一种改进的WSN异常检测和定位算法研究

赖 锴[1] 王新兵[2]
(河南财经政法大学计算机与信息工程学院 郑州 450002)[1]
(上海交通大学电子信息与电气工程学院 上海 200240)[2]

摘 要 异常快速检测和定位对于保证无线传感器网络的有效运行具有重要作用。提出了一种改进的传感器网络异常检测和定位方法。该方法通过两个阶段的探查来收集端到端测量数据以实现异常检测和定位。第1阶段探查的目的是选择可以覆盖最大数量异常链路的探点,缩小可疑区域范围,供第2阶段探查,这一阶段的探点选择问题被建模为预算有限条件下的覆盖范围最大化问题,提出一种基于对偶线性规划的高效近似方法来求解此问题。第2阶段的目的是以最小的通信代价,定位出导致观察到的端到端异常现象的具体链路,并根据多环置信度传播算法(LBP)来预测诊断质量。在不同网络设置下展开实验,结果表明,相比于精确求解方法,提出的算法性能略有下降但运行速度更快。

关键词 无线传感器网络,异常检测,定位,测量数据,探点,线性规划
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.4.017

Research on Improved Anomaly Detection and Localization Algorithm in Wireless Sensor Networks

LAI Kai[1] WANG Xin-bing[2]
(Institute of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou 450002, China)[1]
(School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)[2]

Abstract Fast anomaly detection and localization are critical to ensure effective functioning of wireless sensor networks. This paper presented an improved anomaly detection and localization algorithm in wireless sensor networks, where network heterogeneity is exploited for better bandwidth and energy efficiency. End-to-end measurements are collected through a two-phase probing. The goal of the first phase probing is to select probes that can cover as many anomalous links as possible and narrow down suspicious areas to be examined in the second phase. The probe selection problem in this phase is formulated as a budgeted maximum coverage problem. We proposed an efficient approximation algorithm to solve it based on linear programming duality. The second phase probing is aimed at locating individual links that are responsible for the observed end-to-end anomalies with minimum communication cost. The prediction of diagnosis quality is carried out using the loopy belief propagation (LBP) algorithm. Experimental results show that our algorithm is much faster than the exact solution at the cost of slight performance degradation.

Keywords Wireless sensor networks, Anomaly detection, Localization, Measurements, Probe, Linear programming

1 引言

无线传感器网络在关键基础设施监控、科学数据采集、智能建筑和个人医疗系统等实际领域得到广泛应用[1]。由于环境可能无人监管或者存在不友好因素,传感器网络可能发生链路或结点丢失或被恶意入侵进而发生系统故障,因此,必须持续监测传感器网络的总体状态,以保证其正确高效运行。与互联网诊断协议相比,无线传感器网络由于通信带宽低,能量、内存、计算性能等资源有限,因此面临着新的问题。当前,无线传感器网络异常检测和定位研究[2]基本分为两类:集中式和分布式。在集中式方法中,中央控制器负责监测并跟踪网络异常。然而,资源有限的传感器网络难以支撑周期性集中式采集数据。分布式策略通过鼓励本地决策来解决这一问题,例如,结点可以通过与监视器等相邻结点的协议来检测出行为存在异常的结点。然而,在恶劣环境下,结点可能无法准确传输自身或相邻结点的状态信息,且中间结点可能会故意改变其转发信息。因此,本文将主要针对无线传感器网络的异常检测和定位问题进行研究。

2 相关工作

异常事件和定位问题在无线传感器网络中已经被广泛研究,众多研究者相继提出了一系列有代表性的方案[3-8],如姜旭宝等[3]提出了一种基于变宽直方图的异常数据检测算法,通过数据聚合的方式将网络中的动态感知数据聚合成变宽的直方图来准确检测出异常数据,同时避免不必要的数据传输。肖政宏等[4]提出了一种基于分布式机器学习的异常检测方

到稿日期:2014-05-20 返修日期:2014-07-29 本文受国家自然科学基金重点项目:无线网络资源分配与性能评价(61325012/F020809)资助。
赖 锴(1982—),男,硕士,主要研究方向为信息系统、网络安全;王新兵(1972—),男,博士,教授,博士生导师,主要研究方向为无线网络编码、传感器网络、认知无线电。

