WSN 中一种基于扩展卡尔曼滤波器的虚假数据注入检测算法

熊 伟
(重庆电子工程职业学院 计算机学院, 重庆 401331)

摘 要: 为了有效地检测传感器网络中被注入的虚假数据,提出一种基于扩展卡尔曼滤波器(EKF)的虚假数据注入检测算法。首先通过监控邻近节点行为,使用 EKF 预测邻近节点未来状态;然后给出了使用不同的融合函数(平均、求和、最大、最小)时理论阈值的确定方法;最后为了克服本地检测机制的缺陷,将本地检测方法与系统监控模块有效配合,从而准确地区分出恶性事件和紧急事件。仿真实验结果表明,无论是在合成数据还是实时数据下进行测试,该算法都能为无线传感器网络进行安全的数据融合提供有效的入侵检测功能。

关键词: 无线传感器网络; 数据融合; 虚假数据; 扩展卡尔曼滤波器; 融合函数; 阈值
中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2014)05-1475-06
doi:10.3969/j.issn.1001-3695.2014.05.046

Detection algorithm of false injected data based on extended Kalman filter in wireless sensor networks

XIONG Wei
(*School of Computer*, *Chongqing College of Electronic Engineering*, *Chongqing* 401331, *China*)

Abstract: In order to effectively detect the false injected data, this paper proposed a detection algorithm of false injected data based on an extended Kalman filter. Firstly, by monitoring the behaviors of its neighbors, it used EKF to predict their future states. Secondly, using different aggregation functions (average, sum, max, and min), it presented how to obtain a theoretical threshold. Finally, to overcome the limitations of local detection mechanisms, it illustrated how the proposed local detection approaches worked together with the system monitoring module to differentiate between malicious events and emergency events. Simulation results show that this proposed algorithm is suitable to provide intrusion detection capabilities for secure in-network aggregation in wireless sensor networks, whether the test is processed on the synthetic data or the real-time data.

Key words: wireless sensor networks; data aggregation; false data; extended Kalman filter; aggregation function; threshold

0 引言

无线传感器网络可为健康监控、科学数据收集、环境监控、军事行动等多个领域提供有效而又廉价的解决方案。然而,这些应用场合的传感器节点很容易发生故障,进而向网络注入随机虚假数据。人们已经证明,数据融合是降低无线传感器网络通信开销、节约能源的重要方法,并提出了多种数据融合协议,对协议性能进行了评估[1,2]。然而,只有很少一部分协议考虑了基于预防机制的网络数据安全融合,这些预防机制往往使用了加密、认证和密钥管理技术,一旦某节点被入侵成功,它的所有相关机密都将被攻击者窃取,使基于预防的技术失效。为解决这一问题,人们使用入侵检测系统(IDS)作为第二层防护墙,以有效发现恶性事件。从入侵检测角度解决网络数据安全融合问题的研究却很少。

易叶青等人[3]提出了一种基于协作水印的数据认证算法来识别虚假数据和重复包,算法在每个数据包中嵌入两类水印用于对数据内容进行认证。仿真结果表明,算法在数据包中嵌入水印后,能够对恶意窜改数据具有较高的敏感性,相比基于 MAC 的虚假数据过滤算法具有更低的通信开销和更高的识别与过滤虚假数据的能力。刘志雄等人[4]提出了一种基于地理位置的虚假数据过滤方案 GFFS,它通过对数据分组中包含的 MAC 和地理位置的正确性和合法性进行验证,可以有效地过滤不同地理区域的多个妥协节点协同伪造的虚假数据。冯艳芬等人[5]提出了一种基于分簇的低能耗隐私保护协议(LCCP-DA),它可以在实现数据融合的同时保护数据安全,防止数据被窃听和窜改,保障了数据的隐私性。杨庚等人[6]提出了一种低能耗无线传感器网络数据融合隐私保护算法 ESPART,该算法依靠数据融合树型结构本身的特性和分配的随机时间片,可以在有效保护数据隐私的前提下,花费较少的数据通信量得到精确的数据融合结果。

本文认为,为了提高无线传感器网络安全,应该对系统监控模块和入侵检测模块进行集成。实践中,人们经常部署无线传感器网络来监控森林火灾、战场态势等重要紧急事件。模块的集成有助于恶性事件和重要紧急事件的区分。例如,使用入侵检测模块后,当节点 A 因为事件 E 向节点 B 提出警报时,A 可以在系统监控模块的帮助下对 E 展开进一步调查。具体而言,A 可以唤醒 B 周围的有关传感器节点,向它们询问关于 E 的意见,如果大部分节点认为事件 E 可能发生,A 可以就 E 作出决策,认为 E 是某种紧

收稿日期: 2013-07-09; **修回日期:** 2013-09-06
作者简介: 熊伟(1980-),女,重庆万州人,讲师,硕士,主要研究方向为网络组建、网络安全、网站架构、信息检索(17530356@qq.com).

