无线传感器网络中基于双阈值的分布式监测算法

毕 冉,李建中,高 宏
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 基于单阈值的监测算法降低了警报的准确率,因此研究基于双阈值的监测方法,即带有概率保证的约束违反的监测具有重要意义.首先,基于监测结果的概率阈值语义,研究了节点的双阈值监测问题.其次,给出了感知数据大于监测阈值的概率的紧上界,提出了基于双阈值的分布式监测算法.第三,给出了根据精度要求确定优化样本容量的数学方法,提出了基于抽样的近似簇监测算法.理论分析和实验结果验证了提出的监测算法的高效性.

关键词: 双阈值监测; 抽样算法; 簇监测算法
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2014)08-1594-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.08.021

Dual Threshold Based Distributed Monitoring Algorithm in Wireless Sensor Network

BI Ran, LI Jian-zhong, GAO Hong
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: Sole threshold based monitoring algorithms bring down the accuracy of alarm, hence researching on dual threshold based monitoring technique, that is, constraint violation monitoring with probability guarantee has significant meanings. Firstly, according to the semantics of probability threshold on the monitoring result, dual threshold monitoring of sensor node was investigated. Secondly, a tight upper bound of the probability of the sensing data larger than monitoring threshold was given, and dual threshold based distributed monitoring algorithm was proposed. Thirdly, based on the given accuracy requirement, a mathematical method to determine an optimal sample was provided. A sampling based approximate cluster monitoring algorithm was proposed. The theoretical analysis and performance evaluation demonstrate the efficiency of the proposed algorithms.

Key words: dual threshold monitoring; sampling algorithm; cluster monitoring algorithm

1 引言

无线传感器网络逐渐走进人们的生活,满足了人们低成本地获取环境科学等众多领域的数据的需求[1~3].在基于传感器网络的监测应用中,当感知数据超出用户给定的阈值时,传感器节点向 Sink 节点发送警报信息[4~8].准确的警报信息有助于用户及时地做出正确的分析和决策,有效地避免了人员和财产损失事故的发生.受到感知硬件误差和环境噪声的影响,不确定性和误差广泛地存在于传感器节点采集的感知数据中[9].当噪声扰动或仪器误差引起感知值的严重偏离时,基于单一阈值的监测方法将导致较高的警报误报率和警报漏报率.因此基于双阈值的监测方法,即带有概率保证的返回结果降低了扰动和硬件误差对警报信息的影响,更适合实际的监测应用.

感知数据经常带有噪声和误差,因此以不确定数据形式表示的感知数据比确定的感知值更好地体现了所监测物理变量的实际情况[10~13].文献[14]率先开展了分布式环境下基于概率阈值的感知数据监测的研究,提出了分布式概率阈值监测算法(DPTM).然而在实际的传感器网络的应用中,很难获取感知数据的分布信息.

2 相关工作

无线传感器网络的阈值监测问题引起了人们的广泛关注,提出了有效的阈值监测算法[5~9].针对聚集和的阈值监测问题,文献[15]提出了零松弛的阈值设置方法,当节点的感知值大于 T/n 时,节点向 Sink 节点发送阈值违反信息,然后 Sink 节点收集全网的感知数据考察其聚集和是否大于阈值 T.由于收集全网感知数据消耗较高的通信能量,文献[5]以最小化收集全网感知数

收稿日期:2013-07-25;修回日期:2013-11-18;责任编辑:马兰英
基金项目:国家 973 重点基础研究发展规划(No.2012CB316200);国家自然科学基金(No.61190115, No.61033015, No.60831160525, No.60933001, No.61300225);中央高校基本科研业务费专项基金(No.HIT.NSRIF.201180)

据的概率为优化目标,提出了阈值监测的统一框架.文献[11]提出了自适应的非零松弛阈值设置方法,Meng[16]等开发了基于时间窗口的状态监测系统.然而这些方法将感知值视为确定型数据,没有考虑噪声扰动和误差对警报准确率的影响.文献[14]将感知数据视为不确定数据并假设感知数据的概率分布已知,提出了分布式概率阈值监测算法.然而,在实际的应用中,我们很难精确地获取感知数据的概率分布函数;而且感知值通常是以时间为自变量的函数.更新感知数据的概率分布函数需要消耗较高的通信能量,不适合以电池为能量供给的传感器网络.

3　分布式监测算法

3.1　问题定义

不妨设传感器网络由 M 个节点组成,$s_i(t)$表示 i 节点在 t 时刻的感知值,将 sup、inf 分别记为感知值的上界和下界.由于感知数据具有不确定性,本文将感知数据 $s_i(t)$视为随机变量,$\mu_i(t)$表示 i 节点在 t 时刻感知值的期望,当上下文语义清楚时,$\mu_i(t)$简记为 μ_t.

定义 1　$(\beta,\tau)_t$-双阈值监测:对于给定的监测阈值 β 和概率阈值 τ,i 节点在 t 时刻发出警报当且仅当 $\Pr[s_i(t)>\beta]>\tau$.

由于准确地获取或估计感知数据的概率密度函数或分布函数十分困难,因此本节提出一种新颖、高效的基于双阈值的监测方法.

3.2　数学基础

引理 1　如果 x 为取值[0,1]的随机变量且 $\mathrm{E}[x]=\mu$,那么对 $\forall\, h\neq 0$,$\mathrm{E}[e^{xh}]\leqslant e^h\mu+1-\mu$

证明　令 $f(x)=e^{hx}$,可知 $f(x)$的二阶导数 $f''(x)=h^2e^{hx}$,那么对 $\forall\, h\neq 0$,$f''(x)>0$,因此 $f(x)=e^{hx}$为关于 x 的凸函数,那么对 $\forall\, x_1,x_2\in R$,$\alpha\in[0,1]$有下式成立 $f(\alpha x_1+(1-\alpha)x_2)\leqslant\alpha f(x_1)+(1-\alpha)f(x_2)$.