案。仿真实验结果表明，该方案与当前典型的无线传感器网络入侵检测方案相比具有较高的检测率和较低的误检率。朱翠涛等[5]提出了一种改进的下降迭代检测算法。该算法通过动态调节参数，改变迭代权值，加快了算法收敛速度。实验结果表明：在相同条件下，改进算法的成功检测概率比贝叶斯算法平均提高了13%。Yigal等[6]研究了事件检测中的覆盖空洞问题，为了保证事件检测的可靠性，设计了一个精确的可认证的分布式 k 覆盖方案来保证事件检测的准确性。仿真结果表明，该方案在不需要使用位置信息的情况下，仍能精确地检测出不同网络规模下的事件覆盖空洞。

另外，张玉琴等[7]提出了一种分布式的节点定位异常检测方法。该方法不需要任何已知的部署知识或额外的硬件，每个簇的簇头节点只需根据该簇节点报告的位置和邻居表信息进行过滤计算，更新权值，即可确定和撤销定位异常的节点。杨黎斌等[8]提出了一种基于核聚类的异常检测方案KCAD，以检测路由攻击所导致的流量异常。该方案通过利用Mercer核，将输入空间的流量特征样本隐式地映射到高维特征空间，突出了不同样本间的特征差异，从而更好地完成了聚类，提高了检测准确率，同时还针对流量特征样本做了时间维扩展，使之更能反映近期网络流量状况，减少了由于历史数据集影响所带来的误报。本文在现有工作的基础上，提出了一种改进的传感器网络异常检测和定位方法，并通过仿真实验验证了本文方法的有效性。

3 问题描述

本节首先描述网络模型并定义异常检测和定位问题。我们使用图1中给出的双层异质网路。低层网络是负责环境感应和任务运行的主体网络，它由常规传感器结点构成，各个结点受到能量和通信约束，上层网络通过获取端到端测量数据而监测底层网络的状态，它由计算和通信能力更强的测量结点构成，假设这些结点具有高信任度（比如，使用了防篡改硬件因此不会传输虚假信息[9]）。

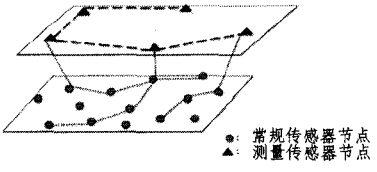


图1 分层网络架构

在定义异常检测和定位这一问题前，我们首先假设异常可能发生于被监测网络的任意链路。各条链路的异常发生是互相独立的。这些异常往往会导致端到端测量值偏离正常情况，我们可以利用这些情况来检测异常。设 $\mathcal{P}$ 表示测量结点到底层网络结点的所有端到端路径。用 ε 表示在 $\mathcal{P}$ 中出现的链路集合，于是我们使用 $|\mathcal{P}|\times|\varepsilon|$ 维矩阵 A（称为路由矩阵）来将路径与链路联系在一起。A 的每一行表示 $\mathcal{P}$ 中的一条路径，每一列表示 ε 中的一条链路。如果路径 P_i 包含链路 e_j，则元素 a_{ij} 等于1。设 x_i 表示路径 P_i 的异常指示器，y_j 表示链路 e_j 的异常指示器，也就是说，$x_i=1$ 表示 P_i 异常，$y_j=1$ 表示 e_j 异常。假设 $|\mathcal{P}|=n_p$ 且 $|\varepsilon|=n_e$，如果网络没有噪声，则如果路径 P_i 的任一链路发生异常，则 P_i 异常，即 $x_i=1-\prod_{j=1}^{n_e}(1-a_{ij}y_j)$，因此我们对 n_e 个变量有最多 n_p 个约束。对于具有多种通信约束的传感器网络，关于路径行为的有限观察数据可能不足以获得 y_js 的唯一解。更一般地，实际网络往往存在噪声[10]，即 $P(x_i=0|y)=(1-\rho_j)\prod_j(a_{ij}\rho_{ij})^{y_j}$，其中 ρ_j 表示泄漏概率，即：即使其所有链路均正常但路径 P_j 仍然异常的事件概率；ρ_{ij} 表示抑制概率，即：路径 P_i 的链路 e_j 实际异常但仍然运行正常的事件概率。综上所述，本文研究的异常检测和定位问题可以描述为（设已知如下信息）：