急事件触发的;否则,A 怀疑 E 是恶性事件。

由于数据传输的报文丢失率高、节点生存环境恶劣、传感器感知的不确定性以及母节点和子节点时间不同步等等,如何有效地检测网络中被注入的虚假数据较为困难。本文通过使用状态空间模型[7],提出基于扩展卡尔曼滤波器的机制来高效地处理这一问题,并在代表性节点 MICA2 Mote 模块[8]上实现了 EKF 滤波器算法。

1 系统模型

1.1 网络模型

为了进行数据融合,往往需要首先建立融合树。图 1(a)给出了数据融合树示例。在该图中,A,B,C D 执行传感任务,获得数据,然后把数据传给母节点 H;H 对接收到的数据进行融合(最小、最大、求和、平均等),再把融合后的数据发送给节点 K。处理过程同样适合如下操作:(E,F,G)-I-J,(M,N)-L-J。这些融合操作基于已经确定的母—子关系,该关系可用图 1(b)建模。在图 1(a)中,基站收集所有数据,必要时通过互联网传输数据。

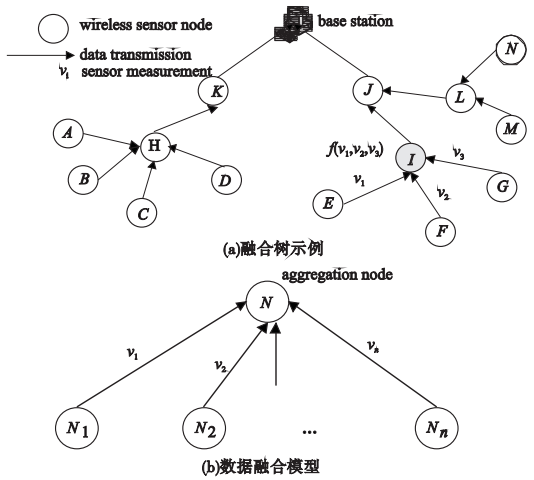


图1 无线传感器网络的数据融合

1.2 相关假设

为了便于后续的阐述,在此给出本文用到的相关假设:a)假设本文算法可以承受母节点和子节点间的时间异步;b)假设节点支持混杂模式,通过支持混杂模式,当某节点如 F,在另一节点(如 I)的无线传输范围内,则 F 可以监听 I 的传输内容,这有利于实现本文的相邻节点监控机制;c)为了节约节点能量,假设节点可以休眠,且任意节点可在任意时刻被瞬间唤醒。

1.3 安全模型

如果节点被敌人入侵,该敌人可取得节点的完全控制权,敌人可以向网络注入虚假数据。本文假设,遭到入侵的节点传输的数据与实际值(比如实际监控的温度均值)有显著区别,进而严重干扰数据融合操作。如果敌人只注入少量虚假数据,且与实际融合数据差别较小。这对部署的应用产生的影响很小。因此,也将考虑敌人持续生成小偏差虚假数据的攻击模型,假设异常事件周围的大部分节点没有被入侵。

2 网络数据的安全融合

2.1 总体思路

本文提出的方案主要包含 IDM 和 SMM 两个模块。IDM 模块的功能是检测被监控节点是否是恶性内部节点;SMM 模块的功能是监控重要紧急事件。其中,SMM 是大部分无线传感器网络的必要组件,IDM 和 SMM 模块只有相互集成才能有效工作。本文不希望只依赖于本地检测,因为每个节点可用信息量非常有限。此外,节点有可能失效,导致完好节点和恶性节点发送的紧急事件难以区分。对于本文提出的算法,每当 IDM 和 SMM 模块检测到异常事件时,就会要求事件周围的更多节点展开合作,作出最终决策。

对 IDM 模块,本文总体思路是:节点 A 以混杂模式窃听相邻节点传输的融合数据,并将其与正常值预测范围作比较。如果监听值不在正常范围内,则要么事件 E 发生,要么相邻节点 N 被入侵。为了确定节点 N 是恶性节点还是爆发了森林火灾等重要紧急事件,节点 A 将唤醒 N 周围的相关节点,询问它们关于 E 的意见,进而让 IDM 和 SMM 展开合作。

当预测网络数据融合正常值范围时,主要面临如下的挑战:a)潜在的不确定性来源过多,难以获得实际融合数据;b)无线传感器网络的报文丢失率较高。例如,在文献[9]中的室内环境,总共部署了 62 个传感器节点,负载为每秒 0.5 个报文,然而大约有 35% 的链路,其 MAC 层的报文丢失率仍然达到了 50% 甚至更高。因此,即使进行合适的链路层丢失恢复,也难以应对高报文丢失率。

考虑到传感器节点的资源有限,这使得无法使用功能强大却成本昂贵的估计和预测算法。为了实现相邻节点监控机制,需要能够被节点高效执行的简单算法。鉴于此,本文提出一种基于 EKF 的检测算法,让每个节点预测和估计相邻节点的未来值,具体内容将在下节详细讨论。

2.2 基于本地检测的扩展卡尔曼滤波器

文献[7]提出的基于状态空间模型的卡尔曼滤波器主要解决如下通用问题:利用与系统状态成线性关系且遭到加性高斯白噪声干扰的测量值,估计遭到高斯白噪声干扰的动态系统的状态。用线性估计理论扩展后,EKF 可以线性模拟小干扰影响,进而用于多种非线性场合。通过为具体的无线传感器网络建立合适的过程模型和测量模型,利用时间更新和测量更新方程递归处理数据,可以使用 EKF 获得状态的准确估计值。