令 $\alpha=1-x$,由于 x 是取值为[0,1]的随机变量所以 $\alpha\in[0,1]$,那么有下式成立:

$$e^{hx}=f(x)=f((1-x)\times 0+x\times 1)\leqslant(1-x)e^0+xe^h$$
$$\Rightarrow \mathrm{E}[e^{hx}]\leqslant\mathrm{E}[(1-x)+xe^h]=e^h\mu+1-\mu.$$

定理 1　不妨设 t 时刻 i 节点的感知数据为 $s_i(t)$且 $\mathrm{E}[s_i(t)]=\mu_t$.当 $\mu_t<\beta$ 时,$\Pr[s_i(t)\geqslant\beta]\leqslant\exp\{-\ell(\hat{\mu}_t+\lambda,\hat{\mu}_t)\}$,其中 $\ell(x_1,x_2)=x_1(\ln x_1-\ln x_2)+(1-x_1)(\ln(1-x_1)-\ln(1-x_2))$,$\hat{\mu}_t=\dfrac{\mu_t}{sup}$,$\lambda=\dfrac{\beta-\mu_t}{sup}$

证明　不妨设 $sup>\beta$.令 $x_i(t)=\dfrac{s_i(t)}{sup}$,$\hat{u}_t=\dfrac{\mu_t}{sup}$,那么 $x_i(t)$为取值[0,1]的随机变量且 $\mathrm{E}[x_i(t)]=\hat{u}_t$.$\Pr[s_i(t)\geqslant\beta]=\Pr[\dfrac{s_i(t)}{sup}\geqslant\dfrac{\beta}{sup}]=\Pr[x_i(t)\geqslant\dfrac{\beta}{sup}]$.令 $\gamma=\dfrac{\beta}{sup}$,对 $\forall\, h\in(0,\infty)$,由马尔可夫不等式知 $\Pr[x_i(t)\geqslant\gamma]=\Pr[e^{x_i(t)h}\geqslant e^{\gamma h}]\leqslant e^{-\gamma h}\mathrm{E}[e^{x_i(t)h}]$.由引理 1 知 $\mathrm{E}[e^{hx_i(t)}]\leqslant e^h\hat{\mu}_t+1-\hat{\mu}_t$,于是对于 $\forall\, h>0$ 我们推出 $\Pr[s_i(t)\geqslant\beta]\leqslant e^{-\gamma h}(e^h\hat{\mu}_t+1-\hat{\mu}_t)$.为了得到 $\Pr[s_i(t)\geqslant\beta]$一个较小的上界,我们需要取最优的 h^*值,使得 $e^{-\gamma h^*}(e^{h^*}\hat{\mu}_t+1-\hat{\mu}_t)=\min\limits_{h>0}\{e^{-\gamma h}(e^h\hat{\mu}_t+1-\hat{\mu}_t)\}$.令 $y(h)=e^{-\gamma h}(e^h\hat{\mu}_t+1-\hat{\mu}_t)=\exp\{\ln(e^h\hat{\mu}_t+1-\hat{\mu}_t)-\gamma h\}$,可知 $h^*=\ln\dfrac{\gamma(1-\hat{\mu}_t)}{(1-\gamma)\hat{\mu}_t}$使得$\dfrac{\partial y}{\partial h}=y(h)\left(\dfrac{\hat{\mu}_te^h}{\hat{\mu}_te^h+1-\hat{\mu}_t}-\gamma\right)=0$.由于 $h^*=\ln\left(\dfrac{\gamma(1-\hat{\mu}_t)}{(1-\gamma)\hat{\mu}_t}\right)\in(0,\infty)$.因此 $h=h^*$时,$y(h)$取得最小值.于是我们可推出下式:

$$\begin{aligned}\Pr[s_i(t)\geqslant\beta]&\leqslant e^{-\gamma h^*}(e^{h^*}\hat{\mu}_t+1-\hat{\mu}_t)\\&=\exp\left\{\ln\left(\frac{\gamma(1-\hat{\mu}_t)}{(1-\gamma)}+1-\hat{\mu}_t\right)-\gamma\ln\left(\frac{\gamma(1-\hat{\mu}_t)}{(1-\gamma)\hat{\mu}_t}\right)\right\}\\&=\exp\left\{-\left(\gamma\ln\left(\frac{\gamma}{\hat{\mu}_t}\right)+(1-\gamma)\ln\left(\frac{1-\gamma}{1-\hat{\mu}_t}\right)\right)\right\}\end{aligned}$$

令 $\ell(x_1,x_2)=x_1\ln\left(\dfrac{x_1}{x_2}\right)+(1-x_1)\ln\left(\dfrac{1-x_1}{1-x_2}\right)$,其中 x_1,$x_2\in(0,1)$,则 $\exp\left\{-\left(\gamma\ln\left(\dfrac{\gamma}{\hat{\mu}_t}\right)+(1-\gamma)\ln\left(\dfrac{1-\gamma}{1-\hat{\mu}_t}\right)\right)\right\}=\exp\{-\ell(\gamma,\hat{\mu}_t)\}$.因此 $\Pr[s_i(t)\geqslant\beta]\leqslant\exp\{-\ell(\gamma,\hat{\mu}_t)\}$.令 $\beta=\mu_t+sup\lambda$,于是 $\gamma=\dfrac{\beta}{sup}=\dfrac{\mu_t+sup\lambda}{sup}=\hat{\mu}_t+\lambda$,所以 $\Pr[s_i(t)\geqslant\beta]\leqslant\exp\{-\ell(\hat{\mu}_t+\lambda,\hat{\mu}_t)\}$.

