)}} = \frac{\sum_{j=1}^{n}(x_{ji}-\overline{X}_i)(y_{jk}-\overline{Y}_i)}{\sqrt{\sum_{j=1}^{n}(x_{ji}-\overline{X}_i)^2}\sqrt{\sum_{j=1}^{n}(y_{jk}-\overline{Y}_i)^2}} \tag{6}$$

其中：x_{ji} 与 y_{jk} 分别表示任意的两条数据流 X_i 和 Y_k 时间序列中的第 j 个值；相干性系数 ρ 是评价多维数据流相干性的一个重要指标，如果 $\rho<0$，数据流间呈负相干关系；若 $\rho>0$，则数据流为正相干；若 $\rho=0$，数据流无相干性关系.

在实际的操作过程中，如果采用宽度为 n 的滑动窗口模型对多条数据流进行相干性分析时，需要分别计算数据流 $\boldsymbol{X}_i=(X_{1i}, X_{2i}, \cdots, X_{pi})$ 的协方差矩阵 $\boldsymbol{S}_{11}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})^{\mathrm{T}}$，$\boldsymbol{Y}_i=(Y_{1i}, Y_{2i}, \cdots, Y_{qi})$ 的协方差矩阵 $\boldsymbol{S}_{22}=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})(Y_i-\overline{Y})^{\mathrm{T}}$ 及 X_i, Y_i 的协方差矩阵 $\boldsymbol{S}_{12}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})^{\mathrm{T}}=\boldsymbol{S}_{21}^{\mathrm{T}}$.

在对数据进行标准化处理后，样本的相干性系数对应为样本的协方差，最后利用卡方检验的方法可得到对应的典型相干性系数和典型相干性变量.

图 3 所示为火灾发生过程中传感器节点采集到

的多模态数据的同步变化情况，从图中可以较为直观地观察到当发生火灾时温度值与 CO_2 浓度值呈明显的正相干性，与湿度值具有负相干性. 当无线传感器网络用于野外环境监测时，如果仅凭单一模态的数据判断是否有事件发生存在较大的误判可能，但是如果同时综合多模态的数据进行判别，则可以极大地提高准确性，避免误判.

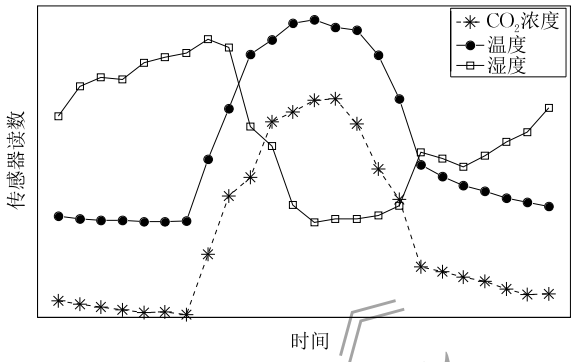

图 3 发生火灾时多模态数据的同步变化情况

综上，无线传感器网络多模态数据之间的相干性及时空相关性为准确高效地进行异常检测提供了理论基础，在此基础之上，本文提出了一种对 WSN 多模态数据流进行异常检测的方法.

4 WSN 多模态数据流异常检测方法

和一般的数据集相比，无线传感器网络的多模态数据集在分布上有其与众不同的特征. 图 4 为传感器三维模态的原始数据分布示意，从图中可以概括出传感器数据集的分布特点：大量的数据点紧密地分布在一起，密度大，重叠度高；少量数据点则较为分散和孤立，数量明显少于密集分布的数据并且差异明显.

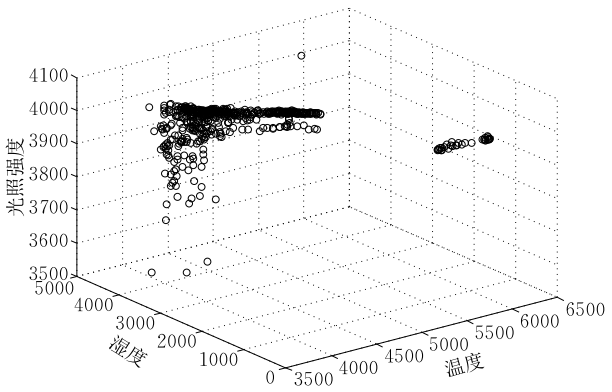

图 4 传感器三维模态原始数据分布

Knorr 等人[31]提出了一种经典的基于距离的异常检测方法，其中异常定义如定义 3.

定义 3. $DB(p,D)$异常是指数据集 S 中存在一个对象 O 且在 O 的距离为 D 的邻域内的其他对象数量与数据集对象总数量之比小于 $1-p$.

但是由于 p 和 D 的取值不易确定，取值的不同会对结果产生明显的影响，需要反复测试选择实验最优解，从而增加了算法的复杂度. 为有效解决该问题，Ramaswamy 等人[32]提出了一种新的基于距离的异常定义如定义 4.

定义 4. $D(n,k)$异常是指在包含有 N 个数据点的 d 维空间中，存在有不超过 $n-1$ 个点 p' 满足条件 $Dk(p')>Dk(p)$. 其中 $Dk(p)$表示点 p 和它的第 k 个最近邻 p' 的距离，n 和 k 均为参数. 如果对数据点按照 $Dk(p)$距离进行排序，则前 n 个点将被视为异常数据. 一般情况下，异常数据的检测是计算其与最近邻居点的距离，故取 $k=1,n=2$.

$D(n,k)$异常解决了 $DB(p,D)$异常需要针对不同的特定数据集进行人工确定参数的问题，极大地提高了可操作性.