1）一组无线链路 $\varepsilon=\{e_1,\cdots,e_{n_e}\}$；

2）所有路径的集合 $\mathcal{P}=\{P_1,\cdots,P_{n_p}\}$；

3）路由矩阵 $A=[a_{ij}]_{n_p\times n_e}$；

4）对通信开销的约束。

我们的目标是从 $\mathcal{P}$ 中选择部分路径，在通信约束条件下采集端到端测量数据，且确定哪些链路的异常概率最大。

4 本文方法

4.1 第1阶段探查

理想状况下，为了监测网络状态，探点应该覆盖网络中的所有链路来检测出所有可能出现的异常。然而，如果对所有链路进行探查会导致严重的通信开销甚至网络拥塞。本文第1阶段探查受到如下结论启发：不同链路的风险程度是不同的，比如，攻击者更可能攻击网络中的“重要”链路。因此，应该用不同的频率和优先级对链路进行探查。根据链路的可信任度及其先前采集到的数据的荒废程度来确定链路的优先级。然后选择合适的探点，以覆盖优先级最高的链路，同时满足已知的通信约束，目标是在检测性能和探测开销间实现折衷。

（1）问题描述：对任一时刻 k 的每条链路 $e_i\in\varepsilon$，设置一个可信任度数值 $t_i(k)\in[0,1)$ 及荒废度数值 $o_i(k)\in[0,1)$。可信任度数值表示 e_i 运行正常的概率；荒废度数值表示 e_i 先前采集到的数据的荒废程度。我们使用文献[11]的方法来获得链路的可信任度，该方法利用 β 分布将反馈综合起来然后获得可信任度值。在贝叶斯统计中，β 分布可看成是二项分布参数 p 的后验概率，其中 p 是成功概率。设 $p_i(k)$ 表示链路 e_i 在时刻 k 运行正常的概率，然后假设 $p_i(k)$ 服从参数为 $\alpha_i(k)$ 和 $\beta_i(k)$ 的 β 分布，即：

$$f(p_i(k)|\alpha_i(k),\beta_i(k))=\frac{\Gamma(\alpha_i(k),\beta_i(k))p_i(k)^{\alpha_i(k)-1}(1-p_i(k))^{\beta_i(k)-1}}{\Gamma(\alpha_i(k))\Gamma(\beta_i(k))}$$

其中，$\Gamma(\cdot)$ 表示伽玛函数。已知直到时间 k 已经观察到的 e_i 状态，$\alpha_i(k)$ 和 $\beta_i(k)$ 可以表示为 $\alpha_i(k)=r_i(k)+1$，$\beta_i(k)=s_i(k)+1$，其中 $r_i(k)$ 表示 e_i 运行正常这一事件的数量，$s_i(k)$ 表示 e_i 异常这一事件的数量。信任值 $t_i(k)$ 定义为 $p_i(k)$ 的期望值，即 $t_i(k)=E[p_i(k)]=(r_i(k)+1)/(r_i(k)+s_i(k)+2)$。

因为链路行为会随着时间的变化而变化，所以时间较久的观察值对当前信任值的相关性不大。可以引入遗忘因子，于是有 $r_i(k+1)=\kappa_1 r_i(k)+I_i(k+1)$ 且 $s_i(k+1)=\kappa_2 s_i(k)+1-I_i(k+1)$，其中，$I_i(k+1)$ 是指示函数，当链路 e_i 的第 $(k+1)$ 个观测值正常时，指示函数为1。遗忘因子 κ_1 和 κ_2 为正且可设为不同值，如果我们希望对异常现象的产生进行严厉的处罚，则可以设置 $\kappa_2>\kappa_1$。

链路的荒废度开始时设为0，然后根据如下规则更新。设 $A_i(k)$ 表示在第 k 个时间间隔没有探查 e_i 这一事件，$A_i^C(k)$ 表示在第 k 个时间间隔 e_i 被探查这一事件，于是：

$$o_i(k+1)=\begin{cases}1-(1-o_i(k))e^{-\tau}, & \text{if } A_i(k)\\ 0, & \text{if } A_i^C(k)\end{cases}$$