在本文中,状态表示将要测量的实际值。某一给定时刻的状态用感兴趣属性的瞬时值来刻画,如无线传感器网络监控的实际温度。在以下内容中,状态与实际值可以互换。由于实际温度受到无线传感器网络各种不确定性因素的影响,融合节点获得的数值很可能不是真实状态数值。融合节点只能获得测量数量,进而估计实际数值。因此,当在不同应用场合下将 EKF 应用于无线传感器网络时,必须要有合适的模型。在本文研究中,采用离散时间 EKF 进行本地检测,它的系统状态在离散时间集合 t_k 内被估计得到,$k=0,1,\cdots$;这些离散时间对应于数值测量和状态估计时间。具体的检测过程如下所示。

2.2.1 系统动态模型

表 1 列出了文中将要使用到的相关符号含义。实际融合数值构成一个动态过程,式(1)给出的过程模型决定该过程的演变。

$$x_{k+1}=F(x_k)+w_k \tag{1}$$

在式(1)中,函数 F 将时间 t_k 时的状态与 t_{k+1} 时的状态联系在一起。$F(x_k)$ 与实际应用环境有关,对给定的应用场合,可能需要不同形式的 $F(x_k)$ 来刻画状态变化,鉴于被监测现象的复杂性,$F(x_k)$ 可能会非常复杂。因此,对部署应用作详细

的分析后,需要对 $F(x_k)$进行简化。

表 1　EKF 滤波器的符号含义

标记	含义	标记	含义
x_k	状态:时间 t_k 时的实际数值	Q_k	时间 t_k 时的 w_k 的方差
F	将 x_{k+1}与 x_k 联系在一起的函数	v_k	时间 t_k 时的测量噪声
$\hat{x}_k^-$	x_k 的先验估计	R_k	时间 t_k 时的 v_k 的方差
$\hat{x}_k^+$	x_k 的后验估计	P_k^-	时间 t_k 时的先验估计误差
z_k	时间 t_k 时的测量值	P_k^+	时间 t_k 时的后验估计误差
H	将 x_k 与 z_k 联系在一起的函数	K_k	时间 t_k 时的卡尔曼增益
w_k	时间 t_k 时的过程噪声		

在式(1)中,w_k 是时间 t_k 时的过程噪声,且经常建模为正态随机分布变量。进一步假设,w_k 服从正态分布 $N(0,Q)$,其中 Q 是常值参数,表示 w_k 的方差。请注意,Q 也可以经基站在网络上广播,以适应环境变化。测量模型用下式给出:

$$z_k = H(x_k) + v_k = x_k + v_k \tag{2}$$

其中:z_k 是时刻 t_k 时的测量值,例如在图 1(a)中,节点 I 在时刻 t_k 向外传输融合数值 z_k,节点 E、F、G 可以窃听到这一数值;$x_k \in \mathrm{R}$ 是时刻 t_k 时被监控的状态,且表示融合节点 I 覆盖区域的实际融合值;v_k 是测量噪声,表示噪声条件下的传感器测量值及无线传感器网络下的各种不确定性因素。同样,对具体应用领域,假设 v_k 服从均值为 0、方差为 R 的正态分布 $N(0,R)$,其中 R 是 v_k 的方差(注意,R 也可通过基站进行调整)。

2.2.2　系统方程

下面列出基于系统模型(式(1)和(2))的重要系统方程。使用时间更新—状态估计等式来预测时间 t_{k+1}时的状态

$$\hat{x}_{k+1}^- = F(\hat{x}_k^+) \tag{3}$$

使用时间更新—误差更新方程来预测时间 t_{k+1}时的误差:

$$P_{k+1}^- = \left.\frac{\partial F}{\partial x}\right|_{x=\hat{x}_k^+} P_k^+ \left.\frac{\partial F}{\partial x}\right|_{\hat{x}_k^+}^T + Q_k = P_k^+ + Q_k \tag{4}$$

使用测量更新—卡尔曼增益方程来计算时间 t_{k+1}时的卡尔曼增益:

$$K_{k+1} = P_{k+1}^-(P_{k+1}^- + R_k)^{-1} = \frac{P_{k+1}^-}{P_{k+1}^- + R_k} \tag{5}$$

使用带有测量值 z_{k+1}的测量更新—估计更新方程,利用 z_{k+1}对估计进行更新:

$$\hat{x}_{k+1}^+ = \hat{x}_{k+1}^- + K_{k+1}(z_{k+1} - \hat{x}_{k+1}^-) \tag{6}$$

使用测量更新—误差协方差更新方程来更新估计误差:

$$P_{k+1}^+ = (I - K_{k+1}H_{k+1})P_{k+1}^{-1} = (1 - K_{k+1})P_{k+1}^{-1} \tag{7}$$

时间更新方程式(3)和(4)负责预测时间 t_{k+1}时的状态($\hat{x}_{k+1}^-$)和估计误差(P_{k+1}^{-1})。在式(4)中,对 $F(x)$进行一阶泰勒级数近似,$\frac{\partial F}{\partial x}|_{x=\hat{x}_k^+}$是 $x=\hat{x}_k^+$ 时 F 关于 x 的一阶偏导数的值。由于本文中的状态是标量变量,因此 $P_{k+1}^- = P_k^+ + Q_k$。

测量更新方程式(5)和(6)负责将新的测量值 z_{k+1}融入到先验估计值 $\hat{x}_{k+1}^-$中,然后计算出后验估计值 $\hat{x}_{k+1}^+$。由于 $H(x_k)=x_k$,$H_k=\frac{\partial H}{\partial x}=1$,因此,式(6)中的 H_k 可被忽略;式(7)可用于更新估计误差。

