基于相似度的无线传感器网络数据复原汇聚方法*

罗永健，史德阳，侯银涛，侯　申
（西安通信学院 电子工程系，西安 710106）

摘　要：针对现有数据复原汇聚算法的复原汇聚精度低和对网络噪声干扰的稳健性差等不足，提出了一种基于相似度的无线传感器网络数据复原汇聚算法。该算法在分布式数据汇聚模型的基础上，利用重心距离和相关系数来衡量各簇节点感知数据的受攻击程度，并把两者统一在加权系数中，通过加权运算提高了算法的数据复原汇聚精度。此外，利用相关系数对噪声干扰敏感的特点提高了算法对网络噪声干扰的稳健性。理论分析和仿真结果表明，新算法的性能优于现有数据复原汇聚算法。

关键词：无线传感器网络；数据复原汇聚；相似度；安全
中图分类号：TP274　　**文献标志码**：A　　**文章编号**：1001-3695(2012)09-3405-03
doi:10.3969/j.issn.1001-3695.2012.09.054

Resilient data aggregation method based on similarity in wireless sensor networks

LUO Yong-jian, SHI De-yang, HOU Yin-tao, HOU Shen
(*Dept. of Electronic Engineering, Xi'an Communications Institute, Xi'an* 710106, *China*)

Abstract: For the low aggregation precision and bad performance against the noise jamming of the existing resilient data aggregation algorithm in wireless sensor networks, this paper proposed an algorithm based on the similarity of the sensor readings. On the base of the distributed model of data aggregation, it applied the gravity distance and correlation coefficient as measurements, and then combined into the weighted operation. Meanwhile, the correlation coefficient was sensitive to the noise jamming, which could improve the robustness of the new algorithm. Theoretical analysis and simulation results show that the performance of the new algorithm is better than that of the existing resilient data aggregation algorithm.

Key words: wireless sensor networks(WSN); resilient data aggregation; similarity; security

数据汇聚是 WSN 中消除数据冗余、降低网络能耗、延长网络寿命的一项关键技术。受节点部署环境和通信信道的影响，数据汇聚在安全问题上面临着严峻的挑战[1]。特别是当 WSN 中部分节点被俘获时，聚合函数的输入值会随之改变，进而会增加聚合函数的输出误差。故在汇聚前需对数据汇聚的输入值进行检测。然而在检测到攻击时，传统的方法将受到攻击节点感知的数据直接丢弃[2,3]，这样对数据资源造成了极大的浪费，降低了网络能量的利用率。

针对以上问题，Wager[1] 最早提出了数据复原汇聚的概念，并指出其所需解决的具体问题，同时相应地提出了几种简单的应对方案，如剪切法、截断法等。Buttyan 等人[4] 在 Wager 工作的基础上，提出了一种基于随机取样一致性检验的数据复原汇聚算法(RANBAC-based resilient aggregation, RANBAR)，但该算法采用集中式数据处理模式，节点能耗较大，且基站处容易形成网络瓶颈。当网络中不存在攻击时，该算法仍会剔除大量边缘数据，降低了数据汇聚精度。文献[5]提出一种基于信息效用测度的方法(information utility measure based resilient data aggregation, IUMRDA)，该方法在分布式数据汇聚模型的基础上，用信息效用测度衡量各簇节点的受攻击程度，根据受攻击程度对各簇节点感知数据赋予相应的权值，然后进行加权聚合运算，这在一定程度上提高了数据复原汇聚精度。但是当敌方对被俘获节点感知数据施加的攻击增量较小时，其复原汇聚精度较低。此外，以上算法均未考虑实际网络中噪声干扰的影响，对网络噪声干扰的稳健性较差。基于此，文献[6]提出一种基于灰色关联度和概率密度距离的数据复原汇聚方法(gray relationship degree and probability density parallel distance based resilient data aggregation, GPDRDA)，该方法的抗噪声能力较强，且复原汇聚精度较高，但仍有待进一步改进。基于此，本文提出一种基于相似度的无线传感器网络数据复原汇聚算法。