本文对 WSN 多模态数据流进行异常检测的方法以基于距离的异常定义为基础，分为异常数据识别、异常来源验证、测量误差检测这 3 个步骤，利用滑动窗口模型和数据流的时间相关性原理对传感器多维数据集合 X 中可能存在的异常数据点进行识别，考虑到数据异常情况通常在时间上都具有一定的延续性，本文引入单模异常概率 $P_j(t_i)$ 和多模异常概率 $P_f(t_i)$ 来评估多模态数据流出现异常的可能性. 这里的单模异常概率与多模异常概率均为变量，以数值的大小表示可能性的高低，而非 0 到 1 之间的百分数，具体的定义和计算方法下文将详细说明. 在检测到异常数据点之后，需要利用空间相关性原理对异常数据点进行进一步验证从而确认并区分其异常来源，之后将数据输入 CPS 决策系统中，对传感器节点的工作状态进行监测，下文将分别进行详细说明.

4.1 异常数据识别

如前文所述，传感器的测量值应当准确再现实际的环境特征，因此测量值 $r_j(t_i)$ 在稳定的环境中表现为一定幅度内的缓慢波动，但出现异常时则会在短时间内出现明显偏差. 如果 $r_j(t_i)$ 满足式(7)，那么该测量值可能为异常数据.

$$\left| r_j(t_i) - \frac{E_{ej}(t)+E_{nj}(t)}{2} \right| > \delta^2 \qquad (7)$$

其中：t 是传感器节点的对应采样时间；$E_{ej}(t)$ 表示正常工作的传感器在事件区域内测量值的数学期望，

$E_{nj}(t)$表示正常区域中测量值的数学期望；一般认为$E_{nj}(t)$在稳定的环境下为常数. 不同环境下的 $E_{ej}(t)$和 $E_{nj}(t)$的值存在差异，需视数据集的情况而确定.

此外，当传感器自身出现故障时(能量耗尽或遭到破坏无法正常工作)，可能在不同的采样时刻连续产生相同的读数[20]，即

$$r_j(t_i) = r_j(t_{i-1}) \tag{8}$$

上述两种情况被称为判断传感器读数是否为异常的判断条件，并以此为基础来计算单模数据流的异常概率 $P_j(t_i)$

$$P_j(t_i) = P_j(t_{i-1}) + c \cdot k^2 \tag{9}$$

单模异常概率 $P_j(t_i)$是一个累加值，采样时刻 t_i出现异常的概率用 $P_j(t_i)$表示，前一采样时刻 t_{i-1}出现异常的概率用 $P_j(t_{i-1})$表示，$\{r_j(t_i)\}$满足判断条件的次数用常数 k 表示，c 为参数. 如果在若干个采样时刻读数 $r_j(t_i)$连续满足判断条件，k 从 0 开始逐次增加，此时 $P_j(t_i)$与 k 呈指数关系；若 $r_j(t_i)$不满足判断条件，则 k，$P_j(t_{i-1})$，$P_j(t_i)$同时清零，当 $r_j(t_i)$满足判断条件时重新开始累加.

传感器节点可以集成多种传感器，在某一采样时刻采集多模态的数据流，产生多组 $P_j(t_i)$值. 然而仅仅通过单一模态的数据流就去判断数据异常的产生原因无疑是不够精确的，需要融合多模态的数据流进行分析和判断. 由多组的单模异常概率 $P_j(t_i)$值可以计算多模异常概率 $P_T(t_i)$

$$P_T(t_i) = \sum_{j=1}^{m} \lambda_j \cdot P_j(t_i),\ \left(\sum_{j=1}^{m} \lambda_j = 1\right) \tag{10}$$

其中，权重系数用 λ_j表示，考虑到不同数据的 $P_j(t_i)$数值可能存在差异，有些偏大而有些相对较小，为了平衡不同的 $P_j(t_i)$对 $P_T(t_i)$的影响，将波动频率快、幅度大的数据的 λ_j设置为较大数值，提高算法的灵敏度；反之则将 λ_j 设为一个较小的值，有效避免因个别数据发生误判. 考虑到 λ_j与数据的波动幅度有关，在取值上可以使其在比例上与数据的标准差比保持一致，即

$$\lambda_1 : \lambda_2 : \cdots : \lambda_j = \sigma_1 : \sigma_2 : \cdots : \sigma_j \tag{11}$$

4.2　异常来源验证

当某个传感器节点的 $P_T(t_i)$ 值达到阈值 $R_{th} = \sum_{j=1}^{m} \lambda_j \cdot \frac{E_{ej}(t)+E_{nj}(t)}{2}$时就认为该节点可能发生了异常，考虑到多模态数据集的差异性和方法的鲁棒性，将 R_{th}设置为一个固定值是不合适的，R_{th}的值应当与数据集的统计特征相关并且突显不同的影响因子. 因此我们将 R_{th}设置为多维数据集均值的加权平均，既体现其与全部数据集的相关性，又确保不同数据集之间存在差异性；为确认异常的来源，需要利用空间相关性进行验证. 当某一节点监测到自身发生疑似异常时，通过无线信道向其邻节点发送请求消息，接收邻节点的 $P_T(t_i)$值. 根据拉依达准则，若本节点的 $P_T(t_i)$值满足 $|P_T(t_i)-\mu| < \delta\sigma$($\mu$ 和 σ 分别为邻居节点 $P_T(t_i)$值的均值和标准差)，则认为误差来源于事件过程中的随机误差，而本节点的状态与邻节点相一致；若不满足则认为本节点与邻节点的状态不一致，出现了故障或测量误差. δ 需要依据具体情况进行取值，但通常可以将事件过程视为一个随机变量符合正态分布的贝努利过程，故可以将该随机变量简化为标准正态分布的随机变量：

$$\begin{aligned} p &= P\left(\frac{|P_T(t_i)-\mu|}{\sigma} \geq \delta\right) \\ &= 1 - P\left(-\delta < \frac{|P_T(t_i)|}{\sigma} < \delta\right) \\ &= 1 - (\Phi(\delta) - \Phi(-\delta)) \\ &= 2 - 2\Phi(\delta) \end{aligned} \tag{12}$$

其中：$\Phi(\delta)$表示标准正态分布，通过查表可求得当 $\Phi(\delta) > 0.975$ 时，$p < 0.05$，当 δ 约大于 1.96 时，$\Phi(\delta) > 0.975$，可以取 $\delta = 2$. 异常来源的具体判断条件如下.