一般地,在时间 t_k,为了预测实际值 x_{k+1},一个节点需要两个数值:a)先验估计值 $\hat{x}_{k+1}^-$,这可根据式(3)得出;b)随机窃听到的测量值 z_{k+1}。然后,根据式(6)可以计算出一个相对准确的估计值 x_{k+1}。式(6)的第二部分表示的是根据先验估计 $\hat{x}_{k+1}^-$和测量值 z_{k+1}之差而对 $\hat{x}_{k+1}^-$进行的调整量。

EKF 可以为相邻节点的未来融合值提供一个相对准确的预测值。为了阐述这一点,使用图 1(b)的网络拓扑结构,融合函数选为平均函数,然后利用 2.2.2 节中的系统方程绘出实际值、测量值和估计值曲线。具体地,模拟了实际温度缓慢上升但遭受高斯白噪声影响的环境。图 1(b)中的传感器节点 N_i 测量温度(为了模拟测量误差,测量温度为受到随机噪声干扰的实际温度),并定期将测量数据发送给它的母节点 N;节点 N 定期计算接收数据(即测量数据)的均值。估计值是每个节点 N_i 测量值基于 EKF 的估计值。在图 1(b)中,假设从 N_i 到 N 的报文丢失率为 0.3。结果显示于图 2 中。从图 2 可以看出,虽然存在多种不确定性,基于 EKF 的算法仍然可以为实际值提供一个相对准确的估计值。

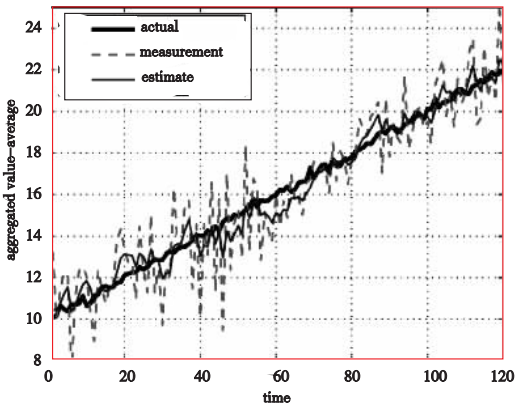
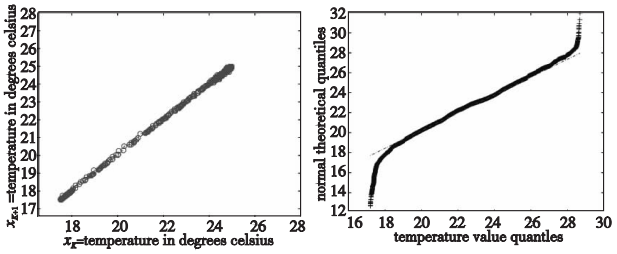


图2　离散时间EKF滤波器的预测准确度

2.2.3　基于实际数据集的模型拟合

式(1)中的 F 视应用场合而定。例如,可以使用常用的英特尔实验室数据[10]来绘制 x_k 和 x_{k+1}间的关系 F,绘制环境与英特尔伯克利研究实验室类似。

随机选择一个传感器节点,过滤掉错误读数(即与上一读数和后续读数存在明显差异的读数),然后选择一个温度读数不断上升的时间周期。根据该时间周期内的读数,绘制出 x_k 和 x_{k+1}间的关系,如图 3(a)所示。从图 3(a)可以看出,F 取线性函数形式 $x_{k+1}=x_k+w_k$ 是合理的。为了阐述 w_k 是否服从固定方差为 Q 的正态分布,绘制出正态分布与($x_{k+1}-x_k$)间的分位数—分位数图形(Q-Q 绘图),如图 3(b)所示。从图 3(b)可以看出,大多数数据点严格沿直线分布,充分表明 w_k 服从正态分布。



(a) x_k和x_{k+1}间的关系　　(b) Q-Q图

图3　英特尔实验室数据测试结果

使用最大似然估计(MLE)[11]来估计$\sqrt{Q}$。使用$N(\mu,\sigma^2)>X$来表示均值为 μ、方差为 σ^2 的正态分布随机过程 X,有 $N(0,Q)\equiv N(0,\sigma^2)>w_k=x_{k+1}-x_k$,似然函数变为

$$\prod_{k=1}^{n} p(w_k|0,\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n} w_k^2\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_{k+1}-x_k)^2\right)$$

其中:n 是拥有的读数数量。因此,Q 的最大似然估计 $\hat{Q}=$

$\frac{1}{n-1}\sum_{k=1}^{n-1}(x_{k+1}-x_k)^2$。将该式用于选择的英特尔实验室数据上,可以估计出 w_k 服从 $N(0,0.2809)$分布。

2.2.4 基于阈值的异常检测机制

本节给出了基于 EKF 的本地检测算法。传感器节点监测其相邻节点的行为,确定相邻节点将来融合值的正常范围,所生成的正常范围以通过 EKF 得出的估计值为中心,如果监测值不在预测的正常范围内,则可以发出警报。这一方法阐述于算法 1 中,其中 Δ 表示事先定义好的阈值。