1　数据复原汇聚模型

首先利用分簇算法将整个网络划分成若干个簇，每个簇选举自己的簇头，传感器节点监测到数据后将数据直接发送给它所在簇的簇头，簇头节点对簇内数据融合处理后发送给基站[7]。其模型如图 1 所示。在上述分布式数据汇聚模型基础上，本文算法作以下假设：

a)基站是足够安全的，攻击者只能对普通传感器节点进行攻击。由于基站通常采用较强的防护措施，故不易受到攻击；而传感器节点是一种价格低廉、结构松散的设备，极易受到

收稿日期：2011-12-23；**修回日期**：2012-02-14　　**基金项目**：国家自然科学基金资助项目(61179002)；陕西省自然科学基础研究计划资助项目(2011JM8030)；国家自然科学青年基金资助项目(61102160)

作者简介：罗永健(1971-)，男，教授，博士，主要研究方向为阵列信号处理(hfl19680710@163.com)；史德阳(1986-)，男，硕士研究生，主要研究方向为无线传感器网络技术；侯银涛(1981-)，男，讲师，硕士，主要研究方向为无线通信技术；侯申(1983-)，男，讲师，硕士，主要研究方向为保密通信.

恶意攻击[8]。

b)节点攻击方式为加性攻击[4]。加性攻击有增量攻击和常量攻击两种。对每个被俘获节点读数都增加一个相同值 *add* 即为增量攻击,修改被俘获节点读数为同一个值 d 即为常量攻击。

c)被俘获节点的数量有限且分布相对集中。假设被俘获节点占所有传感器节点的比例 $k \leqslant 0.5$。事实上,攻击者往往是根据条件在操作方便的范围内选择节点进行攻击,故假设被俘获节点相对集中于网络中的某些簇是合理的。

2 基于相似度的数据复原汇聚方法

为了表述方便,特定义以下符号:n 表示 WSN 中节点的总数目;r 表示 WSN 中分簇的数目;C_i 表示 WSN 的第 i 个簇;m_i 表示簇 C_i 中所含节点的数目;X_{ij} 表示簇 C_i 中第 j 个节点的读数,假定未遭受攻击的传感器节点读数 X_{ij} 服从独立同分布,且期望 μ 和方差 δ^2 均未知;k 表示受攻击节点占所有节点的比例;S_i 表示簇 C_i 中所有节点在某一时刻感知数据的集合,即 $(X_{i1}, X_{i2}, \cdots, X_{ij})$。

考虑聚合函数为求均值的情况,令 X 为聚合函数要得到的目标变量的真值,$\hat{X}$ 为目标变量的估计值,distortion 为估计值与真值之间的绝对偏差,$\hat{X}_i$ 为各簇数据样本的目标变量估计值,有

$$\text{distortion} = |\hat{X} - X| = |\text{average}(S_1, S_2, \cdots, S_r) - X| \tag{1}$$

令 $\overline{X_i}$ 为簇 C_i 中所有节点读数的平均值,即

$$\overline{X_i} = \frac{1}{m_i}\sum_{j=1}^{m_i} X_{ij} \quad i = 1, 2, \cdots, r \tag{2}$$

2.1 相似度定义

首先采用卡方拟合检验法[9]选取一个未遭受攻击或者受攻击最弱的簇作为期望模型 C_z,通过比较 C_i 与 C_z 中数据的相似度大小来判断 C_i 中数据的受攻击程度。

定义 1 令簇 C_i 与 C_z 中数据的均值分别为 $\overline{X_i}$ 和 $\overline{X_z}$,则簇 C_i 与 C_z 的重心距离为

$$d_i = |\overline{X_i} - \overline{X_z}| \tag{3}$$

定义 2 令簇 C_i 与 C_z 中节点的感知数据分别为 $(X_{i1}, X_{i2}, \cdots, X_{ij})$ 和 $(X_{z1}, X_{z2}, \cdots, X_{zj})$,则簇 C_i 与 C_z 的相关系数为

$$r_i = \frac{\sum_{j=1}^{m}(X_{ij} - \overline{X_i}) \times (X_{zj} - \overline{X_z})}{\sqrt{\sum_{j=1}^{m}(X_{ij} - \overline{X_i})^2 \times \sum_{j=1}^{m}(X_{zj} - \overline{X_z})^2}} \tag{4}$$