若 $r_j(t_i) \neq r_j(t_{i-1})$并且 $|P_T(t_i)-\mu| < 2\sigma$，则认为所处区域发生了特定的事件；

若 $r_j(t_i) = r_j(t_{i-1})$并且 $|P_T(t_i)-\mu| \geq 2\sigma$，则认为传感器发生了故障；

若上述条件均不满足，则认为可能出现了测量误差，需要进一步对可能存在的测量误差进行检测并筛选出存在测量误差的数据.

4.3　测量误差检测

出现测量误差的节点首先已经排除了发生故障的可能，应该是可以正常工作的节点. 但其采集到的数据流中又存在与实际的环境特征存在明显偏差的数据，因此对出现测量误差的节点进行检测，剔除其测量误差读数，筛选出正常数据，可有效提高传感器数据流的可靠性.

考虑到 WSN 数据流中的正常数据数量多，分布密集；而异常数据相比而言则数量偏少且差异明显，因此选择数据挖掘中的聚类算法对测量误差读数进行检测，可以快速地筛选出其中的异常数据，在数据集规模较大的情况下同样具有良好的性能.

定义 5.　若 x_i，x_j分别为 d 维空间内的任意两点，则其欧几里得距离(Euclidean distance)可表示为

$$d(x_i,x_j)=\sqrt{\sum_{k=1}^{d}(x_{ik}-x_{jk})^2} \tag{13}$$

如果两个数据点 x_i 和 x_j 之间的欧几里得距离 $d(x_i,x_j)$ 越小，则两者的差异度越小；反之则越大.

由于每个坐标对欧氏距离的贡献是同等的，欧氏距离没有考虑当坐标表示测量值时，它们往往带有大小不等的随机波动，某些维度的数值变化无穷大而有些无穷小. 如果使用不同的计数方法在数值上也会存在较大差异. 为解决这一问题，首先需要对数据进行标准化处理，使变化较大的坐标比变化较小的坐标有较小的权系数. 本文使用的数据集最大维度为四维，考虑到传感器数据流的波动性，采用 Z-score 标准化方法对数据集进行预处理.

定义 6. Z-score 标准化是将 A 的原始值 x 基于原始数据的均值和标准差对应到 x' 的方法.

对于 A 的最大值、最小值未知或存在超出取值范围的情况，Z-score 标准化较为适用，公式如下

$$x'=(x-\mu)/\sigma \tag{14}$$

其中，μ 和 σ 分别为原始数据的均值和标准差.

定义 7. 评价函数 E 可以表示为数据集 S 中各对象 m_i 与其所属聚类中心 m_c 距离的平方和：

$$E=\sum_{i=1}^{k}\sum_{m_i\in S}\sqrt{(m_i-m_c)^2} \tag{15}$$

评价函数 E 在一定程度上体现了聚类结果的准确性，E 的值越小则聚类的效果相对较佳. 当 E 不再随迭代过程变化时则认为此时聚类达到最优解. 而对于算法中 K 值难以确定的问题，可采用遗传算法进行优化[33]. 该方法通过模拟自然界的遗传和变异过程，寻找 K 值的全局最优解，一定程度上降低了不同取值对算法结果的影响，较好地解决了 K-means 算法多次迭代收敛慢且容易陷入局部最优的问题.

K-means 算法中初始的 K 值的选择对性能和效率有直接的影响，可以使用经典的遗传算法(GA)对 K 值进行编码. 考虑到一般情况下 K 是一个大于等于 1 的整数，可以用一个字节的二进制串来表示，即 255 类，该二进制串对应为遗传算法中的染色体.

在遗传算法中，每一个 K 值均对应一条染色体，为了寻找全局最优的 K 值，需要以 K-means 算法最终的聚类结果来评价染色体的适应度，若类内距离越小的同时类间距离越大，则聚类结果越理想.

定义 8. 类内距离 $\mathrm{Dis}(p_{i,j},m_i)$：

$$\mathrm{Dis}(p_{i,j},m_i)=\frac{1}{k}\sum_{i=0}^{k}\frac{\sum_{j=0}^{num_i}d(p_{i,j},m_i)}{num_i} \tag{16}$$

定义 9. 类间距离 $\mathrm{Dis}(m_i,m_j)$：

$$\mathrm{Dis}(m_i,m_j)=\frac{2\sum_{i=0}^{k}\sum_{j=k+1}^{k}d(m_i,m_j)}{k(k-1)} \tag{17}$$

定义 10. 适应度函数 Fit：

$$\mathrm{Fit}=\alpha\frac{\mathrm{Dis}(m_i,m_j)}{1+\mathrm{Dis}(p_{i,j},m_i)}+\beta\frac{1}{NumD} \tag{18}$$

在上述公式中，第 i 类的聚类中心用 m_i 表示；m_i 与 m_j 之间的距离用 $d(m_i,m_j)$ 表示；全部数据集中隶属于 i 类的样本数用 num_i 表示；$p_{i,j}$ 表示隶属于第 i 类中的第 j 个样本；用 $NumD$ 统计不同类之间样本个数的差别，调节类间距离与类内距离的权重系数分别用 α 和 β 表示.

首先，以伪随机数的方式产生聚类中心的初始位置，接着形成初始染色体组，计算染色体的适应函数 Fit，再进化部分适应性好的染色体，通过交叉、变异，最后选择适应性强的染色体形成新一代染色体组. 经过一定次数的迭代，找到最优的 K 值.