其中，$\tau>0$ 是控制信息衰减速度的参数。换句话说，链路探查的时间距今越短，则荒废度越小。设 $w_i(k)=\rho\cdot(1-t_i(k))+(1-\rho)\cdot o_i(k)$，$\rho$ 表示调节信任度与荒废度相对重要性的参数，于是可以使用 $w_i(k)$ 作为权重来表示探查链路 e_i 的紧急程度。

第 1 阶段探查问题可以描述为一个优化问题。设 h_i 表示路径 P_i 的长度，h_0 表示通信约束，则探点遍历的链路数量不得大于 h_0。设 u_i 表示选择路径 P_i 的指标函数，且 $u_i=1$ 表示路径 P_i 被选；v_j 表示选择链路 e_j 的指示函数。于是，优化问题可以定义为：

$$\max z=\sum_{j\in\varepsilon}w_j\cdot v_j \tag{1}$$

$$\text{s. t. }\forall i\in\mathcal{P},u_i\in\{0,1\},\sum_{j\in\mathcal{P}}h_iu_i\leqslant h_0 \tag{2}$$

$$\forall j\in\varepsilon,v_j\in\{0,1\},\sum_{j\in\mathcal{P}}a_{ij}u_i-v_j\geqslant 0 \tag{3}$$

约束(2)表示通信约束，约束(3)是因为一条链路可能属于多条路径。上述优化问题属于预算有限条件下的最大覆盖问题类别[12]，且该问题是 NP 难题。Khuller 等[12]提出了一种$(1-1/e)$近似算法，实现了最优近似率。然而，该算法需要列举 $\mathcal{P}$ 中基数为 k 的所有子集且 $k\geqslant 3$，对本文来说计算量太大。我们针对该问题提出一种计算复杂度极低的高效近似算法，该算法不仅计算开销低，而且近似因子高。

(2)近似算法：近似算法及其分析基于线性规划对偶。为了求解优化问题，首先将式(2)和式(3)中的整数约束放松为一对线性约束 $0\leqslant u_i\leqslant 1$ 且 $0\leqslant v_j\leqslant 1$。可以进一步证明，$0\leqslant u_i\leqslant 1$ 和 $0\leqslant v_j\leqslant 1$ 可以等价变换为 $u_i\geqslant 0$ 和 $v_j\leqslant 1$。于是，原先优化问题的对偶问题可写为：

$$\min_{\lambda,\gamma}\lambda h_0+\sum_{j\in\varepsilon}\gamma_j \tag{4}$$

$$\text{s. t. }\forall i\in\mathcal{P},\lambda h_i+\sum_{j\in\varepsilon}a_{ij}\gamma_j\geqslant\sum_{j\in\varepsilon}a_{ij}w_j=\widetilde{w}_i \tag{5}$$

其中，$j=1,\cdots,n_e$ 时，λ、γ_j 为拉格朗日乘子，$\widetilde{w}_i$ 是路径 P_i 链路权重之和。

我们将被选路径集合表示为 χ，将被 χ 覆盖的链路集合表示为 $\varepsilon(\chi)$，χ 中路径总跳数表示为 $hops(\chi)$。需要注意，因为路径可能有链路重叠现象，所以 $hops(\chi)$ 往往大于 $\varepsilon(\chi)$。设 $q_0=\arg\max_{i\in\mathcal{P}}\widetilde{w}_i/h_i$，算法开始时有可行性对偶解 $\chi=\{P_{q_0}\}$，$hops(\chi)=h_{q_0}$，$\lambda=\widetilde{w}_{q_0}/h_{q_0}$，$\gamma_j=0$，且 $j\in\varepsilon(\chi)$。于是，在每次迭代中，基本思路是：在提高原始目标值的同时，降低对偶目标值。开始时，对偶问题的目标值是 λh_0，式(5)中只有第 q_0 个约束是活跃的(起作用)。为了降低每次迭代时的对偶目标值，我们选择式(5)中的非活跃约束，比如第 i 个约束，然后使其活跃。我们将 λ 值减去常数 β，对 γ_j 加上同样数值 β。设 $overlap=|\varepsilon(\{P_i\})\cap\varepsilon(\chi)|$ 表示 P_i 和 χ 间的重叠链路数量；改变第 i 个约束的左侧后有 $-h_i\cdot\beta+overlap\cdot\beta<0$ 且 $h_i=|\varepsilon(\{P_i\})|\geqslant overlap$。因此，妥善选择合适的正值 β 后第 i 个约束活跃。为了使对偶解具有可行性，应该避免违反其他约束，于是被选择的约束应该关联所有非活跃约束中的最小 β 值。对于已经处于活跃状态的约束，改变它们左侧等式后会有 $-h_i\cdot\beta+\sum_j a_{ij}\cdot\beta=0$，于是它们仍然活跃且不会被违反。每次迭代后，对偶目标值的变化量为 $\beta\cdot(|\varepsilon(\chi)|-h_0)$。因为 $|\varepsilon(\chi)|<hops(\chi)$，所以只要原始解可行，对偶目标值就会下降，即 $hops(\chi)\leqslant h_0$。只要有通信约束被违反，即 $hops(\chi)>h_0$，算法就会终止。第 1 阶段探点选择算法如下所示。