定义 3 簇 C_i 与 C_z 中节点数据的相似度为

$$\theta_i = \lambda \times r_i + \frac{\eta}{d_i} \tag{5}$$

其中:$\lambda + \eta = 1$。

相似度用来表示两个数据或者样本之间的相似程度。由式(5)可知,两样本的相关系数越大,相似度越大;两样本间的重心距离越大,相似度越小。由式(3)(4)可知,重心距离反映的是簇 C_i 与 C_z 均值间的差异,相关系数体现的是 C_i 与 C_z 中各节点感知数据之间的差异。当攻击增量较小时,相关系数能反映 C_i 中受攻击数据和期望模型的差异。当网络中存在噪声干扰时,各节点感知数据的扰动也能准确地体现相关系数的变化;而当攻击增量较大时,重心距离则能准确地反映簇 C_i 与 C_z 的相似度。

2.2 新算法的基本原理

借鉴 GPDRDA 方法的思想,为簇 C_i 定义一个新的加权系数,并将重心距离和相关系数统一在该加权系数中,即

$$Z_i = a \times w_i + (1 - a) \times v_i \tag{6}$$

其中:w_i 为簇 C_i 相关系数所对应权值,v_i 为簇 C_i 重心距离所对应的权值。在攻击增量较小时,令 w_i 的系数 $a = 1$;而在攻击增量较大时,则令 $a = 0$。

在此引入预判机制对攻击增量的大小进行判断。各节点可存储其前 t 时刻的感知数据,t 值由各传感器节点的存储空间和计算能力决定。令簇 C_i 中第 j 个节点在前 t 时刻的感知数据为 $X_{ij}(1), X_{ij}(2), \cdots, X_{ij}(t)$。可利用其历史时刻数据来判别当前时刻的感知数据是否可信,具体方法为

$$x_{ij0} = \frac{1}{t}\sum_{p=1}^{t} X_{ij}(p) \tag{7}$$

$$\delta_{ij0} = \sqrt{\frac{\sum_{p=1}^{t}(X_{ij}(p) - x_{ij0})^2}{t-1}} \tag{8}$$

利用格罗贝斯(Grubbs)计算统计量为[10]

$$g_{ij} = \frac{X_{ij}(t) - x_{ij0}}{\delta_{ij0}} \tag{9}$$

在给定的显著性水平 α 下,可以求出格罗贝斯统计量的临界值 $g_{ij\alpha}$。把当前时刻的感知数据代入到式(9)中,当 g_{ij} 超过 $g_{ij\alpha}$ 时,则当前节点不可信,并把自身的信任度上传给簇头,最终在基站处得到整个网络中不可信节点的个数 K。进而在基站处,可估算出攻击者所施加的攻击增量为

$$d = \frac{(\hat{X}(t) - \hat{X}(t-1))}{k} \tag{10}$$

其中:$k = \frac{K}{n}$;$\hat{X}(t)$ 为当前时刻的数据汇聚结果(利用聚合函数直接汇聚得到);$\hat{X}(t-1)$ 为前一时刻网络中所有节点均可信时的汇聚结果。

式(6)中的 w_i 和 v_i 可利用拉格朗日极值法求得[9]

$$w_i = \frac{r_i^2}{\sum_{j=1}^{r-1} r_j^2} \tag{11}$$

$$v_i = \frac{1/d_i^2}{\sum_{j=1}^{r-1}(1/d_j^2)} \tag{12}$$

除了构成期望模型的簇外,其余簇均需要计算权系数。式(11)(12)中 $r-1$ 为 WSN 中参与加权聚合运算的簇数目。

利用式(6)将 w_i 和 v_i 统一起来即可得到簇 C_i 的最终加权系数。对各簇样本的估计值进行如下加权运算:

$$\hat{X}' = \sum_{i=1}^{r-1} Z_i \times \hat{X}_i \tag{13}$$

最后,将构成期望模型的簇与利用式(13)求得的值再进行一次聚合,即可得到最终的聚合结果:

$$\hat{X} = f(\hat{X}_z, \hat{X}') \tag{14}$$

3 计算机仿真

将 1 000 个传感器节点随机部署在 100 m × 100 m 的区域内,将 WSN 均匀划分为 10 个簇。假定在 WSN 中不存在攻击时,传感器节点的感知数据服从 $N(0, \delta^2)$,显著性水平 $\alpha = 0.05$。在下面的仿真实验中,聚合函数为求均值,攻击方式为