在算法 1 中,节点 A 的作用是决定 z_{k+1}是否正常。节点 A 可以窃听节点 B 在 t_{k+1}时传输的内容 z_{k+1},在得出时间 t_k 时的估计值 $\hat{x}_k^+$ 后,A 可以根据式(3)预测出时间 t_{k+1}时的节点 B 的传输值 $\hat{x}_{k+1}^-$;在时间 t_{k+1},A 窃听到 B 的传输值 z_{k+1},将其与 $\hat{x}_{k+1}^-$作比较,以决定 B 是否正常。如果 $\hat{x}_{k+1}^-$与 z_{k+1}间的差别(在算法 1 中表示为 Diff)大于预先定义好的阈值 Δ,则 A 可以就 B 提出警报;否则,A 认为 B 运转正常。

算法 1 基于 EKF 的本地检测算法

假设:节点 A 可以窃听到节点 B 的传输内容,A 认为在 t_k 时及 t_k 前 B 是正常节点。

输入:节点 B 传输且被 A 窃听的 z_{k+1}。

输出:A 是否就 z_{k+1}提出警报。

在时间 t_k,A 根据式(6)计算 $\hat{x}_k^+$(请注意,$\hat{x}_k^-$ 存储于节点 A);

A 利用式(3)根据 $\hat{x}_k^+$ 计算 $\hat{x}_{k+1}^-$;

A 计算 Diff = $|\hat{x}_{k+1}^- - z_{k+1}|$;

if (Δ < Diff) then

  A 就 B 提出警报

else

  A 认为 B 状态正常

end if

实际上,基于异常的 IDS 的虚警率较高,可以使用后处理机制来降低潜在的虚警率。例如,可以更新算法 1 的第 4 行,当 Δ 的连续几个观测值均小于 Diff 时,A 再就 B 提出警报。这主要是因为入侵行为往往表现出一定的局部性,即较短时间窗口内的多个警报。于是,可以使用多个警报来生成一个入侵报告,进而有效降低虚警率。

3 阈值分析

在无线传感器网络数据融合协议中的多种因素(比如报文丢失、报文冲突和时间异步等),都会造成数据融合值存在不确定性。设 U 表示这种不确定性的方差,根据文献[12]所述,Δ 可以设为 $3\sqrt{U}$。下面给出 U 的分析。

图 1(b)中,每个节点 N_i 表示一个传感器节点,每个节点 N_i 根据事先定义好的数据融合协议发射数据 v_i 给其母节点 N。假设每个数值 v_i 的期望为 $E[v_i]=\mu_i$,每个数值 v_i 的方差为 $\mathrm{var}(v_i)=\sigma_i^2$;还假设每条链路的报文丢失率为 $0<p<1$(p 为 N 因报文丢失、报文冲突等原因,没有从其子节点处接收到报文的概率);用随机变量 X 表示节点 N 处的数据融合值。在不同报文丢失概率下,分析随机变量 X 的方差。

a)平均值。设 E_m 表示 $m(m\geqslant 0\wedge m<n)$个报文发生丢失的事件,$V_m$ 表示这 m 个报文丢失后对应的数据融合数值,$p(E_m)$表示事件 E_m 的概率,且 $\sum p(E_m)V_m=\frac{(n-1)!}{m!\ (n-m)!}\times(1-p)^{n-m}p^m\sum_{i=1}^{n}v_i$。然后,关于 X 的概率函数 P_X 为

$$P_X=\begin{cases}(1-p)^n & \text{如果 } X=\dfrac{\sum_{i=1}^{n}v_i}{n}\\ (1-p)^{n-1}p & \text{如果}\dfrac{\sum_{i=1\wedge i\neq 1}^{n}v_i}{n-1}\\ (1-p)^{n-1}p & \text{如果}\dfrac{\sum_{i=1\wedge i\neq 2}^{n}v_i}{n-1}\\ \cdots & \cdots\end{cases}$$

为了节约篇幅,省略了推导过程。对于执行过数据融合操作,被监测数值 v_i 互相之间比较类似的应用场合(比如,一个地区的监测温度并没有很大差别),有 $E(v_i)=\mu$ 且 $\mathrm{var}(v_i)=\sigma^2$。因此,考虑报文丢失、报文冲突等影响,有

$$E(X)=\sum_{m=0}^{n}(1-p)^{n-m}p^m\frac{n!}{m!\ (n-m)!}\mu$$

$$E(X^2)=\sum_{m=0}^{n}\left[\frac{(1-p)^{n-m}p^m}{(n-m)^2}\left(\begin{array}{l}(C_n^m-C_{n-1}^{m-1})n(\mu^2+\sigma^2)+\\ 2(C_n^m-2C_{n-1}^{m-1}+C_{n-2}^{m-2})\dfrac{n(n-1)\mu^2}{2}\end{array}\right)\right]$$

其中:$C_n^m=\frac{n!}{m!\ (n-m)!}$。鉴于篇幅所限,省略推导过程。这里有 $\mathrm{var}(X)=E[X^2]-E^2[X]$,$U$ 可设为 $\sqrt{\mathrm{var}(X)}$。

b)求和。对于融合求和,当出现 $m(m\geqslant 0\wedge m\leqslant n)$个报文丢失时,有

$$\sum p(E_m)V_m=\frac{(n-1)!}{m!\ (n-m-1)!}(1-p)^{n-m}p^m\sum_{i=1}^{n}v_i$$

于是

$$P_X=\begin{cases}(1-p)^n & \text{如果 } X=\sum_{i=1}^{n}v_i\\ (1-p)^{n-1}p & \text{如果 } X=\sum_{i=1\wedge i\neq 1}^{n}v_i\\ (1-p)^{n-1}p & \text{如果 } X=\sum_{i=1\wedge i\neq 2}^{n}v_i\\ \cdots & \cdots\end{cases}$$