算法 1　第 1 阶段探点选择算法

1. 初始化：$q_0=\arg\max_{x_i\in\mathcal{P}}\widetilde{w}_i/h_i$，$\chi=\{P_{q_0}\}$，$hops(\chi)=h_{q_0}$，$\lambda=\widetilde{w}_{q_0}/h_{q_0}$，$\gamma_j=0$
2. While $hops(\chi)<h_0$
3. $\beta=inf$，$idx=0$
4. for each $i\notin\chi$
5. $overlap=|\varepsilon(\{P_i\})\cap\varepsilon(\chi)|$
6. $\beta_i=\dfrac{\lambda\cdot h_i+\sum_{j\in s}a_{ij}\gamma_j-\widetilde{w}_i}{h_i-overlap}$
7. if $\beta>\beta_i$
8. $\beta=\beta_i$，$idx=i$；
9. end
10. end
11. if $hops(\chi)+h_{idx}\leqslant h_0$
12. $\lambda=\lambda-\beta$
13. $\forall i\in\varepsilon(\chi),\gamma_i=\gamma_i+\beta$
14. $\forall i\notin\varepsilon(\chi),\gamma_i=\gamma_i$
15. $\chi=\chi\cup P_{idx}$
16. else
17. terminate(终止)
18. end
19. end

在算法 1 中，每个测量结点只负责网络的部分情况(即构成其被监测路径的链路信息)及对应的链路权重，权重则通过信任度进行本地更新。除了第 1 行的计算复杂度($\max\widetilde{w}_i/h_i$)和第 4－10 行计算复杂度($\min\beta_i$)，其他计算过程均在本地进行，也就是说，各测量结点可以运行这些步骤且无需交换信息。通过在测量结点间维护一个生成树，即可实现最小值的分布式计算。此外，第 2－19 行'While-end'语句迭代可在 $h_0/h_{\min}$ 次循环内完成，其中 $h_{\min}$ 表示 $\mathcal{P}$ 中最短路径的长度。因为测量结点的数量较小，所以该算法的通信开销很小。

4.2　第 2 阶段探测

第 2 阶段探测的目的是确定到底哪些链路可能对观察到的路径异常负责。假设第 1 阶段观察到的路径状态表示为 x_1。将第 2 阶段的探点选择策略和 ε 中的链路状态分别表示为 π 和 Y，则 π 的诊断质量可表示为 $f(\pi)=H(Y|x_1)-H(Y|x_1,\phi(\pi))$，其中 $\phi(\pi)$ 表示通过部署 π 而采集到的端到端测量数据。设 $h(\pi)$ 表示通过部署 π 而导致的通信开销，衡量为被选探点遍历的链路数量。$\tilde{h}_0$ 表示第 2 阶段的通信约束，于是该阶段的探点选择问题转化为确定策略 π^*，使：

$$\pi^*=\arg\max_{\pi}E_\pi[H(Y|x_1)-H(Y|x_1,\phi(\pi))]$$

$$\text{s. t. }h(\pi)\leqslant\tilde{h}_0$$

求解该问题等价于求解指数状态空间的有限阶段马尔科夫决策过程[13]，且该问题为 NP 难题。求解这一问题的常用方法是启发式贪婪策略，通过迭代方式选择每一步可尽量降低不确定性的探点。设 χ_C 表示先前发送的探点(包括第 1 阶段发送的探点)，且假设观察值为 $\chi_C=x_C$，通信开销为 $h(\chi_C)$，下一探点选择为：

$$i=\arg\max_{j,h_j\leqslant\tilde{h}_0-h(\chi_C)}H(Y|x_C)-H(Y|x_C,X_j) \tag{6}$$