推论 1　对于用户给定的监测阈值(β,τ),当感知数据的期望 $\mu_t<\beta$ 时,如果下式成立:

$$\frac{\beta}{sup}\ln\left(\frac{\beta}{\mu_t}\right)+\left(1-\frac{\beta}{sup}\right)\ln\left(\frac{sup-\beta}{sup-\mu_t}\right)\geqslant\ln\tau^{-1}$$

那么传感器节点不需向 sink 节点发送警报.

定理 2　不妨设 t 时刻 i 节点的感知数据为 $s_i(t)$且 $\mathrm{E}[s_i(t)]=\mu_t$.当 $\mu_t>\beta$ 时,$\Pr[s_i(t)\leqslant\beta]\leqslant\exp\{-\ell(\hat{\mu}_t-\lambda,\hat{\mu}_t\}$,其中 $\hat{\mu}_t=\dfrac{\mu_t}{sup}$,$\lambda=\dfrac{\mu_t-\beta}{sup}$.

证明　定理 2 的证明过程与定理 1 类似,下面我们简述定理 2 的证明过程.令 $x_i(t)=\dfrac{s_i(t)}{sup}$,$\hat{\mu}_t=\dfrac{\mu_t}{sup}$,那么 $\Pr[s_i(t)\leqslant\beta]=\Pr[x_i(t)\leqslant\dfrac{\beta}{sup}]$.令 $\gamma=\dfrac{\beta}{sup}$,对 $\forall\, h\in(-\infty,0)$,由马尔可夫不等式知 $\Pr[x_i(t)\leqslant\gamma]\leqslant e^{-\gamma h}\mathrm{E}[e^{x_i(t)h}]$.由引理 1 知 $\mathrm{E}[e^{hx_i(t)}]\leqslant e^h\hat{\mu}_t+1-\hat{\mu}_t$,因此对于 $\forall\, h<0$ 我们推出下面的不等式:$\Pr[s_i(t)\leqslant\beta]\leqslant e^{-\gamma h}(e^h\hat{\mu}_t+1-\hat{\mu}_t)$.我们需要取最优的 h^*值,使得 $e^{-\gamma h^*}(e^{h^*}\hat{\mu}_t+1-\hat{\mu}_t)=\min\limits_{h<0}\{e^{-\gamma h}(e^h\hat{\mu}_t+1-\hat{\mu}_t)\}$.令 $y(h)=e^{-\gamma h}(e^h\hat{\mu}_t+1-\hat{\mu}_t)$可知 $h^*=\ln\dfrac{\gamma(1-\hat{\mu}_t)}{(1-\gamma)\hat{\mu}_t}$使得$\dfrac{\partial y}{\partial h}=0$.

$\because \gamma < \hat{\mu}_t \therefore \frac{\gamma(1-\hat{\mu}_t)}{(1-\gamma)\hat{\mu}_t} < 1, h^* = \ln\left(\frac{\gamma(1-\hat{\mu}_t)}{(1-\gamma)\hat{\mu}_t}\right) \in (-\infty, 0)$,

因此当 $h = h^*$ 时,$y(h)$取得最小值.于是我们可推出下式:

$$\begin{aligned}\Pr[s_i(t) \leqslant \beta] &\leqslant e^{-\gamma h^*}(e^{h^*}\hat{\mu}_t + 1 - \hat{\mu}_t) \\ &= \exp\left\{-\left(\gamma\ln\left(\frac{\gamma}{\hat{\mu}_t}\right) + (1-\gamma)\ln\left(\frac{1-\gamma}{1-\hat{\mu}_t}\right)\right)\right\} \\ &= \exp\{-\ell(\gamma, \hat{\mu}_t)\}\end{aligned}$$

因此 $\Pr[s_i(t) \leqslant \beta] \leqslant \exp\{-\ell(\gamma, \hat{\mu}_t)\}$.由于 $\beta < \mu_t$,令 $\beta = \mu_t - sup\lambda$,于是 $\gamma = \frac{\beta}{sup} = \frac{\mu_t - sup\lambda}{sup} = \hat{\mu}_t - \lambda$,所以 $\Pr[s_i(t) \leqslant \beta] \leqslant \exp\{-\ell(\hat{\mu}_t - \lambda, \hat{\mu}_t)\}$.

推论 2 对于用户给定的监测阈值(β, τ),当感知数据的期望 $\mu_t > \beta$ 时,如果下式成立:

$$\frac{\beta}{sup}\ln\left(\frac{\beta}{\mu_t}\right) + \left(1 - \frac{\beta}{sup}\right)\ln\left(\frac{sup - \beta}{sup - \mu_t}\right) > \ln(1-\tau)^{-1}$$

那么传感器节点需向 sink 节点发送警报信息.

3.3 基于双阈值的分布式监测算法

根据定理 1 和定理 2 给出的尾部概率界估计公式,$(\beta, \tau)_t$-双阈值监测问题可如下定义:

输入:

(1)监测阈值 β,概率阈值 τ.

(2)感知值 $s_i(t)$及其期望 $\mu_i(t)$,节点状态.

基于双阈值的分布式监测算法(Dual Threshold based Distributed Monitoring Algorithm)的主要思想如下.

第一种情况: $\mathrm{E}[s_i(t)] < \beta$

由定理 1 知 $\exp\{-\ell(\hat{\mu}_t + \lambda, \hat{\mu}_t)\}$为 $\Pr[s_i(t) \geqslant \beta]$的上界,那么 $\Pr[s_i(t) > \beta] \leqslant \exp\{-\ell(\hat{\mu}_t + \lambda, \hat{\mu}_t)\}$

(Ⅰ)当 $\exp\{-\ell(\hat{\mu}_t + \lambda, \hat{\mu}_t)\} \leqslant \tau$ 时

可知 $\Pr[s_i(t) > \beta] \leqslant \tau$,那么节点 i 不必发出警报并将自身设置为 Normal(正常)状态.