综上，本文方法对应的伪代码描述如下.

```
//步骤 1：异常数据识别
Input：{r_j(t_i)},v_n(t);E_e(t),E_n(t);P_j(t_i),P_T(t_i),c;
if ( |r_j(t_i) − (E_ej(t)+E_nj(t))/2| > δ^2 or r_j(t_i)=r_j(t_{i−1})) {
P_j(t_i)=P_j(t_{i−1})+c·k^2;
P_T(t_i)=Σ_{j=1}^{n} λ_j · P_j(t_i), (Σ_{j=1}^{n} λ_j = 1);
}
//步骤 2：异常来源验证
if (P_T(t_i)>R_th) {
  status=anomaly;
  broadcasting P_T(t_i) to neighbors;
}
Receiving P_T(t_i) from neighbors, compute μ and σ;
if (r_j(t_i)≠r_j(t_{i−1}) and |P_T(t_i)−μ|<2σ )
  status=Event;
if (r_j(t_i)=r_j(t_{i−1}) and P_T(t_i)−μ|≥2σ)
  status=Fault;
else go to Step 3
//步骤 3：测量误差检测
do {
  for (j=1 to n){
    Do assign each x_j to the closest cluster;
    Replace the average distance of cluster;
  }
Compute Fit and E;
k=k+1;
}
while(E remains changing)
```

```
for (i=1 to n)
{
  Compute average distance of cluster: d_avg;
  if (d(x_i, m_i) ≥ d_avg)
    Put x_i into Suspect;
  Compute variance of distance between x_i and m_i as v_d;
    if (|d(x_i, m_i) − d_avg| > 1.67 × sqrt(v_d))
  Consider x_i as measurement error data;
}
```

4.4　性能分析

为了对本文方法的可靠性进行准确的定量评估，此处引入节点的检测率(True Positive Rate)及误报率(False Positive Rate)这两项指标，下面以事件节点为例，分别给出定义[34].

定义 11.　事件节点的检测率 $a(C)$：

$$a(C)=\frac{|C\cap E|}{|E|} \tag{19}$$

定义 12.　事件节点的误报率 $e(C)$：

$$e(C)=\frac{|C|-|C\cap A|}{|S|-|A|} \tag{20}$$

其中：S 为全部节点的集合；E 为实际事件节点的集合；A 为异常节点的集合.

接下来进行时间复杂度的分析. 在滑动窗口中，如果输入的数据流维度为 m，各维度的数据流规模为 n，滑动窗口的长度为 w，$n>w$. 滑动窗口每次滑动一格，则完成全部数据流的输入需要 $n-w$ 次，由于 m 维数据流在 m 个滑动窗口中同时输入，每次输入之后需要完成均值及方差的更新，故滑动窗口的时间复杂度为 $O(n-w)$，空间复杂度为 $O(mw)$. 由此可见，在时间复杂度上由于采用了并行的方法，规避了时间因为数据流维度成倍增长的问题；在空间复杂度上，因为滑动窗口先进先出的特性解决了以往集中式方法数据集规模过大的问题. 对 K-means 而言，由于算法迭代的特性，时间复杂度为 $O(nkt)$，其中 n 表示数据集样本数目，k 表示初始聚类数，t 表示迭代次数，即时间复杂度与数据集规模为线性关系；与此类似的文献[10]采用了基于密度的DBSCAN算法，需人工确定点 P 的邻域半径 Eps 和邻域中包含对象的最小个数 $MinPts$ 两个参数，采用索引结构依据邻域内包含对象的数目进行聚类划分，算法的时间复杂度为 $O(n\log n)$；但是此方法在高维度的数据集下，因数据的索引结构失效，时间复杂度退化为 $O(n^2)$. 由此可知，在大规模、高维度的数据集下，本文算法与基于密度的方法相比在时间复杂度上具有优势.

5　仿真实验及结果分析

如上文所述，滑动窗口长度的选择会对算法的效率和检测结果产生显著影响. 如果滑动窗口过小，会造成较大的偏差，降低算法的可靠程度；过大的窗口长度不仅不能提高算法性能，还会导致有限资源的浪费，影响算法效率. 为了确定滑动窗口的长度，本文通过实验进行选择.

实验 1.　为考察不同长度的滑动窗口对数据流统计特征的影响，选取无线传感器网络的温度、湿度、光照强度、CO_2 浓度四维数据各 5000 组，分别用不同长度的滑动窗口对方差进行计算，结果如表 1. 数据流的方差随着窗口长度的增加趋于稳定，综合 4 种数据的结果分析，滑动窗口的长度在 240～260 较为合适，这里选择 240.

表 1　不同滑动窗口长度下数据流的方差

滑动窗口长度	方差			
	温度	湿度	光照强度	CO_2 浓度
10	5.68	4.42	8.51	2.33
100	6.74	3.92	8.84	2.27
150	5.98	4.38	8.93	2.24
180	6.34	4.27	8.69	2.26
200	6.09	4.34	8.81	2.15
220	6.15	4.31	8.74	2.21
240	6.17	4.29	8.76	2.18
260	6.19	4.30	8.76	2.16
270	6.18	4.29	8.78	2.17

为验证本文方法的有效性并评估其性能，利用无线传感器网络野外监测系统获取的温度、湿度、光照强度、CO_2浓度等四维数据共 15000 组，对本文方法和现有其它典型方法进行对比实验.

实验 2.　比较本文方法在不同数据维度下的异常数据检测率.