其中，X_j 表示路径 P_j 未知状态的随机变量，h_j 表示路径 P_j 的跳数。本文提出一种启发式算法来进行第 2 阶段探测，如下所示(其中，$\tilde{h}_0$ 表示与 x_i 相对应的路径的跳数)。

算法 2 基于 LBP 的改进算法(用于第 2 阶段探点选择)

1. 初始化：执行 LBP 算法以选择首个探点 x_1，获得初始置信信息 $\mu_{\times\to\times}^{[0]}$，$k=1$，$h_k=h(x_1)$，converge=false，$S^{[0]}=\{x_1\}$
2. while $h_k\leqslant\tilde{h}_0$
3. while converged==false
4. converged=true，$S^{[1]}=j$
5. for $s=S^{[0]}$
6. for $t\in N(s)$
7. 根据新观察到的探测结果或从相邻结点新接收到的信息计算 $\mu_{s\to t}^{[1]}$
8. if $|\mu_{s\to t}^{[1]}-\mu_{s\to t}^{[0]}|>\epsilon$
9. converged=false，send $\mu_{s\to t}^{[1]}$ to t
10. $\mu_{s\to t}^{[0]}=\mu_{s\to t}^{[1]}$，$S^{[1]}=S^{[1]}\cup t$
11. end if
12. end for
13. end for
14. $S^{[0]}=S^{[1]}$
15. end while
16. $k=k+1$
17. 选择第 k 个探点 x_k，$h_k=h_{k-1}+h(x_k)$
18. $S^{[0]}=\{x_k\}$
19. end while

5 实验评估

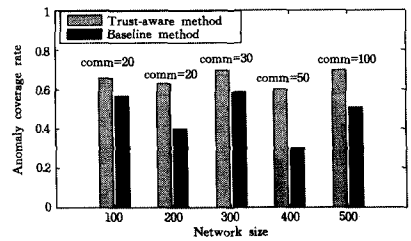
本文采用 MATLAB2012 进行仿真实验。测试所使用的软硬件环境如下：Inter(R) Core(M) CPU 4.0GHz，320G 硬盘，8.0GB 内存，Microsoft Windows 7 操作系统。表 1 总结了不同规模网络的相关参数。

表 1 网络参数

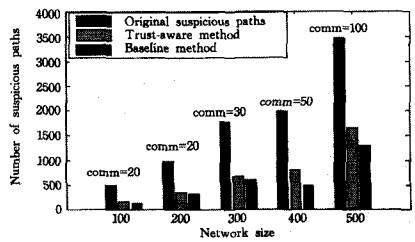
网络规模	平均结点度	测量结点数量	链路数量	路径数量	平均路径长度
100	5.36	5	268	485	3.30
200	5.06	5	506	985	5.21
300	4.68	5	702	1779	5.30
400	5.09	5	1049	1985	6.39
500	6.27	7	1568	3472	7.83

在实验中，网络每条链路出现的异常建模为一个伯努利过程。出于简便考虑，我们假设一旦出现异常，该异常会一直持续到被检测到为止。我们给出不同的实验设置(网络规模、通信约束和异常密度)条件下两个探测阶段的结果。由于当前网络往往没有考虑严格的通信开销，因此设计一种受到通信约束的简单的基准算法，然后将本文算法与该算法做比较。假设基准算法无法感知信任度信息。它的步骤与基于信任度的算法相同，但是假设每条链路的异常概率均为 0.5。

(1)第 1 阶段探测：目标是在通信约束下覆盖尽可能多的异常链路，同时缩小网络的可疑区域范围。我们提出使用两个指标评估算法性能。第一个指标是异常覆盖率，为候选探点覆盖的异常链路与所有异常链路数量之比。第二个指标是可疑区域，表示为第 1 阶段探测之后的可疑路径数量。因为第 1 阶段只会检测到端到端异常行为(路径异常)，所以可疑路径表示与被检测出来的异常路径相交的路径。它们也是可能用于第 2 阶段探测以定位具体异常链路的候选路径。图 2(a)给出了给定通信约束条件时，本文算法和基准算法在不同网络下的异常覆盖率。图 2(b)给出了这些网络在第 1 阶段探测后的可疑路径数量。两个图中的 X 轴表示网络规模，其他网络参数如表 1 所列。柱状图上方的数据表示第 1 阶段使用的相应通信开销。网络规模越大，我们设立的通信约束也越大。这些网络的异常密度为 20%左右，也就是说，网络中 20%左右的链路存在异常。