对于被测数值 v_i 比较类似的应用场合:

$$E(X)=\sum_{m=0}^{n}(1-p)^{n-m}p^m\frac{n!}{m!\ (n-m-1)!}\mu$$

$$E(X^2)=\sum_{m=0}^{n}\left[(1-p)^{n-m}p^m\left(\begin{array}{l}(C_n^m-C_{n-1}^{m-1})n(\mu^2+\sigma^2)+\\ 2(C_n^m-2C_{n-1}^{m-1}+C_{n-2}^{m-2})\dfrac{n(n-1)\mu^2}{2}\end{array}\right)\right]$$

然后可以计算出 U。

c)最大/最小。最大值数据融合分析与最小值类似,为了节约篇幅,只提供最小值分析。对于数据集 $v_i(1\leqslant i\leqslant n)$,不失一般性,对 v_i 排序后有 $v_1\leqslant v_2\leqslant\cdots\leqslant v_n$。此时,$X$ 可作如下定义,同时 P_X 为

$$P_X=\begin{cases}(1-p) & \text{如果 } X=v_1\\ p(1-p) & \text{如果 } X=v_2\\ p^2(1-p) & \text{如果 } X=v_3\\ \vdots & \vdots\end{cases}$$

对于被测数值 v_i 比较类似的应用场合,假设 v_i 的概率密度函数为 $f(x)$,累积分布函数为 $F(x)$。根据顺序统计学理论,可以计算出 X 的概率密度函数:

$$f_X=\frac{n!}{(r-1)!\ (n-r)!}[F(x)]^{r-1}[1-F(x)]^{n-r}f(x)=$$
$$n[1-F(x)]^{n-1}f(x)$$

其中:r 表示第 r 个顺序统计量,对于最小值数据融合,$r=1$。根据式(8),可以进一步计算出 $E(X)$和 $E(X^2)$,进而求出 U。

4　IDM 和 SMM 模块集成

人们经常部署无线传感器网络来监测紧急事件(如森林火灾爆发),有效节点可以触发重要事件,生成异常而又重要的信息;同时,传感器节点容易损坏,即使是正常的节点也会发生故障而生成异常信息。因此,仅仅依靠本地检测,存在虚警率高这一问题。传感器网络需要进行节点协作来提高异常事件的决策准确度。

本文将 IDM 和 SMM 模块进行集成,以克服本地检测机制的缺陷,从而让传感器节点更加高效地工作。当节点 A 由于事件 E 对节点 B 发出警报时,为了确定 E 是恶性事件还是紧急事件,A 可以与当前 SMM 模块展开合作,对 E 进行进一步调查。无线传感器网络部署密度较高,就是为了对一些事件展开协作式监测。为了节约能量,定期调度部分节点使其进入睡眠状态。以此为基础,节点 A 可以唤醒 B 周围的这些节点(在图 4 表示为联合检测节点),询问它们关于事件 E 行为的意见。由于 E 周围的大部分节点没有发生故障,A 收集完这些节点的信息后,如果 A 觉得大部分节点认为事件 E 可能发生,则 A 作出决策,认为 E 由一些紧急事件触发;另一方面,如果 A 发现大部分节点认为 E 不应该发生,则 A 认为 E 是被恶性节点或发生故障的正常节点触发。通过这种方法,A 可以继续唤醒事件 E 周围节点,询问它们关于 E 的意见。如果 A 仍然发现大部分节点觉得 E 不该发生,则 A 怀疑 E 是恶性事件。A 作出最终决策后,把该事件报告给基站。无论该事件是紧急事件还是恶性事件,相关操作人员都会妥善处理。实际上,SMM 可以使用一些高效的方法来收集事件 E 周围节点的信息。例如,Wang 等人在文献[13]中提出了一种高效算法,它可以生成一个涵盖可疑节点(本例中为节点 B)的所有相邻节点的支配树,然后由发起节点(本例为 A)加以利用,收集可疑节点信息。

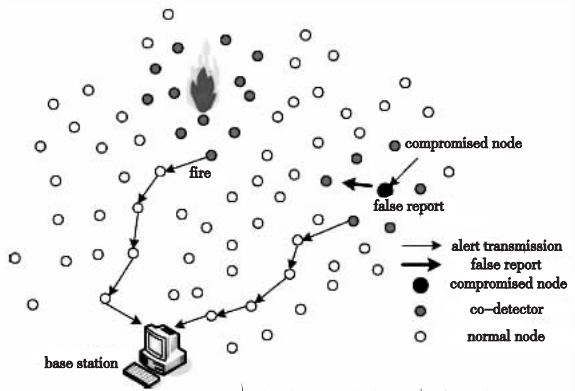


图4　IDM和SMM进行协作,将恶性事件和紧急事件区分开

5　实验评估

本章使用实时数据和合成数据来评估基于 EKF 的检测算法性能。实时数据的好处就是可以反映真实情况,然而,实时数据只涵盖有限的几种情况,且这些情况的参数不能改变;另外,获得真实攻击数据的难度较大。而合成数据在正常和异常情况下的参数都可以被仔细控制,因此具有一定的优势。