为验证上文所述的传感器多维数据之间的关联性，对同一传感器节点采集到的多维数据进行异常

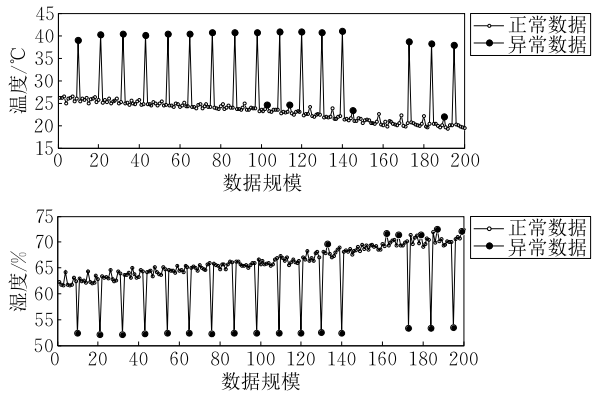

图 5　温度和湿度异常数据的检测结果

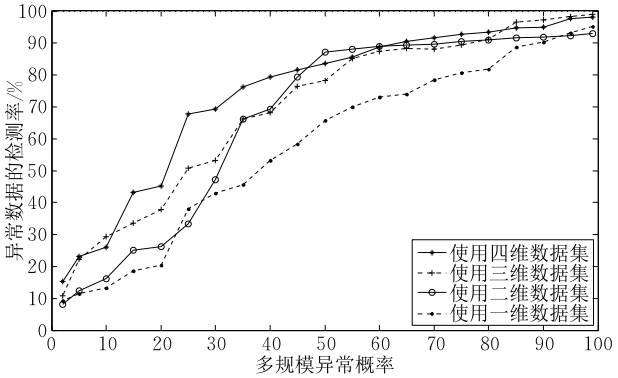

图 6 异常数据的检测率随 $P_T(t_i)$的变化规律

检测.以温度和湿度为例,图 5 给出了算法检测的结果.可以看出,当温度处于正常区间时对应的湿度数据基本处于正常范围,而当温度发生异常时湿度数据也相应出现了明显的偏差.这一实验结果说明了传感器不同模态数据之间并非孤立的,它们具有明显的相干性.

图 6 对应描述了传感器异常数据的检测率随多模异常概率 $P_T(t_i)$变化的规律.当二维数据集的 λ_j 分别为温度 0.6,湿度 0.4;三维数据集的 λ_j 分别为温度 0.3,湿度 0.3,光照强度 0.4;四维数据集的 λ_j 分别为温度 0.3,湿度 0.2,光照强度 0.4,CO_2 浓度 0.1;$c=0.002$ 时,异常读数的检测率均随着 $P_T(t_i)$ 的增加而增加:当 $P_T(t_i)$为 5%时,检测率为 10%;当 $P_T(t_i)$为 95%时,检测率为 94%.此外,通过对不同维度数据集的对比可知,当 $P_T(t_i)$相同时,四维数据集的检测率高于单一维度数据集的检测率.

异常数据的检测率随数据规模变化的规律如图 7 所示.当数据规模较小(小于 3500)时,数据集的密度低,分布较为分散,单一数据点对计算距离均值有明显的影响,此时,将平均距离作为测度的聚类结果不理想,平均检测率只有 85%左右;但随着数据样本规模的增大,单一离群点对结果的影响逐渐减弱,距离均值对于数据样本特征的代表性逐渐增强,检测率稳定在 97%.

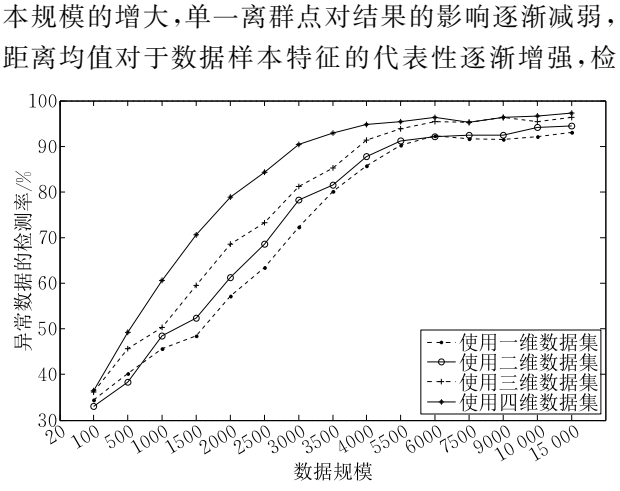

图 7 异常数据的检测率随数据规模的变化规律

实验 3. 对比经 GA 改进后的 *K*-means 算法与 DBSCAN 算法对测量误差读数的检测效果.

与传统 *K*-means 算法相比,本文对 *K* 值的选择用遗传算法进行改进,使其可以根据不同的数据集进行自动优化.实验中用到的遗传算法参数为:染色体交叉率为0.83;突变率为0.02;子代染色体中新染色体所占比例为0.75;初代染色体的数目 50;子染色体的代数为 100.表 2 的结果表明经过改进后的 *K*-means 算法在相同的数据规模下对误差读数具有更高的检测率和更低的误报率.

表 2 改进前后 *K*-means 算法的检测结果对比

数据规模	*K*-means 算法		改进后的 *K*-means 算法	
	检测率/%	误报率/%	检测率/%	误报率/%
1300	92.27	4.68	94.16	4.15
1800	92.34	4.69	94.32	4.07
2200	92.74	4.55	94.41	4.02
2600	92.87	4.56	94.55	3.87
3000	92.93	4.51	94.61	3.51
5000	93.29	4.48	95.73	3.26
10000	93.32	4.36	96.19	2.83
15000	93.46	4.33	96.33	2.77

为比较两种算法的检测效果,引入 ROC 曲线(Receiver Operating Characteristic Curve)[35] 作为标准.一般而言,好的算法应同时具备较高的检测率和较低的误报率,那么它的 ROC 曲线应尽可能趋近左上角.即算法 ROC 曲线下方的面积 AUC (Area Under the ROC Curve)越大,那么可以认为该算法的性能越佳.如图 8 所示,在不同的数据规模下,改进后的 *K*-means 算法相比于 DBSCAN 算法具有更高的检测率及更低的误报率,其 ROC 曲线下方面积更大,因此对传感器数据流中的测量误差读数具有更好的检测效果.

实验 4. 为了验证本文方法对节点异常来源验证的准确度,仿真实验中,在 200 m×200 m 的区域内设置了无线通信半径为 20 m 的节点 200 个,

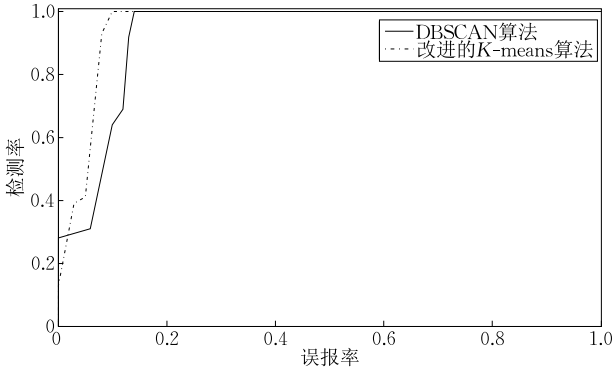

图 8 两种算法的 ROC 曲线

$c=0.002$，将四维模拟数据集注入 200 个节点中，人工设置故障节点、事件节点、正常节点和误差节点各 12 个、56 个、112 个及 20 个. 图 9 是对应的各类型节点的分布情况，通过实验共检测出故障节点 11 个(检测率 91.7%)、事件节点 54 个(检测率 96.4%)、正常节点 109 个(检测率 97.3%)及误差节点 19 个(检测率 95.0%)，出现误报的节点总计 7 个(误报率 3.5%).