(a) Anomaly coverage rate

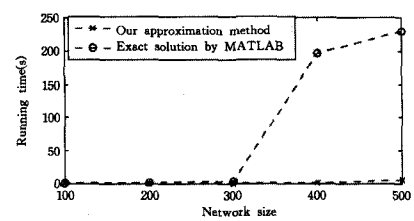


(b) Number of suspicious paths

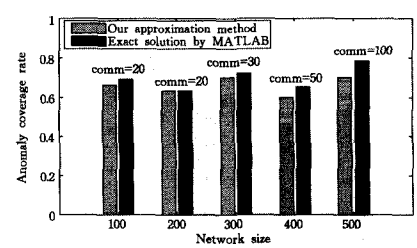
图 2 本文算法和基准算法的比较

图 2(a)表明，在通信约束相同时，本文算法覆盖的异常链路多于基准算法。图 2(b)表明，本文算法和基准算法均缩小了可疑区域范围：候选探点数量显著降低。我们也注意到，基准算法的可疑区域下降幅度较小，这是因为它检测出的异常路径数量很少，相交的异常路径数量也变少。

如前文所述，第一阶段探点选择的另一个因素就是算法的计算复杂度。我们在 MATLAB 中从速度和性能两个角度，将本文近似算法与 MATLAB 整数规划精确求解方法进行对比。实验设置与前文相同。图 3(a)给出了不同网络条件下本文算法和精确求解方法的运行时间。图 3(b)给出了两种算法的异常覆盖率。



(a) Running time for the two methods



(b) Anomlay coverage rate for the two methods

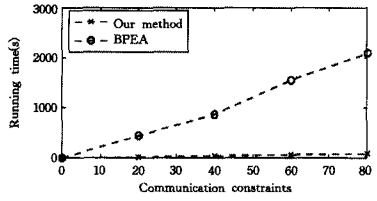
图 3 近似算法与准确求解算法的比较

对于小规模网络，由于通信约束较小，预算有限条件下的最大覆盖问题的规模也较小，因此精确求解方法的速度比近似方法慢3倍左右。当网络规模较大时，近似算法相对精确方法的性能优势更加明显。例如，当网络规模超过300时，近似算法的速度比精确方法快出一个数量级，而异常覆盖率下降幅度不到0.1。考虑到精确方法的运行时间很长(分钟)，近似算法的性能衰减是可以接受的。

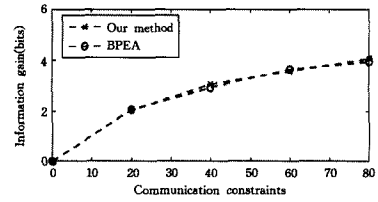
(2)第2阶段探测：第2阶段探查的目的是确定到底哪些链路导致观察到的链路异常。本文从漏警率(MDR)和虚警率(FAR)两方面进行性能评估。漏警率表示未被检测到的异常链路的比例，虚警率表示被错误诊断为异常的正常链路的比例。设ε表示被检测的链路集合，ε_A表示异常链路集合，ε_D表示被正常定位的异常链路集合，于是：

$$MDR=1-\frac{|\varepsilon_A\cap\varepsilon_D|}{|\varepsilon_A|},FAR=\frac{|\varepsilon_D\cap(\varepsilon-\varepsilon_A)|}{|\varepsilon-\varepsilon_A|}$$

