

文章编号:1001-9081(2014)08-2145-03

doi:10.11772/j.issn.1001-9081.2014.08.2145

## 面向无线传感器网络的自适应数据清洗方法

夏英\*, 毕海洋, 雷建军, 裴海英

(重庆邮电大学 空间信息系统研究中心, 重庆 400065)

(\*通信作者电子邮箱 xiaying@cqupt.edu.cn)

**摘要:**针对无线传感器网络(WSN)数据不精确和不可靠的问题,根据感知数据的空间相关性定义了弹性空间模型,并在此基础上提出一种自适应近邻空间清洗方法(ANSA)。该方法根据感知数据波动动态调整近邻空间大小,并通过计算近邻节点测量数据的加权平均对本地数据清洗。实验结果表明,感知数据清洗后误差控制在0.5以内,与经典的加权移动平均(WMA)方法相比,所提方法的精确度更高,同时能量损耗减少约36%。

**关键词:**无线传感器网络;空间相关性;数据清洗;孤立点检测;数据可靠性

**中图分类号:** TP311; TP393.032    **文献标志码:**A

### Adaptive approach for data cleansing in wireless sensor networks

XIA Ying\*, BI Haiyang, LEI Jianjun, BAE Haeyoung

(Research Center of Spatial Information System, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

**Abstract:** Since the data gathered in Wireless Sensor Network (WSN) are inaccurate and unreliable, a flexible space model based on the spatial correlation of sensor data was defined, and an adaptive neighbor-space approach for data cleansing (ANSA) was proposed. The approach adjusted neighbor-space dynamically according to sensor data fluctuation and calculated the weighted average of neighbors' measurements to clean local raw data. The experimental results show that, the sensor data error after cleansing by the proposed approach is less than 0.5, and compared to the classic Weighted Moving Average (WMA), it is more accurate and the energy consumption is reduced by about 36%.

**Key words:** Wireless Sensor Network (WSN); spatial correlation; data cleansing; outlier detection; data reliability

### 0 引言

近年来,无线传感器网络(Wireless Sensor Network, WSN)被广泛应用于环境感知、智能监测、行为分析等领域。但由于传感器易受环境等因素的影响,WSN中获得的感知数据往往含有噪声和误差,并经常出现数据丢失、重复和不一致现象。由于感知数据的不精确性和不完整性<sup>[1]</sup>,WSN难以对现实世界中的状态或事件进行精确监测,因此在利用感知数据之前,先对其进行清洗以保证数据的精确性和可靠性是极其重要的。

感知数据的时空相关性<sup>[2]</sup>是WSN区别于传统数据流的一个重要的特性,也是被研究人员广泛用于数据清洗的特性,这种特性可以概括如下。

1)时间相关性。WSN在环境监测、事件跟踪等应用中,感知节点所在环境发生的现象或事件会持续一段时间,不会立即消失,即该感知节点的测量值在这段时间内(几个或多个测量周期)应该保持相对稳定,不会发生急剧变化。

2)空间相关性。在WSN应用中,现实世界发生的现象或事件会覆盖WSN的某个区域(称为事件区域(Event Area)),而不是一个位置点,即在现象或事件发生区域内的感知节点,在同一时刻应具有相近的测量值,且在一段时间内

(几个或多个测量周期)其测量值具有相似的变化趋势。

在WSN应用领域,基于时空相关性的传感数据清洗技术得到广泛研究,主要分为集中式清洗方法和网内清洗方法。

1)集中式数据清洗方法。如Jeffery等<sup>[3]</sup>建立的传感器数据清洗模型,利用感知数据的时空相关性来恢复缺失数据和去除孤立点;Sheng等<sup>[4]</sup>提出使用直方图表示数据的分布式提示信息(hints)并使用hints滤出不必要的数据和识别可能的孤立点。