常量攻击,计算机仿真中均进行 200 次的蒙特卡洛实验。仿真图中横坐标为受攻击节点所占比例,纵坐标为聚合结果与真值之间的绝对偏差。

3.1　新算法与 RANBAR 的性能比较

图 2 为 $\delta^2=5$,$d=10$ 时新方法与 RANBAR 的性能对比,此时 $a=0$,新方法采用重心距离做测度。由图 2 可知,新算法的性能要好于 RANBAR,并且受攻击节点的分布越集中,新方法复原汇聚的绝对偏差越小。RANBAR 通过随机取样与样本数据进行分布函数的一致性检验,以此来构建期望模型,并剔除部分外围数据,增加了数据汇聚的精度。但 RANBAR 仅利用了很少的一部分样本数据,而新算法则合理利用了所有节点的感知数据,故新算法的复原汇聚性能较好。

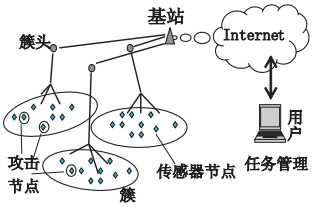

图1　分布式数据汇聚模型

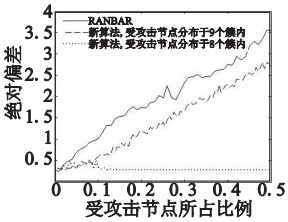

图2　新方法与RANBAR的性能比较

3.2　新算法与 IUMRDA 和 GPDRDA 的比较

3.2.1　复原汇聚性能比较

图 3 和 4 分别为 $d=0.5$ 和 $d=10$ 时新方法与 IUMRDA 和 GPDRDA 的复原汇聚效果对比。仿真环境为受攻击节点分布在 WSN 的 8 个簇中,$\delta^2=5$。$d=0.5$ 时 $a=1$,$d=10$ 时 $a=0$。

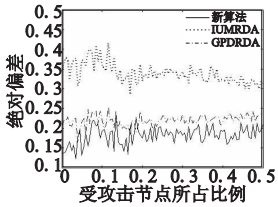

图3　d=0.5时新方法与IUMRDA和GPDRDA的比较

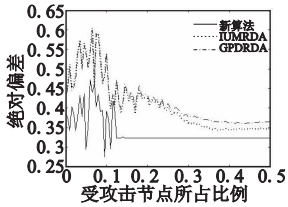

图4　d=10时新方法与IUMRDA和GPDRDA的比较

由图 3 可知,新方法的性能要好于 IUMRDA。这是因为 IUMRDA 采用信息效用测度来确定加权系数,信息效用测度实际上是各簇均值与期望模型均值差的平方。当攻击增量较小时,信息效用测度无法准确体现各簇数据的受攻击程度。新算法与 GPDRDA 相比稍有改进,这是因为 GPDRDA 此时采用灰色关联度做测度,利用了各个节点与期望模型各节点数据的差异,能够在一定程度上改善 IUMRDA 的不足;而新方法通过相关系数来衡量各簇数据与期望模型的相似度,在攻击增量较小时,各节点感知数据的微小变动都能体现在相关系数的变化上。

由图 4 可知,新算法的复原汇聚效果要好于 IUMRDA 和 GPDRDA。这是因为此时新算法采用重心距离做测度来确定各簇数据的加权系数,而 IUMRDA 和 GPDRDA 此时采用重心距离的平方来衡量各簇数据间的差异,这相当于增加了受攻击数据在复原汇聚时的比重,故新算法的性能较好。