(Ⅱ)当 $\exp\{-\ell(\hat{\mu}_t + \lambda, \hat{\mu}_t)\} > \tau$ 时

由于 $\exp\{-\ell(\hat{\mu}_t + \lambda, \hat{\mu}_t)\}$为 $\Pr[s_i(t) > \beta]$的上界,根据定理 1 不能确定 $s_i(t) > \beta$ 的概率是否大于用户指定的概率阈值 τ,此时节点将自身状态设置为 A_c(警报候选)状态.若 $\exp\{-\ell(\hat{\mu}_{t+1} + \lambda, \hat{\mu}_{t+1})\} \leqslant \tau$,那么节点 i 不必发出警报并将自身状态设置为 Normal 状态.当 $\exp\{-\ell(\hat{\mu}_{t+1} + \lambda, \hat{\mu}_{t+1})\} > \tau$ 时,如果用户不希望漏报任何警报,那么节点将采取积极的预警策略,此时节点向 Sink 节点发送警报并将自身设置为 Alarm 状态.如果用户对警报的正确率有较高的要求,那么节点考察 $t+1$ 时刻的感知值,如果 $s_i(t+1) > \beta$,那么节点发送警报并将自身设置为 Alarm 状态,否则节点不发送警报并将自身设置为 A_c 状态.

第二种情况: $\mathrm{E}[s_i(t)] > \beta$

由定理 2 知 $\Pr[s_i(t) \leqslant \beta] \leqslant \exp\{-\ell(\hat{\mu}_t - \lambda, \hat{\mu}_t)\}$,那么 $\Pr[s_i(t) > \beta] \geqslant 1 - \exp\{-\ell(\hat{\mu}_t - \lambda, \hat{\mu}_t)\}$

(Ⅰ)当 $\exp\{-\ell(\hat{\mu}_t - \lambda, \hat{\mu}_t)\} < 1 - \tau$ 时

可知 $\Pr[s_i(t) > \beta] > \tau$,那么节点 i 发出警报并将自身设置为 Alarm 状态.

(Ⅱ)当 $\exp\{-\ell(\hat{\mu}_t - \lambda, \hat{\mu}_t)\} \geqslant 1 - \tau$ 时

由于 $\exp\{-\ell(\hat{\mu}_t - \lambda, \hat{\mu}_t)\}$为 $\Pr[s_i(t) \leqslant \beta]$的上界,那么 $1 - \exp\{-\ell(\hat{\mu}_t - \lambda, \hat{\mu}_t)\}$为 $\Pr[s_i(t) > \beta]$的下界,所以当 $1 - \exp\{-\ell(\hat{\mu}_t - \lambda, \hat{\mu}_t)\} \leqslant \tau$ 时,根据定理 2 不能确定 $s_i(t) > \beta$ 的概率是否大于概率阈值 τ,此时节点将自身状态设置为 N_c(正常候选)状态.若 $\exp\{-\ell(\hat{\mu}_{t+1} - \lambda, \hat{\mu}_{t+1})\} < 1 - \tau$,那么节点 i 发出警报并将自身状态设置为 Alarm 状态.当 $\exp\{-\ell(\hat{\mu}_{t+1} - \lambda, \hat{\mu}_{t+1})\} \geqslant \tau$ 时,如果用户不希望漏报任何警报,那么节点将采取积极预警策略,此时节点向 Sink 节点发送警报;如果用户对警报的正确率有较高的要求,那么节点考察 $t+1$ 时刻的感知值,若 $s_i(t+1) > \beta$ 则节点发送警报否则节点不发送警报并将自身设置为 N_c 状态.

第三种情况: $\mathrm{E}[s_i(t)] = \beta$

鉴于已有的尾部概率界估计失效,我们根据用户的不同应用需求,提出积极的预警策略和消极的预警策略.如果用户容忍较高的警报误报率,节点将采取积极的预警策略:令 $\beta' = \beta - \varepsilon$,节点考察 $\Pr[s(t) > \beta']$是否大于阈值 τ;若用户对警报的正确率有较高的要求,节点将采取消极的警报策略:令 $\beta' = \beta + \varepsilon$,节点考察 $\Pr[s(t) > \beta']$是否大于阈值 τ.

3.4 算法性能分析

对于用户给定的监测阈值 β 和概率阈值 τ,由推论 1 和推论 2 可知尾部概率界的计算复杂度为 $O(1)$.与已有的方法相比[11],本节给出的尾部概率界的估计公式仅需感知数据的期望,降低了因参数估计不准确导致的尾部概率界的计算误差,降低了扰动或者仪器误差引起的警报误报率.

4 (α, τ)-簇阈值监测算法

4.1 问题定义

定义 2 (α, τ)-簇阈值监测[11]:不妨设簇 C 由 n 个传感器节点组成,令 $S(t) = \sum_{i=1}^{n} s_i(t)$.对于给定的监测阈值 α 以及概率阈值 τ,簇头节点在 t 时刻发出警报当且仅当 $\Pr[S(t) > \alpha] > \tau$.