2)网内数据清洗方法。如Zhuang等<sup>[5]</sup>提出的加权移动平均(Weighted Moving Average, WMA)方法,通过本地节点测试和邻居节点测试相结合的方法进行感知数据的去噪,从而减少本地节点采样的能量消耗,并加快感知数据的响应速度;郭龙江等<sup>[6]</sup>在WMA基础上提出基于节点密度的混合式方法(Density-based Hybrid Approach, DHA),根据节点密度(单跳通信范围内近邻节点个数)来动态调整算法,以达到有效去噪和节省能量的目的;Branch等<sup>[7]</sup>提出一种网内孤立点检测方法(全局检测和半全局检测),只使用单跳通信获取近邻节点数据,通过计算感知数据的估计值识别孤立点。

集中式清洗方法由于要将大量感知数据传送至Sink节点进行集中处理,因此不能很好地满足WSN应用的实时性要求;同时,由于大量数据的传输引起的能量消耗,也使得集中

收稿日期:2014-04-07;修回日期:2014-05-07。 基金项目:重庆市自然科学基金资助项目(cstc2012jjA4014);重庆市基础与前沿研究项目(cstc2013jcyjA40023);重庆邮电大学青年科学项目(A2012-90)。

作者简介:夏英(1972-),女,重庆人,教授,博士,CCF会员,主要研究方向:数据库与数据挖掘、移动定位与位置服务、云计算与云服务;毕海洋(1989-),男,河北玉田人,硕士研究生,主要研究方向:数据挖掘、空间信息处理;雷建军(1976-),男,重庆人,副教授,博士,主要研究方向:无线传感器网络;裴海英(1948-),男,韩国人,教授,主要研究方向:大数据分析与处理、空间数据库、地理信息系统、多媒体数据库。

式清洗方法低效而难以普及。网内数据清洗方法与集中式数据清洗方法相比,感知数据直接在节点内部清洗,具有良好的实时性和节能特性,但这些方法<sup>[5~7]</sup>没有考虑近邻节点的空间相关度。

综上所述,本文定义了一种数据清洗的弹性空间模型,并在此基础上提出网内的自适应近邻空间清洗方法(Adaptive Neighbor-Space Approach, ANSA)。ANSA 通过动态调整近邻空间大小控制能耗,并使用空间相关度较高的近邻节点测量数据的加权平均识别孤立点和去噪。最终,通过仿真验证了 ANSA 方法的可行性和效果。

## 1 面向数据清洗的自适应近邻空间

### 1.1 近邻空间及其对数据清洗的影响

感知节点的近邻空间是指某一节点根据某一度量依据与其他节点共同形成的空间,这些节点互为近邻节点。近邻空间的度量依据可以是地理距离或路由跳数等信息,也可以人为指定。在 WSN 中,由于感知节点密集分布且可以利用某些定位技术获得其地理位置信息,因此,往往根据地理位置将网络节点划分在不同的空间。

基于近邻空间的数据清洗过程,是以近邻节点的感知数据作为依据,消除本地节点数据不确定性的过程。在 WSN 中,节点的近邻空间越大,近邻节点数量越多,节点间感知数据的整体空间相关度就越低。空间相关度越低,近邻节点所提供的信息不确定性越高,可能导致数据清洗结果不准确。另外,更多的节点意味着更多的数据通信和更多的能量消耗。当节点近邻空间过小时,由于近邻节点提供的信息不足,也不能很好地完成数据清洗任务。因此,根据本地传感器数值的不确定程度和节点的空间相关性选择合适大小的近邻空间,是有效完成数据清洗任务的前提。

### 1.2 弹性空间模型

在 WSN 应用中,某一感知节点当前传感器数值与上一周期比较具有较大的波动时,此次测量值具有较大的不确定性。因而,为消除此次测量值的不确定性需要更多近邻节点的信息。弹性空间是指在保证各节点具有较高空间相关性的前提下,随感知数据的波动而变化的近邻空间。弹性空间模型定义如下:

$$S = \lceil R \cdot e^{\lambda \cdot \Delta^2} \rceil \quad (1)$$

其中: $S$  为近邻空间大小,表示数据清洗需要使用的关系程度最高的近邻节点数量; $R$  为当前近邻空间的整体相关度( $0 < R \leq 1$ ); $e$  为数学常数; $\Delta$  表示测量值变化,是当前测量值与上一周期修正值的差; $\lambda$  为波动调节参数( $\lambda > 0$ )。在某些 WSN 中,由于传感器质量较差,测量值波动较大,可将  $\lambda$  设为小于 1 的值,从而减小空间规模,控制能耗。