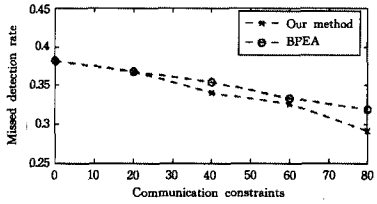
我们利用Kevin Murphy贝叶斯网络工具箱[15]部署了第2阶段的探点选择算法。为了证明本文启发式算法可以提升算法效率，我们部署了Zheng等人在文献[14]中提出的BPEA算法。BPEA算法也是重复运行LBP算法进行在线探点选择，该算法的推断准确率较高。然而，BPEA算法没有利用LBP算法重复运行时的冗余，因此计算复杂度很大。图4给出了400结点网络时本文算法和BPEA算法的性能比较。网络其他参数设置见表1。在该实验中，异常密度为30%左右。第一阶段的探点选择通信约束为50。图4(a)比较了两种算法的运行时间。图4(b)比较了两种算法的信息增益(比特)。图4(c)给出了漏警率。从图4可以看到，两种算法的虚警率均较低，在0.04左右，发送更多的探点对虚警率没有太大影响，所以此处略去。图中X轴对应于第二阶段的通信约束。



(a)Speed of the two methods



(b)Information gain of the two methods



(c)Missed detection rate of the two methods

图4 BPEA和本文算法的比较

可以看出，本文的启发式算法确实提高了探点选择过程的速度，且比BFEA速度高出一个数量级。当通信开销为80跳时，BPEA算法需要35分钟，对于实践中的探点选择显然不可行，而本文方法只要1分半钟的时间。当通信增益相同时，两种算法降低不确定性的信息增益基本相同，且本文方法的漏警率略低。

结束语 本文提出一种改进的异常检测和定位策略，利用端到端测量值检测并定位出资源有限无线传感器网络的异常。与先前研究专注于检测和定位精度不同，本文研究重点是在推断精度和传感器网络资源消耗间实现折衷。我们提出了一种高效的双阶段探测方案，利用链路信任度信息在推断精度和推测开销间实现良好折衷。仿真结果证明了本文算法的有效性。在下步工作中，我们将在实际传感器网络测试床上部署本文算法，对其性能做进一步评估。

参考文献

[1] 钱志鸿，王义君. 面向物联网的无线传感器网络综述[J]. 电子与信息学报，2013，35(1)：215-227

[2] 张鹏，冯欣，周建国. 无线传感器网络中基于空间关联性的聚类异常检测算法[J]. 计算机应用研究，2013，30(5)：1370-1372

[3] 姜旭宝，李光耀，连朔. 基于变宽直方图的无线传感器网络异常数据检测算法[J]. 计算机应用，2011，31(3)：694-697

[4] 肖政宏，陈志刚，李庆华. WSN中基于分布式机器学习的异常检测仿真研究[J]. 系统仿真学报，2011，23(1)：121-127

[5] 朱翠涛，瞿毅. 基于压缩感知的稀疏事件检测[J]. 中南民族大学学报：自然科学版，2011，30(1)：81-83

[6] Yigal Bejerano. Coverage Verification without Location Information [J]. IEEE Transactions on Mobile Computing, 2012, 11(4)：631-643

[7] 张玉琴，秦拯. 无线传感器网络中基于分簇的节点定位异常检测[J]. 计算机应用，2010，27(3)：1139-1141

[8] 杨黎斌，慕德俊，蔡晓妍. 基于核聚类的无线传感器网络异常检测方案[J]. 传感技术学报，2008，21(8)：1442-1447

[9] Lu K，Qian Y，Guizani M，et al. A framework for a distributed key management scheme in heterogeneous wireless sensor networks [J]. IEEE Transactions on Wireless Communications，2008，7(2)：639-647

[10] Cheng L，Qiu X，Meng L，et al. efficient active probing for fault diagnosis in large scale and noisy networks[C]// INFOCOM，2010 Proceedings IEEE. IEEE，2010：1-9

[11] Jsang A，Ismail R. The beta reputation system[C]// Proceedings of the 15th Bled Electronic Commerce Conference. 2002：41-55

[12] Khuller S，Moss A，Naor J S. The budgeted maximum coverage problem [J]. Information Processing Letters，1999，70(1)：39-45

[13] Golovin D，Krause A. Adaptive Sub-modularity：A New Approach to Active Learning and Stochastic Optimization[C]// COLT. 2010：333-345

[14] Zheng A X，Rish I，Beygelzimer A. Efficient test selection in active diagnosis via entropy approximation [J]. arXiv preprint arXiv：1207.1418，2012

[15] Murphy K. The bayes net toolbox for matlab [J]. Computing science and statistics，2001，33(2)：1024-1034