精确计算 $\Pr[S(t) > \alpha]$的开销较高[11],本节研究基于抽样的聚集值监测的近似算法.令 $Sum(t_1, t_N) = \{S(t_1), S(t_2), \cdots, S(t_N)\}$,$\alpha Sum(t_1, t_N) = \{S(t_j) \mid S(t_j) > \alpha, j = 1, \cdots, N\}$.当上下文语义清楚时,多重集合 $\alpha Sum(t_1, t_N)$简记为 $\alpha(t_1, t_N)$.由泊松大数定律知,频率收敛于概率的平均值,因此在 $t_1 \sim t_N$ 时间区间内,$\Pr[S$

$(t)>\alpha]\approx\frac{|\alpha(t_1,t_N)|}{N}$. 因此,令 $Pr[S(t_1,t_N)>\alpha]=\frac{|\alpha(t_1,t_N)|}{N}$. 如果 $\frac{|\alpha(t_1,t_N)|}{N}>\tau$,那么簇头向 Sink 节点发送警报;否则簇头不向 Sink 节点发送警报[11].

定义 3 (ε,δ) - 近似估计:对于给定的 $\varepsilon>0,\delta>0$,$\hat{Pr}[S(t_1,t_N)>\alpha]$称为 $Pr[S(t_1,t_N)>\alpha]$的(ε,δ) - 近似估计,如果 $\hat{Pr}[S(t_1,t_N)>\alpha]$满足下式:

$$\Pr[|\hat{Pr}[S(t_1,t_N)>\alpha]-Pr[S(t_1,t_N)>\alpha]|\geqslant\varepsilon]\leqslant\delta$$

定义 4 (ε,δ) - 近似(α,τ)双阈值监测:如果 $\hat{Pr}[S(t_1,t_N)>\alpha]$是 $Pr[S(t_1,t_N)>\alpha]$的(ε,δ) - 近似估计,当 $\hat{Pr}[S(t_1,t_N)>\alpha]>\tau$ 时,簇头发送警报信息.

4.2 数学基础

不妨设 $U_m=\{S(t_{k1}),S(t_{k2}),\cdots,S(t_{km})\}$是对 $Sum(t_1,t_N)$进行均衡随机抽样获得的容量为 m 的样本,令 $\alpha(U_m)=\{S(t_{kj})|S(t_{kj})>\alpha,j=1,\cdots,m\}$,那么 $Pr[S(t_1,t_N)>\alpha]$的估计器 $\hat{Pr}[S(t_1,t_N)>\alpha]$可按下面的公式计算:

$$\hat{Pr}[S(t_1,t_N)>\alpha]=\frac{|\alpha(U_m)|}{m}$$

定理 3 $\hat{Pr}[S(t_1,t_N)>\alpha]$是 $Pr[S(t_1,t_N)>\alpha]$的无偏估计,即 $E[\hat{Pr}[S(t_1,t_N)>\alpha]]=Pr[S(t_1,t_N)>\alpha]$.

证明 由均衡随机抽样的性质知,$\forall S(t_j)\in Sum(t_1,t_N)$被抽到样本 U_m 的概率为 $p=\frac{m}{N}$. 令 $\alpha(t_1,t_N)=\{S(t_{h1}),\cdots,S(t_{h|\alpha(t_1,t_N)|})\}$,那么 $\forall S(t_{hj})\in\alpha(t_1,t_N)$被随机抽到样本集合 U_m 的概率亦为 p. 令随机变量 $y_j=1$,当且仅当 $S(t_{hj})$属于随机样本集合 U_m;否则 $y_j=0$. 于是 $Y=\sum_{j=1}^{|\alpha(t_1,t_N)|}y_j$ 为样本 U_m 中聚集和大于 α 的个数. 易知 Y 服从 $B(|\alpha(t_1,t_N)|,p)$ 的二项分布并且 $E[Y]=|\alpha(t_1,t_N)|p$,则 $E[\hat{Pr}[S(t_1,t_N)>\alpha]]=\frac{E[Y]}{m}=\frac{|\alpha(t_1,t_N)|}{N}$,即 $\hat{Pr}[S(t_1,t_N)>\alpha]$为 $Pr[S(t_1,t_N)>\alpha]$的无偏估计.

定理 4 对于给定的 $\varepsilon>0,\delta>0$,如果样本容量满足 $m\geqslant\frac{1}{\delta\varepsilon^2}$,那么根据样本 U_m 计算的估计器 $\hat{Pr}[S(t_1,t_N)>\alpha]$满足下式:

$$\Pr[|\hat{Pr}[S(t_1,t_N)>\alpha]-Pr[S(t_1,t_N)>\alpha]|\geqslant\varepsilon]\leqslant\delta$$

证明 由于 $Y=\sum_{j=1}^{|\alpha(t_1,t_N)|}y_j$ 为样本 U_m 中聚集和大于 α 的个数并且服从 $B(|\alpha(t_1,t_N)|,p)$ 的二项分布,其中 $p=\frac{m}{N}$. 因此 $E[Y]=p|\alpha(t_1,t_N)|$,$Var[Y]=|\alpha(t_1,t_N)|p(1-p)<|\alpha(t_1,t_N)|p$. 由切比雪夫不等式知 $\Pr[|Y-E[Y]|\geqslant a]\leqslant\frac{Var[Y]}{a^2}$,令 $a=\sqrt{\frac{Var[Y]}{\delta}}$,可知 $\Pr\left[|Y-E[Y]|\geqslant\sqrt{\frac{Var[Y]}{\delta}}\right]\leqslant\delta$ 成立,易知 $\Pr\left[|p^{-1}Y-E[p^{-1}Y]|\geqslant\sqrt{\frac{Var[p^{-1}Y]}{\delta}}\right]\leqslant\delta$.