在实际应用中,为保证数据可靠性,设定弹性空间下限;同时,为避免近邻空间过大,设定弹性空间的上限,并将近邻空间大于上限的测量值标记为孤立点。因此,该模型不仅具有良好的自适应性,而且具有识别孤立点的特性。

## 2 自适应的近邻空间清洗方法

### 2.1 节点间的空间相关性度量

在 WSN 中,两节点位置越接近,那么两节点相关性可能

越大,因此可使用径向基函数根据两节点间的欧氏距离来度量其相关性。本文使用经典的径向基函数——尺度高斯核函数来度量节点间的空间相关性:

$$r(i, j, \sigma) = e^{-\frac{dis(i, j)^2}{2\sigma^2}} \quad (2)$$

其中: $e$  为数学常数; $dis(i, j)$  表示两节点  $i$  与  $j$  之间的距离; $\sigma$  是函数的宽度参数,控制函数的径向作用范围。通过调节参数  $\sigma$ ,可以避免距离较远、相关度较低的节点感知数据进入数据清洗过程,从而保证数据清洗质量。

为了衡量整个近邻空间的空间相关度,可以计算各节点与本地节点  $i$  空间相关度的平均值,如式(3)所示:

$$R(i) = \left( \sum_{j \in N(i)} r(i, j, \sigma) \right) / |N(i)| \quad (3)$$

其中: $R(i)$  是节点  $i$  的近邻空间整体相关度; $N(i)$  是节点  $i$  的近邻节点集,  $|N(i)|$  表示近邻节点数; $j \in N(i)$ , 是节点  $i$  的近邻节点。

### 2.2 数据去噪与孤立点识别

设定弹性空间的下限值  $\omega$  和上限值  $\alpha$ ,本地节点感知数据的去噪和孤立点识别方法如下。

1) 对于近邻空间  $S$  大于  $\alpha$  的节点测量值,将其识别为孤立点,因其具有较大的不确定性,故只使用近邻节点测量值的加权平均去噪:

$$\bar{x}(i, t) = \left( \sum_{j \in N(i)} x(j, t) \cdot r(i, j, \sigma) \right) / \alpha \quad (4)$$

其中: $x(j, t)$  代表节点  $j$  在  $t$  时刻的原始测量值; $\bar{x}(i, t)$  是节点  $i$  在  $t$  时刻的修正值; $r(i, j, \sigma)$  由式(2)计算所得,作为近邻节点  $j$  测量值的权重。

2) 对于近邻空间  $S$  小于或等于  $\alpha$  的测量值,使用本地节点测量值(权重为 1)和近邻节点测量值的加权平均去噪:

$$\bar{x}(i, t) = \frac{x(i, t) + \sum_{j \in N(i)} x(j, t) \cdot r(j, i, \sigma)}{|N(i)| + 1} \quad (5)$$

### 2.3 ANSA

ANSA 将初始近邻空间设定为最小近邻空间,即由相关性最高的前  $\omega$  个节点和本地节点组成的空间,并假设附近节点与本地节点的空间相关度已经通过计算获得。

ANSA 伪代码如下。

```

Input 弹性空间的上限值  $\alpha$  和下限值  $\omega$ , 当前测量值和上一周期
      修正值。
Output 清洗后的本地节点感知数据。
Begin:
    1) 计算近邻空间整体相关程度  $R$ ;
    2) 计算节点前后两次的观测偏差  $\Delta$ ;
    3) 计算近邻空间大小  $S$ ;
    4) If  $S > \alpha$ 
         $N = \alpha$ ;
    Else If  $S < \omega$ 
         $N = \omega$ ;
    Else
         $N = S$ ;
    5) 获取空间相关性最高的  $N$  个近邻节点感知数据;
    6) If  $S > \alpha$ 
        计算近邻节点测量值的加权平均;
    Else
        计算本地节点和近邻节点测量值的加权平均;

```

End

算法中第1)~3)行计算近邻空间大小;第4)行判断近邻空间是否超出弹性空间范围,若超出则进行调整;第5)~6)行进行感知数据去噪和孤立点识别。由于本地节点数据清洗需要与近邻节点交换感知数据,因此算法的时间性能主要取决于通信延迟。