实验5.　为评估本文方法对不同类型异常节点的检测效果，分别使用一维至四维的数据集对算法进行测试，并与文献[16]基于 DBSCAN 的密度算法、文献[20]高可靠性的异常数据检验方法以及文献[23]核密度估计的方法进行对比. 下文以事件节点为例，给出检测率、误报率以及执行时间这 3 项性能指标的对比.

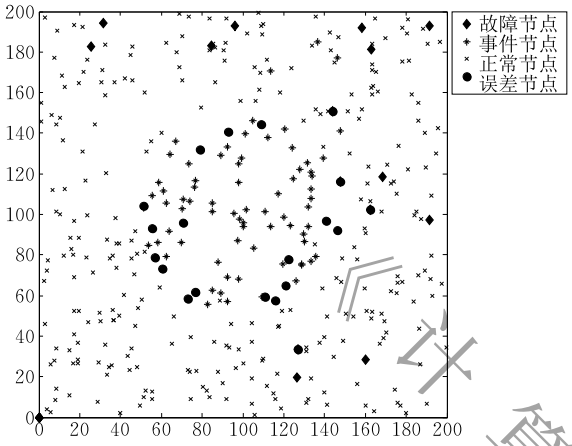

图 9　仿真实验检测结果

图 10 为不同方法的事件节点检测率的对比. 与以往方法相比，本文方法运用了节点的时空相关性和多模态数据流的相干性原理，结合多模态数据的异常数据对节点的异常类型进行判断，因此在高维度数据集(二维以上)的情况下，对事件节点的检测率达到 95% 以上，与以往方法相比提高了 2%～5%.

图 11 为不同方法的事件节点误报率的对比. 由于本文方法综合了多模态数据流的异常数据检测结果识别节点的异常来源的方法，同时考虑了不同模态的数据流对结果的影响，有效解决了以往存在的无差别处理不同模态数据集的问题，从而大幅降低了节点的误报率. 以往方法的误报率平均为 6%～8%，而本文方法稳定在 3.5%左右.

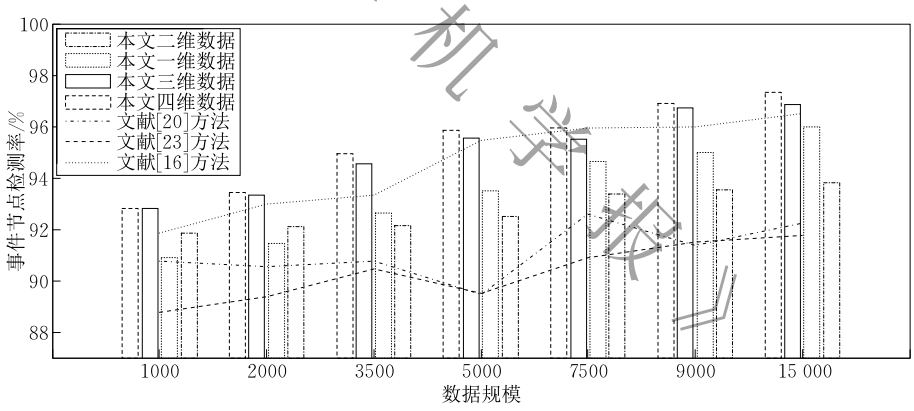

图 10　事件节点检测率的对比

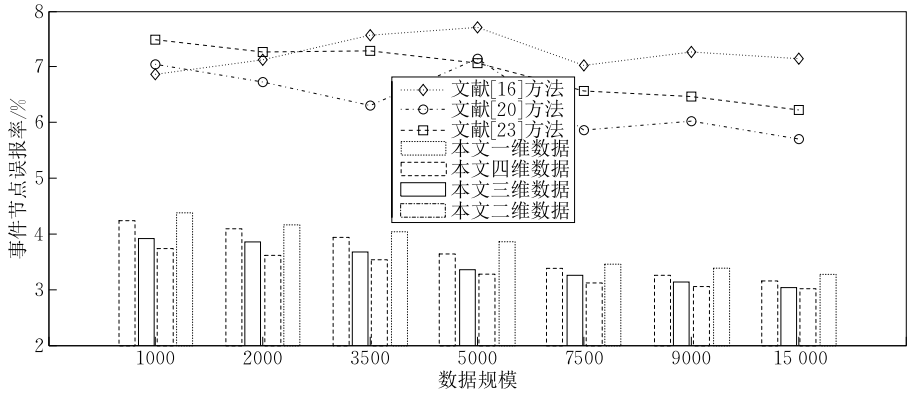

图 11　事件节点误报率的对比

表 3 为不同方法执行时间的对比. 由于本文的方法将传感器网络的多维数据集视作多维空间中的点集，采用基于欧氏距离的定义对异常数据点进行区分，有效减少了重复运算，提高了算法效率. 实验

结果也表明本文方法的执行时间随着数据维度变化的增幅明显小于传统方法,说明本文方法对于高维度数据集的处理具有良好的效果.