$Var[p^{-1}Y]<|\alpha(t_1,t_N)|p^{-1}<\frac{N^2}{m}$,于是有下式成立:

$$\Pr\left[|Yp^{-1}-E[Yp^{-1}]|\geqslant\frac{1}{\sqrt{\delta}}\frac{N}{\sqrt{m}}\right]\leqslant$$

$$\Pr\left[|Yp^{-1}-E[Yp^{-1}]|\geqslant\sqrt{\frac{Var[p^{-1}Y]}{\delta}}\right]\leqslant\delta$$

由上式知 $\Pr\left[\left|\frac{Yp^{-1}}{N}-\frac{E[Yp^{-1}]}{N}\right|\geqslant\frac{1}{\sqrt{\delta m}}\right]\leqslant\delta$,那么 $\Pr\left[\left|\frac{Y}{m}-\frac{|\alpha(t_1,t_N)|}{N}\right|\geqslant\frac{1}{\sqrt{\delta m}}\right]\leqslant\delta$. 由于 $m\geqslant\frac{1}{\delta\varepsilon^2}$,可推出 $\Pr\left[\left|\frac{Y}{m}-\frac{|\alpha(t_1,t_N)|}{N}\right|\geqslant\varepsilon\right]\leqslant\delta$. 因此 $\Pr[|\hat{Pr}[S(t_1,t_N)>\alpha]-Pr[S(t_1,t_N)>\alpha]|\geqslant\varepsilon]\leqslant\delta$,即根据样本 U_m 计算的估计器 $\hat{Pr}[S(t_1,t_N)>\alpha]$满足 $\Pr[|\hat{Pr}[S(t_1,t_N)>\alpha]-Pr[S(t_1,t_N)>\alpha]|\geqslant\varepsilon]\leqslant\delta$.

4.3 基于均衡抽样的近似簇监测算法

为了便于算法描述,本文采用基于簇结构的网络拓扑,但是提出的算法和证明的理论结果适用于一般的信道模型和网络协议下的传感器网络. i 节点在 $t_1\sim t_N$ 时间内的感知数据集合记为 $D_i(t_1,t_N)=\{s_i(t_1),s_i(t_2),\cdots,s_i(t_N)\}$,由$(\varepsilon,\delta)$-近似$(\alpha,\tau)$双阈值监测的定义知,近似$(\alpha,\tau)$双阈值监测问题可如下定义:

输入:

(1)监测阈值 α,概率阈值 τ.

(2)$t_1\sim t_N$ 时间内簇成员节点感知数据集合 $D(t_1,t_N)=\{D_{i1}(t_1,t_N),D_{i2}(t_1,t_N),\cdots,D_{in}(t_1,t_N)\}$.

(3)误差上界 ε,失败概率上界 δ.

对于给定的 $\varepsilon>0,\delta>0$,基于均衡抽样的近似簇监测算法的具体步骤如下所示.

(1)根据给定的 ε、δ,簇头根据定理 4 计算优化的样本容量 $m=(\delta\varepsilon^2)^{-1}$并随机独立不重复地产生 m 个 $1\sim N$ 范围内的自然数.

(2)簇头将 m 个随机数广播至簇内成员节点.

(3)成员节点接收到随机数序列后,该节点将其对应时刻的感知数据发送至簇头节点.

(4)簇头接收到成员节点发送的样本数据后,簇头计算感知数据聚集和大于阈值 α 的个数$|\alpha(U^m)|$.

(5)若$|\alpha(U^m)|>\tau m$,则簇头节点向 Sink 节点发送警报.

4.4 算法性能分析

计算复杂度:在算法的第 1 步中,计算优化的样本

容量的复杂度为 $O(1)$,簇头节点产生 m 个随机数的计算复杂度为 $O(m)$;在算法的第 4 步中,簇头节点分别计算 m 个时刻感知数据聚集和的计算复杂度为 $O(nm)$.对于给定的 $\varepsilon>0$、$\delta>0$,由于 $m=O((\delta\varepsilon^2)^{-1})$,因此该算法的计算复杂度为 $O(n\delta^{-1}\varepsilon^{-2})$,其中 n 为簇成员节点数量.

通信复杂度:簇头节点广播随机数和收集样本数据的通信复杂度为 $O(nm)$;簇成员节点将感知数据发送至簇头节点的通信复杂度为 $O(nm)$.因此该算法的通信复杂度为 $O(nm)$.由于 $m=O((\delta\varepsilon^2)^{-1})$,因此该算法的通信复杂度为 $O(n\delta^{-1}\varepsilon^{-2})$.

5 实验结果

感知数据来源于海量数据计算研究中心的真实传感器网络实测.第一组实验考查 $\mathrm{E}[s(t)]<\beta(\beta=40℃)$ 时,分布式监测算法的有效性.图 1 显示了温度感知值及其期望值的对比结果.由实验结果可知,期望值有效地降低了因环境噪声和仪器误差对感知值的影响.如图 1 所示,因外界干扰或感知硬件的影响,传感器节点在第 6 时刻的感知值是 39.68℃,与实际不符.图 1 显示出的实验结果亦辅证了将感知数据视为不确定数据能更好地反映所监测物理变量的真实情况.此外,双阈值监测算法和朴素监测算法进行了对比实验.朴素监测算法的思想为:节点发送警报,当且仅当 $s(t)>\beta$.根据图 1 的温度期望值,图 2 给出了根据定理 1 计算的 $\Pr[s(t)>40]$的上界.对比实验结果如图 3 所示:o 代表运行双阈值监测算法的结果,$*$ 代表运行朴素监测算法结果.由图 2 所示,$\Pr[s(t)>40]$上界小于概率阈值 0.96,因此节点不必发送警报.由图 3 可知,朴素监测算法向用户发送了 4 个误报的警报.实验结果说明双阈值监测算法降低了因环境噪声和仪器误差对节点感知值的影响,降低了警报的误报率和漏报率.