3.2.2　抗噪声性能比较

当 $d=0.01$ 时,图 5、6 分别是信噪比为 0 dB 和 −5 dB 时新算法与 IUMRDA、GPDRDA 的抗噪声性能对比曲线。仿真环境为各传感器节点感知数据服从 $N(0,1)$,对每个节点感知信号施加一个高斯白噪声信号 $N(0,\delta_{噪}^2)$。

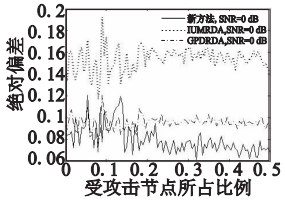

图5　SNR=0 dB时新方法与IUMRDA和GPDRDA的抗噪声性能

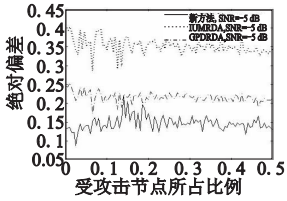

图6　SNR=−5 dB时新方法与IUMRDA和GPDRDA的抗噪声性能

由图 5、6 可知,新算法在信噪比为 0 dB 和 −5 dB 时的复原汇聚误差比 IUMRDA 和 GPDRDA 都要小,这意味着新算法对噪声干扰的稳健性较强。这是因为新算法采用相关系数做测度来衡量各簇数据与期望模型数据的相似度,而相关系数对噪声干扰的敏感度更高,故新算法利用相关系数对各簇数据进行加权聚合,以此来降低噪声干扰的影响。

4　结束语

本文提出一种基于相似度的复原汇聚算法,以重心距离和相关系数做测度来衡量各簇节点的受攻击程度,并引入预判机制来判断攻击增量的大小,根据判断结果采用重心距离或相关系数来确定各簇的加权系数,对各簇节点的感知数据进行加权聚合。新算法的复原汇聚精度较高,且对网络噪声干扰的稳健性较强,整体性能优于现有的复原汇聚算法。

参考文献:

[1] WAGER D. Resilient aggregation in sensor networks[C]//Proc of the 2nd ACM Workshop on Security in Ad hoc and Sensor Networks. New York: ACM Press,2004:78-87.

[2] BUTTYAN L, SCHAFFER P, VAJDA I. Resilient aggregation with attack detection in sensor networks[C]//Proc of the 4th Annual IEEE International Conference on Pervasive Computing and Communications. Washington DC: IEEE Computer Society, 2006: 332-336.

[3] REN Shu-qin, PARK J S. Density mining based resilient data aggregation for wireless sensor networks[C]//Proc of the 4th International Conference on Networked Computing and Advanced Information Management. [S.l.]: IEEE Computer Society, 2008: 261-266.

[4] BUTTYAN L, SCHAFFER P, VAJDA I. RANBAR: RANSAC-based resilient aggregation in sensor networks[C]//Proc of the 4th ACM Workshop on Security in Ad hoc and Sensor Network. New York: ACM Press, 2006: 83-90.

[5] LUO Yong-jian, YANG Xin, ZHANG Xu. An effective resilient data aggregation algorithm in wireless sensor networks[C]//Proc of International Conference on Wireless Communication, Networking and Mobile Computing. [S.l.]: IEEE Press,2007: 2642-2645.

[6] 罗永健,丁小勇,罗相根,等. 一种有效的无线传感器网络数据复原汇聚方法[J]. 数据采集与处理, 2011,26(1): 90-94.

[7] 陈正宇, 杨庚, 陈蕾, 等. 无线传感器网络数据融合研究综述[J]. 计算机应用研究,2011,28(5):1601-1604,1613.

[8] ROBERTO D P, PIETRO M, REFIK M. Confidentiality and integrity for data aggregation in WSN using peer monitoring[J]. Security and Communication Networks,2009,2(2):181-194.

[9] 杨鑫. 无线传感器网络中的数据复原汇聚方法[D]. 西安:西安通信学院, 2008.

[10] 项新建. 基于模糊数学与统计理论集成的多传感器数据融合方法[J]. 传感技术学报,2004,17(2):197-199.