### 3 实验仿真

为测试ANSA在WSN数据清洗中的可行性和效果,仿真实验使用Matlab模拟环境温度监测,并将结果与经典的WMA方法进行对比。WMA同时采用时间和空间两维数据,并计算它们的加权移动平均值,对于光滑噪声具有比较显著的效果。

#### 3.1 去噪效果实验

实验中,弹性空间的上限值和下限值分别设定为12和3,距本地节点最近的15个节点的空间相关度依次设为0.98,0.95,0.92,0.88,0.87,0.84,0.83,0.78,0.78,0.74,0.72,0.72,0.70,0.70。实验假设本地节点位置的真实温度为20°C,空间相关度在0.9以上的节点位置真实温度为 $(20 \pm 0.1)$ °C,空间相关度在[0.8,0.9]和[0.7~0.8]之间的节点所处位置的真实温度分别为 $(20 \pm 0.3)$ °C和 $(20 \pm 0.5)$ °C。各节点在上述范围内随机产生温度值,并在此基础上添加均值为0、方差为1的高斯白噪声。每个节点产生100个数据,仿真各节点连续100个周期的测量值。同时,为了对比ANSA的特性,将WMA的时间滑动窗口设定为一般大小10,而将其近邻节点数设为小于弹性空间上限的值8。

去噪效果如图1所示,原始数据(noisy data)误差较高,最高误差超出2.5;ANSA与WMA去噪后感知数据的误差都控制在0.5以下,去噪效果良好。

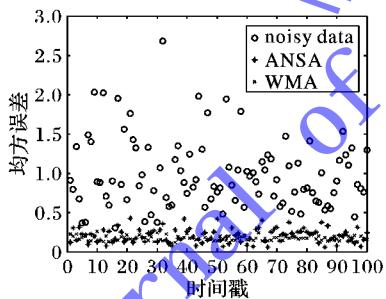


图1 两种算法的去噪效果对比

#### 3.2 孤立点检测实验

在去噪实验的基础上,对40~60时间截的温度数据增加2°C模拟高温事件,仿真结果如图2所示。

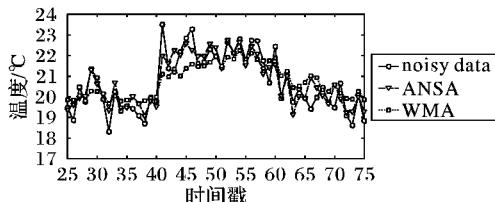


图2 两种算法的孤立点检测结果对比

第41个时间截,温度骤然上升,原始数据出现明显的孤立点,ANSA去噪之后修正值接近真实温度(22°C),而WMA去噪结果明显低于真实值,较长时间后才接近真实值。同样,

高温事件消失时,ANSA迅速下降,也优于WMA算法。

实验分析可知,由于WMA使用历史数据,在一定程度上影响了对当前事件的判断,出现了过平滑现象。而ANSA方法对于孤立点的识别则仅使用近邻节点测量数据,因此对于孤立点的识别更为精确、对突发事件反应更快,更能有效反映现实世界真实状况。

#### 3.3 通信能耗评估

在数据清洗中,使用进行数据交换的近邻节点数量来评估ANSA的通信能耗。近邻空间越小,通信的近邻节点数量越少,意味着通信能耗越低。通过对100次实验进行统计,ANSA平均需要和5.1个近邻节点通信,而WMA始终需要和8个近邻节点通信,在采用相同通信协议的情况下,ANSA能相对减少约36%的通信能耗。在不同网络部署环境和不同参数设定下,通信能耗统计结果可能会有差异,但与其他固定近邻节点数的清洗方法相比,在达到同样清洗效果的情况下ANSA更节省能耗。

### 4 结语

本文首先探讨了无线传感器网络中感知数据的不确定性和节点间感知数据的空间相关性,进而提出一种可以提高数据可靠性且节省能耗的自适应清洗方法。本文方法综合考虑感知数据波动、节点间的空间相关度和节点的通信能耗,从而确定数据清洗使用的近邻节点数量。实验结果表明该方法可行、有效,更适合实时性要求较高的WSN应用。下一步研究将考虑结合各节点感知数据的时间相关性对感知数据进行清洗。