5.1　合成数据实验结果

5.1.1　实验配置

本文使用图 1(b)作为抽象的网络模型来评估基于 EKF 的异常检测算法的性能。对每条链路,使用不同的报文丢失率 P,分别为 0.1, 0.25 和 0.5。对每种报文丢失率,模拟两组传感器数值 v_i,v_i 表示子节点发送给母节点的初始数据。对于第一组传感器数值,设所有的 v_i 在预定义范围[min,max]内随机分布。选择[min,max]的目的是绘制出合适的虚警率和检测率。这是为了模拟具有相同属性(如温度)的无线传感器网络监测区域。对于第二组传感器数值,设置不同的 v_i 在不同的[min,max]区间随机分布,即 $\forall i, 1 \leqslant i \leqslant n$,[min,max]区间满足 $\min_i < v_i < \max_i$,$\min_{i+1} = \min_i + T$ 且 $\max_{i+1} = \max_i + T$。T 是常值参数,其目的是为了模拟具有不同属性的无线传感器网络监测区域。例如在实际应用中,由于受地理条件影响,不同的区域可能有不同的温度。

本文设置子节点的数量为 6。对于密集部署的传感器网络,这个数值是合理的。此外,发现子节点数量对性能结果的影响非常有限。本文使用正态分布来生成过程噪声和测量噪声,对每组参数,测量它在不同数据融合函数下的性能。

本文设置节点 N 为问题节点,即节点 N 可能向网络注入虚假融合数据。一个关键问题是,如何模拟攻击数据。直观来讲,攻击数据与正常数据的差别越大,则攻击检测的难度就越低,因此引入破坏程度这一概念,用 D 表示。D 定义为攻击数据与正常数据间的差异。例如,在图 1(b)中,如果本文假设节点 N 的正常融合值为 C,而 N 发出的恶性融合值为 M,这时有 $D = M - C$。基于不同的 D 值来评估本文本地检测算法的性能。

5.1.2　评价指标

使用下面两个指标来评估基于 EKF 的算法 1 的性能。

a)虚警率。针对正常数据测量该指标。假设测量了 m 个正常数据,其中有 n 个数据被确定为异常,则虚警率为 n/m。

b)检测率。针对异常数据测量该指标。假设测量了 m 个异常数据,其中有 n 个数据被检测出来,则检测率为 n/m。

对相同的模拟参数集合,获得了 5 000 个正常数据和 5 000 个恶性数据;对这些数据项,使用不同的阈值,然后测量相应的虚警率和检测率。换句话说,对于给定的阈值,使用算法 1 中的本地检测算法来获得一个虚警率和检测率;然后使用不同的阈值,获得一组虚警率和检测率。对同一融合函数,使用相同的一组阈值以便于比较。

5.1.3　基于 EKF 的检测算法的仿真结果

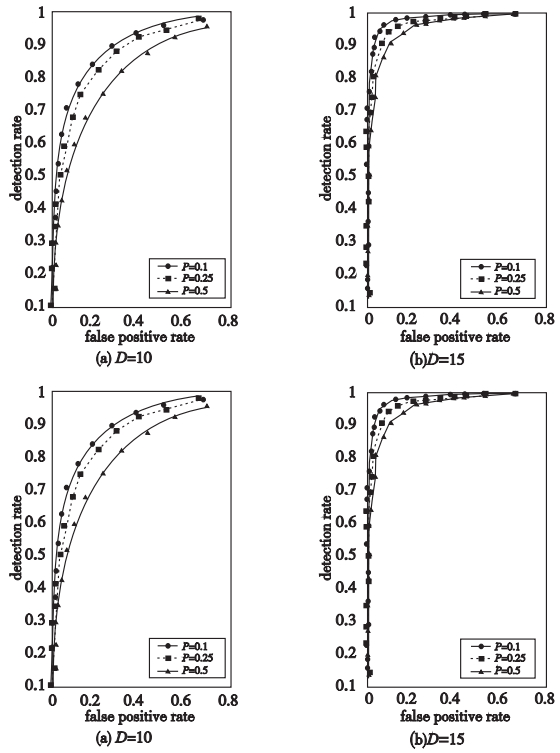
对基于 EKF 的检测算法进行性能评估,当 v_i 服从相同分布时,让所有的 v_i 在预定义的范围[min,max]内随机分布。当 v_i 分布不同时,设置 v_i 在不同的[min,max]内随机分布。由于平均、求和、最大和最小融合函数的仿真结果类似,此处只给出均值融合函数仿真结果。

1)平均值

从图 5(a)和(b)中可以看出,算法性能(虚警率和检测率)受到报文丢失率的影响很小。这是因为融合函数为均值函数,融合节点 N 计算出接收数值的均值,所有的 v_i 落入相同的范围,即使丢失几个 v_i 报文也不会对整体性能产生很大影响。比较图 5(a)和(b)可以看出,随着攻击数据的损失程度上升,算法性能将会上升(曲线向左上方发展)。这正是期望的情况。

从图 6(a)和(b)可以观察到不同报文丢失率的影响。当报文丢失率上升时,检测率将会下降,而虚警率基本保持不变。

这是因为,当报文丢失率上升时,数据较小的 v_i 的丢失率将会上升,导致时间 t_k 时的测量值 z_k 变大。根据式(6),$\hat{x}_k^+$ 变大;在本文仿真环境下,根据式(3)计算出来的 $\hat{x}_{k+1}^-$ 也会变大;根据算法 1,Diff 变小,导致检测率变小;类似地可以看出,图 6(b)的总体性能优于图 6(a)。