图 4 显示了优化的样本容量与 ε、δ 的关系.当 $\varepsilon=0.2$,$\delta=0.15$ 时,抽样算法仅需收集 167 个时刻的感知数据就使得 $\hat{Pr}[S(t_1,t_N)>\alpha]$与精确值 $\Pr[S(t_1,t_N)>\alpha]$之间的绝对误差大于 0.2 的概率小于 0.15.由于样本容量与 N 无关,因此基于均衡抽样的簇监测算法适用于高频数据采集的传感器网络.

第二组实验考查基于抽样的近似簇监测算法的精度与抽样比例的关系.由图 5 可知,近似簇监测算法仅需少量的样本就能输出高精度的近似 $Pr[S(t_1,t_N)>\alpha]$.例如当抽样比例为 16% 时,近似 $\hat{Pr}[S(t_1,t_{800})>\alpha_1]=0.16$ 与精确 $Pr[S(t_1,t_{800})>\alpha_1]=0.16$ 之间的绝对误差为 0.053.图 6 给出了感知数据规模不同时,近似簇监测算法的精度与抽样比例的关系.由图 6 所示,随抽样比例的增加,近似值与估计值的绝对误差下降得十分明显.

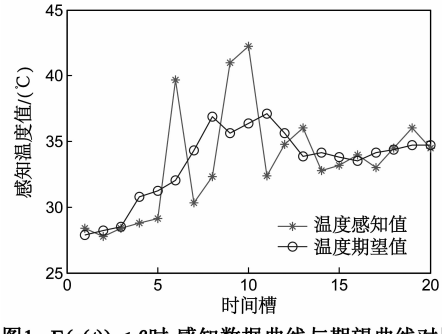


图1 $\mathrm{E}(s(t))<\beta$时,感知数据曲线与期望曲线对比

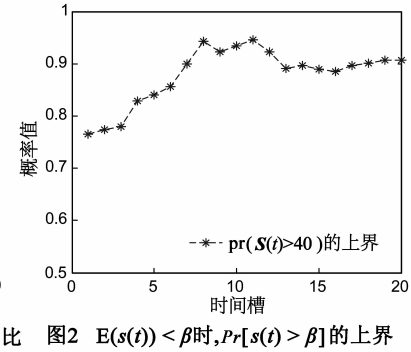


图2 $\mathrm{E}(s(t))<\beta$时,$Pr[s(t)>\beta]$的上界

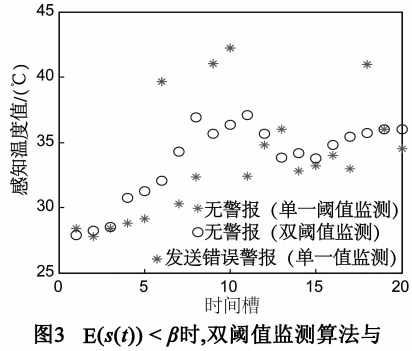


图3 $\mathrm{E}(s(t))<\beta$时,双阈值监测算法与朴素算法的对比

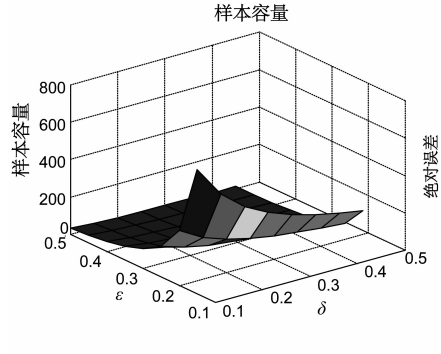


图4 样本容量与ε、δ的关系

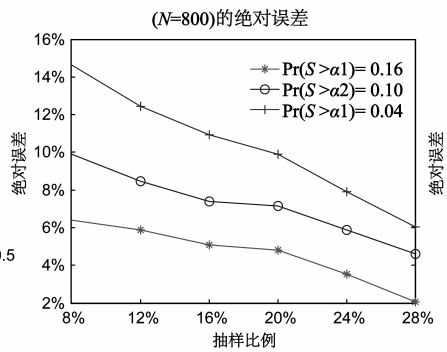


图5 N= 800时,算法精度与抽样比例的关系

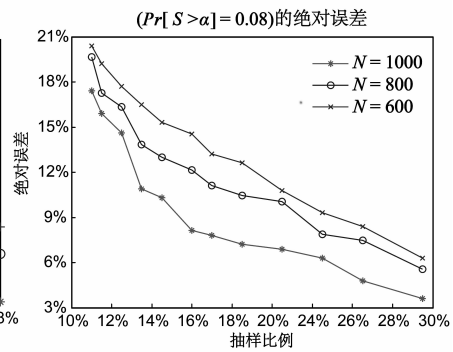


图6 $Pr[S>\alpha]=0.08$时,算法精度与抽样比例的关系

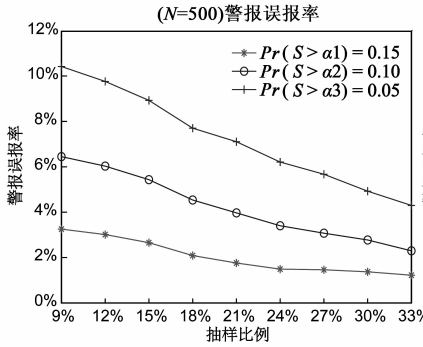


图7　$N=500$时,警报误报率与抽样比例的关系

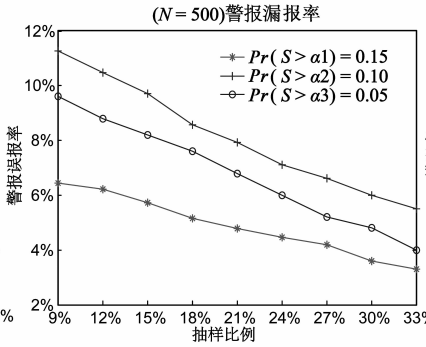


图8　$N=500$时,警报漏报率与抽样比例的关系

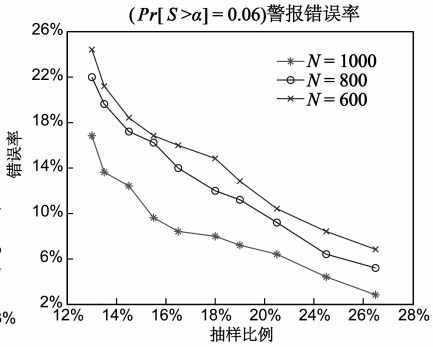


图9　$Pr[S>\alpha]=0.06$时,警报错误率与抽样比例的关系

第三组实验考查基于抽样的近似簇监测算法的警报错误率与抽样比例的关系,实验结果如图 7~9 所示.警报错误率为警报误报率与警报漏报率之和,其中警报误报(或漏报)率的计算公式如下.

$$\text{警报误报/或漏报率}=\frac{\#\text{误报/或漏报的警报}}{\#\text{正确的警报}}$$

如图 7~8 所示,近似簇监测算法仅需少量的样本就可以保证较低的误报率和漏报率.例如当样本比例为 15% 时,误报率小于 0.06.图 8 给出了漏报率随样本容量增加的变化情况.图 9 显示了感知数据规模不同时,警报的错误率与样本比例的关系.如图 9 所示,感知数据规模较大时,警报的错误率下降得十分明显.第二组和第三组实验说明,近似簇监测算法仅需少量的样本就能输出比较精确的近似 $Pr[S(t_1,t_N)>\alpha]$结果并且可以保证较低的警报误报率,因此基于抽样的近似监测算法可以有效地降低通信开销延长网络寿命,适用于高频数据采集的传感器网络.

6　结论

受节点硬件和环境噪声的影响,基于单阈值的监测方法降低了警报的准确率.鉴于上述原因,本文开展了基于双阈值的监测问题的研究.首先,本文给出了$(\beta,\tau)_t$-双阈值监测的定义及其概率语义,给出了计算尾部概率上界的数学方法并设计了分布式监测算法.其次,给出了近似(α,τ)-双阈值监测的定义,证明了算法中需要的数学结果,介绍了基于抽样的近似簇监测算法.最后,通过实验分析了各参数对监测算法的影响,验证了提出的算法的高效性.在未来的工作中,我们考虑概率值估计的其他方法,开展基于(α,τ)-监测的最优阈值设置方法以及基于非均衡抽样的近似簇监测算法的研究.

参考文献

[1] 梁俊斌,王建新,陈建二.在传感器网络中构造延迟限定的最大化生命周期树[J].电子学报,2010,38(2):345-351.
Liang J, Wang J, Chen J. On the construction of a delay-constrained maximum lifetime tree in wireless sensor networks[J]. Acta Electronica Sinica, 2010, 38(2): 345-351. (in Chinese)

[2] 奎晓燕,杜华坤,梁俊斌.无线传感器网络中一种能量均衡的基于连通支配集的数据收集算法[J].电子学报,2013,41(8):1512-1528.