#### 参考文献:

- [1] MARTINCIC F, SCHWIEBERT L. Distributed event detection in sensor networks [C]// ICSNC'06: Proceedings of the 2006 IEEE International Conference on Systems and Networks Communications. Piscataway: IEEE Press, 2006: 43.
- [2] AKYILDIZ I F, VURAN M C, AKAN O B. On exploiting spatial and temporal correlation in wireless sensor networks [C]// WiOpt'04: Proceedings of the 2004 Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks. Cambridge: University of Cambridge, 2004: 71~80.
- [3] JEFFERY S R, ALONSO G, FRANKLIN M J, et al. Declarative support for sensor data cleaning [C]// Proceedings of the 4th International Conference on Pervasive Computing. Berlin: Springer, 2006: 83~100.
- [4] SHENG B, LI Q, MAO W, et al. Outlier detection in sensor networks [C]// Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing. New York: ACM Press, 2007: 219~228.
- [5] ZHUANG Y, CHEN L, WANG X S, et al. A weighted moving average-based approach for cleaning sensor data [C]// ICDCS'07: Proceedings of the 27th IEEE International Conference on Distributed Computing Systems. Piscataway: IEEE Press, 2007: 38~38.
- [6] GUO L, FU H, ZHANG Z. Adaptive method for cleaning sensory data in wireless sensor networks [J]. Computer Engineering and Applications, 2009, 45(13): 150~155. (郭龙江,付惠娟,张中兆. 传感器网络感知数据自适应去噪方法[J]. 计算机工程与应用, 2009, 45(13): 150~155.)

(下转第2154页)

有效地均衡了网络中簇头节点的分布,整体上均衡了网络的负载,降低了网络能耗并延长了网络生命周期。

## 5 结语

本文提出一种基于 LEACH 的混合优化的改进路由协议——HOBDE-LEACH,从分簇机制及数据通信两方面对 LEACH 协议进行了优化。协议在网络管理方面采用集中式与分布式相结合的方法,给出覆盖半径种子扫描成簇算法进行快速分簇,保证对区域的全覆盖。网络运行机制结合能量和距离考虑负载均衡,合理地将整个网络的运行分为稳定和轮循两个阶段,对这两个阶段的簇头选举和通信机制采用不同的运行策略。在簇头与基站进行通信时,采用了基于分布式的传输管理策略,同时考虑了距离的影响。通过与相关算法的比较分析实验表明,HOBDE-LEACH 减少了网络总能耗,增强了网络的鲁棒性,延长了网络生命周期。当节点的初始能量比较大的时候,网络的效率将会有更加显著的提高。在此基础上,今后将通过实践进一步研究无线传感网络的广泛应用。