表 3 不同方法执行时间的对比

数据规模/组	执行时间/s						
	本文方法				文献[16]方法	文献[20]方法	文献[23]方法
	一维数据	二维数据	三维数据	四维数据	二维数据	一维数据	四维数据
1000	18.32	18.79	19.24	20.86	20.34	21.27	24.73
2000	24.87	25.24	27.97	29.89	28.78	26.14	34.94
3500	30.73	31.13	33.03	35.13	36.33	37.57	39.67
5000	36.67	38.78	39.82	40.78	40.41	40.72	45.71
7500	44.23	45.24	48.04	49.35	49.31	50.32	54.68
9000	52.31	54.86	56.02	58.03	57.46	59.52	65.92
15000	88.87	90.32	91.78	93.02	93.13	94.57	104.55

6 结束语

本文以提高 CPS 系统中无线传感器网络的可靠性为研究目标,以环境监测场景中的异常数据检测问题作为研究对象,结合现有的研究方法和成果,以多维数据和滑动窗口模型为基础,提出了一种无线传感器网络中多模态数据流的异常检测方法.该方法通过计算无线传感器网络数据集的统计特征来评价数据点的离散程度,利用传感器网络数据流之间的时空相关性以及多模态数据集之间的相干性,对异常数据的来源进行识别和验证,有效提高了算法的性能.实验结果表明,本文方法在大数据规模和高数据维度的情况下具有理想的效果.

参 考 文 献

[1] Yang Fan, Liu Yan, Li Ren-Fa, et al. A modeling method research based on data in cyber-physical system. Chinese Journal of Computers, 2016, 39(5): 961-972(in Chinese)
(杨帆,刘彦,李仁发等. 一种基于数据驱动的 CPS 建模方法研究. 计算机学报, 2016, 39(5): 961-972)

[2] Chen Hai-Ming, Cui Li. Design and model checking of service oriented software architecture for Internet of Things: A survey. Chinese Journal of Computers, 2016, 39(5): 853-871(in Chinese)
(陈海明,崔莉. 面向服务的物联网软件体系结构设计与模型检测. 计算机学报, 2016, 39(5): 853-871)

[3] Chen Hai-Ming, Cui Li, Xie Kai-Bin. A comparative study on architectures and implementation methodologies of Internet of Things. Chinese Journal of Computers, 2013, 36(1): 168-188(in Chinese)
(陈海明,崔莉,谢开斌. 物联网体系结构与实现方法的比较研究. 计算机学报,2013, 36(1): 168-188)

[4] Wu F J, Kao Y F, Tseng Y C. From wireless sensor networks towards cyber physical systems. Pervasive & Mobile Computing, 2011, 7(4): 397-413

[5] Lin C Y, Zeadally S, Chen T S, et al. Enabling cyber physical systems with wireless sensor networking technologies. International Journal of Distributed Sensor Networks, 2012, 2012(2012): 184-195

[6] Li Jian-Zhong, Gao Hong. Survey on sensor network research. Journal of Computer Research and Development, 2008, 45(1): 1-15(in Chinese)
(李建中,高宏. 无线传感器网络的研究进展. 计算机研究与发展, 2008, 45(1): 1-15)

[7] Wu Peng-Fei, Li Guang-Hui, Zhu Hong. Event boundary detection method based on wireless sensor network and linear neural network. Pattern Recognition and Artificial Intelligence, 2015, 28(4): 377-384(in Chinese)
(吴鹏飞,李光辉,朱虹等. 基于无线传感器网络和线性神经网络的事件边界检测方法. 模式识别与人工智能, 2015, 28(4): 377-384)

[8] Ma Zu-Chang, Sun Yi-Ning, Mei Tao. Survey on wireless sensors network. Journal on Communications, 2004, 25(4): 114-124(in Chinese)
(马祖长,孙怡宁,梅涛. 无线传感器网络综述. 通信学报, 2004, 25(4): 114-124)

[9] Mahapatro A, Khilar P M. Fault diagnosis in wireless sensor networks: A survey. IEEE Communications Surveys & Tutorials, 2013, 15(4): 2000-2026

[10] Peng Shang-Lian, Li Zhan-Huai, Chen Qun. Complex event processing over live archived data streams. Chinese Journal of Computers, 2012, 35(3): 540-554(in Chinese)
(彭商濂,李战怀,陈群等. 在线-离线数据流上复杂事件检测. 计算机学报, 2012, 35(3): 540-554)

[11] Pham D S, Venkatesh S, Lazarescu M, et al. Anomaly detection in large-scale data stream networks. Data Mining & Knowledge Discovery, 2014, 28(1): 145-189(45)

[12] Mamun Q, Islam R, Kaosar M. Anomaly detection in wireless sensor network. Journal of Networks, 2014, 9(11): 2914-2924

[13] Xie M, Han S, Tian B, et al. Anomaly detection in wireless sensor networks: A survey. Journal of Network & Computer Applications, 2011, 34(4): 1302-1325

[14] Koushanfar F, Potkonjak M. On-line fault detection of sensor measurements. IEEE Sensors Proceedings, 2003, 2: 974-979

[15] Krishnamachari B, Iyengar S. Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks. IEEE Transactions on Computers, 2004, 53(3): 241-250

[16] Pan Yuan-Yang, Li Guang-Hui, Xu Yong-Jun. Abnormal data detection method for environment wireless sensor networks based on DBSCAN. Computer Applications and Software, 2012, 29(11): 69-72(in Chinese)
(潘渊洋，李光辉，徐勇军. 基于 DBSCAN 的环境传感器网络异常数据检测方法. 计算机应用与软件，2012，29(11)：69-72)

[17] Samparthi V S K, Verma H K. Outlier detection of data in wireless sensor networks using kernel density estimation. International Journal of Computer Applications, 2010, 5(7): 28-32

[18] Zhang Jian-Ping, Li Bin, Liu Xue-Jun. Abnormal time series detection in wireless sensor network based on hadoop. Chinese Journal of Sensors and Actuators, 2014(12): 1659-1665(in Chinese)
(张建平，李斌，刘学军等. 基于 Hadoop 的异常传感数据时间序列检测. 传感技术学报，2014(12)：1659-1665)

[19] Lee M H, Choi Y H. Fault detection of wireless sensor networks. Computer Communications, 2008, 31(14): 3469-3475