2)估计误差

在仿真时使用常值 Q 和 R,于是估计误差 P_k^+ 会很快稳定,因此 P_k^+ 的初始值并不重要。图 7(a)给出了具有随机初始值的 P_k^+ 运行情况。

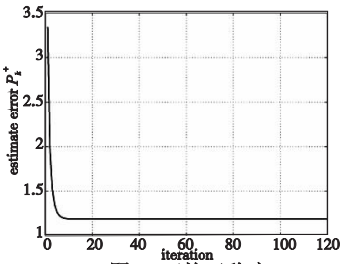


图7 P_k^+趋于稳定

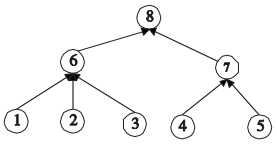


图8 测试床

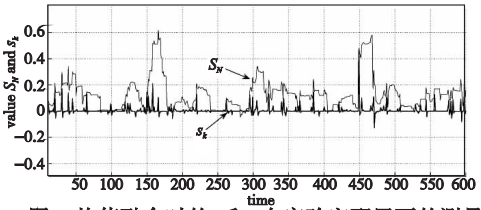


图9 均值融合时的s_N和s_k在实验室配置下的测量

5.2 实时数据实验结果

本文利用 MICA2 Mote 模块在 TinyOS 系统上实现 EKF,从 ROM 和 RAM 大小两方面评估其开销性能。对单个 MICA2 Mote 模块,EKF 需要 1 204 个 ROM 字节, 32 个 RAM 字节。使用 8 个 MICA2 传感器节点来建立一个双层测试床,如图 8 所示。在温度基本恒定的实验室中,叶节点(节点 1 ~5)以每分钟一个样本的速率采集温度,定期将采集的数值发送给母节点(节点 6 和 7)。母节点计算出接收数据的均值,然后把融合数据发往节点 8,由节点 8 对接收数据作进一步融合。

本文利用节点 8 采集到的数据,绘制出 s_N 和 s_k 的动态特性,如图 9 所示。观察结果与 5.1.3 节的平均融合结果类似;对其他融合函数,观察结果与 5.1.3 节相同。为了节约篇幅,本文没有给出其他函数的实验结果。

6 结束语

数据安全融合问题是目前无线传感网中的研究热点。在本文中,提出基于 EKF 的算法来检测虚假注入数据,阐述了如何使用 EKF 处理无线传感器网络的各种不确定性,并给出了一种有效的本地检测机制。同时,证明了本文提出的 IDM 可以与 SMM 配合,将恶性事件与紧急事件区分开来。利用基于 MICA2 Mote 模块的实时实验及大规模合成数据对本文算法进行性能评估。实验结果表明,本文算法可以为无线传感器网络进行安全的数据融合提供有效的入侵检测功能。

下一步将从本文算法的多方面着手提高其稳定性和效率,包括使基于 EKF 的本地检测算法考虑更多的参数因素,当大部分相邻节点出现问题时如何发现入侵者,以及对本文算法进行更加系统的评估。

参考文献:

[1] OZDEMIR S. Concealed data aggregation in heterogeneous sensor networks using privacy homomorphism[C]// Proc of IEEE International Conference on Pervasive Services. [S. l.]: IEEE Press, 2007: 165-168.

[2] CHEN J Y, PANDURANGAN G, XU Dong-yan. Robust computation of aggregates in wireless sensor networks: distributed randomized algorithms and analysis [J]. IEEE Trans on Parallel and Distributed Systems, 2006,17(9): 987-1000.

[3] 易叶青,林亚平,李小龙,等. WSN 中基于协作水印的虚假数据过滤算法[J]. 软件学报,2010,21(1):107-118.

[4] 刘志雄,王建新. 传感器网络中一种基于地理位置的虚假数据过滤方案[J]. 通信学报,2012,33(2):156-163.

[5] 冯艳芬,刘宴兵. 基于分簇的低能耗数据融合隐私保护协议[J]. 计算机应用研究,2013,30(3):885-888.

[6] 杨庚,王安琪,陈正宇,等. 一种低耗能的数据融合隐私保护算法[J]. 计算机学报,2011,34(5):792-800.

[7] GREWAL M S, ANDREWS A P. Kalman filtering: theory and practice using MATLAB[M]. Hoboken: Wiley, 2011.

[8] MICA2 Mote [EB/OL]. http://www. memsic. com.

[9] ZHAO J, GOVINDAN R. Understanding packet delivery performance in dense wireless sensor networks[C]//Proc of the 1st International Conference on Embedded Networked Sensor Systems. New York: ACM Press, 2003: 1-13.

[10] Intel lab data [EB/OL]. (2004-04-05). http://db. csail. mit. edu/labdata/labdata. html.

[11] DUDA R O, HART P E, STORK D G. Pattern classification [M]. Hoboken: Wiley, 2012.

[12] MONTGOMERY D C. Introduction to statistical quality control [M]. Hoboken: Wiley, 2007.

[13] WANG Gui-ling, ZHANG Wen-sheng, CAO Guo-hong, *et al*. On supporting distributed collaboration in sensor networks[C]//Proc of Military Communications Conference. Washington DC: IEEE Computer Socieyt, 2003: 752-757.