## 参考文献:

- [1] YU H, ZENG P, LIANG W. Intelligent wireless sensor networks [M]. Beijing: Science Press, 2006. (于海斌,曾鹏,梁桦.智能无线传感器网络系统[M].北京:科学出版社,2006.)
- [2] YU H, ZENG P, WANG Z, et al. Study of communication protocol of distributed sensor network [J]. Journal on Communications, 2004, 25(10): 102–110. (于海斌,曾鹏,王忠锋,等.分布式无线传感器网络通信协议研究[J].通信学报,2004,25(10):102–110.)
- [3] HEINZELMAN W. Application-specific protocol architectures for wireless networks [D]. Boston: Massachusetts Institute of Technology, 2000.
- [4] WU C, XU B, YE Y. Routing method research based on energy intensity and clustering [J]. Control Engineering of China, 2012, 19(4): 566–570. (邬春学,许博威,叶胤鹏.基于能量强度的簇区分路由方法研究[J].控制工程,2012,19(4):566–570.)
- [5] CHEN Z, LUO P, YUE W, et al. An energy-aware topology control algorithm for wireless sensor networks [J]. Chinese Journal of Sensors and Actuators, 2013, 26(3): 382–387. (陈志,骆平,岳文静.一种能量感知的无线传感网拓扑控制算法[J].传感技术学报,2013,26(3):382–387.)
- [6] HEINZELMAN W, CHANDRAKASAN A, BALAKRISHNAN H. Energy-efficient communication protocol for wireless microsensor networks [C]// Proceedings of the 33rd Annual Hawaii International Conference on System Sciences. Washington, DC: IEEE Computer Society, 2000: 3005–3014.
- [7] LU Y, CHEN Y, CHEN M. The improvement and simulation research of wireless sensor network LEACH protocol [J]. Journal of Anhui University of Engineering, 2012, 27(4): 42–45. (鲁玉定,陈跃东,陈孟元.无线传感器网络 LEACH 协议改进与仿真研究[J].安徽工程大学学报,2012,27(4):42–44.)
- [8] GARGARI E A, HASHEMZADEH F, RAJABIOUN R. Colonial competitive algorithm a novel approach for PID controller design in MIMO distillation column process [J]. International Journal of Intelligent Computing and Cybernetics, 2008, 1(3): 337–355.
- [9] HUANG H, YAO D, SHEN J, et al. A multi-weight based clustering algorithm for wireless sensor networks [J]. Journal of Electronics and Information Technology, 2008, 30(6): 1489–1492. (黄河清,姚道远,沈杰,等.一种基于多权值优化的无线传感网分簇算法的研究[J].电子与信息学报,2008,30(6):1489–1492.)
- [10] BAI F, WANG L, MA Y, et al. Algorithm analysis of routing protocols-LEACH for wireless sensor networks [J]. Journal of Taiyuan University of Technology, 2009, 40(4): 248–252. (白凤娥,王莉莉,马艳艳,等.无线传感器网络路由协议 LEACH 的算法分析[J].太原理工大学学报,2009,40(4),248–252.)
- [11] WU H. An improved energy and distance LEACH clustering algorithm in wireless sensor network [J]. Chinese Test, 2012, 38(5): 62–66. (邬厚民.无线传感网络中能量和距离改良的 LEACH 分簇算法[J].中国测试,2012,38(5):62–66.)
- [12] FAN Z, JIN Z, XIE D. Energy-efficient clustering algorithm for wireless sensor networks [J]. Journal of Chinese Computer Systems, 2013, 34(3): 535–539. (樊志平,金政哲,谢冬青.基于能量效率的无线传感网络分簇算法[J].小型微型计算机系统,2013,34(3):535–539.)
- [13] LI J, CAO B, WANG L, et al. An improved cluster routing algorithm for wireless sensor network [J]. Journal of Hunan University of Arts and Science: Natural Science, 2012, 24(2): 51–55. (李建奇,曹斌芳,王立,等.一种基于 LEACH 的无线传感器网络改进路由算法[J].湖南文理学院学报:自然科学版,2012,24(2):51–55.)
- [14] IQBAL A, AKBAR M. Advanced LEACH: a static clustering based heterogeneous routing protocol for WSNs [J]. Journal of Basic and Applied Scientific Research, 2013, 3(5): 864–872.

(上接第 2147 页)

- [7] BRANCH J W, GIANNELLA C, SZYMANSKI B, et al. In-network outlier detection in wireless sensor networks [J]. Knowledge and Information Systems, 2013, 34(1): 23–54.
- [8] ZHANG Y, MERATNIA N, HAVINGA P. Outlier detection techniques for wireless sensor networks: a survey [J]. IEEE Communications Surveys and Tutorials, 2010, 12(2): 159–170.
- [9] FRANKE C, GERTZ M. ORDEN: outlier region detection and exploration in sensor networks [C]// Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2009: 1075–1078.
- [10] JEFFERY S R, GAROFALAKIS M, FRANKLIN M J. Adaptive

- cleaning for RFID data streams [C]// Proceedings of the 32nd International Conference on Very Large Data Bases. New York: ACM Press, 2006: 163–174.
- [11] ZHANG Z, YANG D, ZHANG T, et al. A study on the method for cleaning and repairing the probe vehicle data [J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(1): 419–427.
- [12] FANG L, DOBSON S. In-network sensor data modelling methods for fault detection [C]// Aml-2013: Proceedings of the Fourth International Joint Conference on Ambient Intelligence. Berlin: Springer, 2013: 176–189.