[20] Lee D W, Kim J H. High reliable in-network data verification in wireless sensor networks. Wireless Personal Communications, 2010, 54(3): 501-519

[21] Bi Ran, Li Jian-Zhong. Energy efficient Top-k monitoring algorithm in wireless sensor networks. Journal of Computer Research and Development, 2014, 51(11): 2361-2373(in Chinese)
(毕冉，李建中. 无线传感器网络中能量高效的 Top-k 监测算法. 计算机研究与发展，2014，51(11)：2361-2373)

[22] Hu Shi, Li Guang-Hui, Feng Hai-Lin. Top-k(σ) outlier detection algorithm for wireless sensor networks. Journal of Nanjing University (Natural Sciences), 2016, 52(2): 261-269)(in Chinese)
(胡石，李光辉，冯海林. 基于 Top-k(σ)的无线传感器网络异常数据检测算法. 南京大学学报(自然科学)，2016，52(2)：261-269)

[23] Subramaniam S, Palpanas T, Papadopoulos D. Online outlier detection in sensor data using non-parametric models// Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul, Korea, 2006: 187-198

[24] Palpanas T, Papadopoulos D, Kalogeraki V, et al. Distributed deviation detection in sensor networks. ACM SIGMOD Record, 2003, 32(4): 77-82

[25] Ren Qian-Qian, Li Jian-Zhong, Cheng Si-Yao. Fault-tolerant event monitoring in wireless sensor networks. Chinese Journal of Computers, 2012, 35(3): 581-590(in Chinese)
(任倩倩，李建中，程思遥. 无线传感器网络中可容错的事件检测算法. 计算机学报，2012，35(3)：581-590)

[26] Zhang Shu-Kui, Wang Yi-Huai, Cui Zhi-Ming. Event region fault-tolerant detection algorithm based on aggregation tree. Journal on Communications, 2010(9): 74-87(in Chinese)
(张书奎，王宜怀，崔志明. 基于融合树的事件区域检测容错算法. 通信学报，2010(9)：74-87)

[27] Cao Dong-Lei, Cao Jian-Nong, Jin Bei-Hong. A fault-tolerant algorithm for event region detection in wireless sensor networks. Chinese Journal of Computers, 2007, 30(10): 1770-1776(in Chinese)
(曹冬磊，曹建农，金蓓弘. 一种无线传感器网络中事件区域检测的容错算法. 计算机学报，2007，30(10)：1770-1776)

[28] Akyildiz I F, Vuran M C, Akan O B. On exploiting spatial and temporal correlation in wireless sensor networks// Proceedings of Wiopt Modeling & Optimization in Mobile Ad Hoc & Wireless Networks. Avignon, France, 2010: 71-80

[29] Fei Huan. Anomaly Detection Method Research of Sensor Data Stream Based on Multi-Dimensional Data Model [M. S. dissertation]. Zhejiang A&F University, Lin'an, Zhejiang, 2015(in Chinese)
(费欢. 基于多维数据模型的传感器数据流异常检测方法的研究[硕士学位论文]. 浙江农林大学，浙江，临安，2015)

[30] Richard A J, Dean W W. Applied Multivariate Statistical Analysis. 6th Edition. New York, USA: Prentice Hall, 2007

[31] Knorr EM, Ng RT. Algorithms for mining distance-based outliers in large datasets//Proceedings of VLDB 1998 C J. New York, USA, 1998: 392-403

[32] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. Acm Sigmod Record, 2000, 29(2): 427-438

[33] Chang W. K-means clustering based on genetic algorithm. Computer Science, 2003, 37(3): 2696-2704

[34] Zhang Y, Hamm N A S, Meratnia N. Statistics-based outlier detection for wireless sensor networks. International Journal of Geographical Information Science, 2012, 26(8): 1373-1392

[35] Zhang Y, Meratnia N, Havinga P. Outlier detection techniques for wireless sensor networks: A survey. Communications Surveys & Tutorials, IEEE, 2010, 12(2): 159-170

FEI Huan, born in 1990, Ph. D. candidate. His research interests is wireless sensor networks.

XIAO Fu, born in 1980, Ph. D., professor, Ph. D. supervisor. His research interests include wireless sensor networks and Internet of Things.

LI Guang-Hui, born in 1970, Ph. D., professor, Ph. D. supervisor. His research interests include wireless sensor networks and intelligent nondestructive testing.

SUN Li-Juan, born in 1963, Ph. D., professor, Ph. D.

supervisor. Her research interest is wireless sensor networks.

WANG Ru-Chuan, born in 1943, Ph. D., professor, Ph. D. supervisor. His research interests include Internet of Things and network security.

Background

This paper focuses on the research about the anomaly data detection in wireless sensor networks (WSN), which is a highly significant part of Cyber Physical System (CPS). Many researches paid more attention on the method of single-dimensional data set and increase the detection accuracy, yet ignored the relevance among multiple-dimensional data. Actually, the wireless sensor nodes are often integrated multiple sensors that can get different types of data at same time. Special events usually make multi-dimensional data change simultaneously while multi-dimensional data is more isolated in sensor fault. Therefore, we can distinguish the source of anomaly data by proposed method in this paper. A distributed event region detection algorithm, Bayesian fault recognition algorithm (BFRA) is proposed by Krishnamachari B et al, BFRA based on the premise that event is spatial correlation but sensor fault is spatially uncorrelated. Every sensor exchanges readings to its all neighbors in the detected events to determine whether the occurrence of the events or not. Lee D W et al. proposed a high reliability wireless sensor network data validation methods, according to the node probability of anomaly, to filter the abnormal data and evaluate the sensor work state. However, this method only considers the one-dimensional data, there is no specific solution for multi-dimensional data. The method proposed in this paper meliorates the insufficient.

This research work is supported by the following funds: National Natural Science Foundation of China under Grant Nos. 61472368, 61373137, and 61572260; Key University Science Research Project of Jiangsu Province, China, with No. 14KJA520002; Six talent peaks project in Jiangsu Province, China, with No. 2013-DZXX-014.