Kui X, Du H, Liang J. An energy-balanced connected dominating sets for data gathering in wireless sensor networks[J]. Acta Electronica Sinica, 2013, 41(8): 1521-1528. (in Chinese)

[3] 陈零,王建新,张士庚,奎晓燕.无线传感器网络中基于树的能量高效分布式精确数据收集算法[J].电子学报,2013,41(9):1738-1743.
Chen L, Wang J, Zhang S, et al. A distributed tree-based energy-efficient algorithm for precise data gathering in wireless sensor networks[J]. Acta Electronica Sinica, 2013, 41(9): 1738-1743. (in Chinese)

[4] Noury N, Hervé T, Rialle V, et al. Monitoring behavior in home using a smart fall sensor and position sensors[A]. 2000 1st Annual International, Conference On Microtechnologies in Medicine and Biology[C]. Piscataway, NJ: IEEE, 2000. 607-610.

[5] Agrawal S, Deb S, Naidu K V M, et al. Efficient detection of distributed constraint violations[A]. 2007 23rd IEEE International Conference on Data Engineering[C]. Piscataway, NJ: IEEE, 2007. 1320-1324.

[6] Sharfman I, Schuster A, Keren D. A geometric approach to monitoring threshold functions over distributed data streams [J]. ACM Transactions on Database Systems, 2007, 32(4): 23-34.

[7] Kashyap S, Ramamirtham J, Rastogi R, et al. Efficient constraint monitoring using adaptive thresholds[A]. 2008 24th International Conference on Data Engineering[C]. Piscataway, NJ: IEEE, 2008. 526-535.

[8] Dilman M, Raz D. Efficient reactive monitoring[J]. IEEE Journal on Selected Areas in Communications, 2002, 20(4): 668-676.

[9] Cormode G, Muthukrishnan S, Yi K. Algorithms for distributed functional monitoring [J]. ACM Transactions on Algorithms, 2011, 7(2): 21 – 40.

[10] Cheng R, Kalashnikov D V, Prabhakar S. Evaluating probabilistic queries over imprecise data [A]. Proceedings of the 2003 ACM SIGMOD International Conference on Management of data [C]. New York: ACM, 2003. 551 – 562.

[11] Deshpande A, Guestrin C, Madden S R, et al. Model-driven data acquisition in sensor networks [A]. Proceedings of the 30th International Conference on Very Large Data bases [C]. San Francisco: Morgan Kaufmann, 2004. 588 – 599.

[12] Gruenwald L, Chok H, Aboukhamis M. Using data mining to estimate missing sensor data [A]. 2007 7th IEEE International Conference on Data Mining Workshops [C]. Piscataway, NJ: IEEE, 2007. 207 – 212.

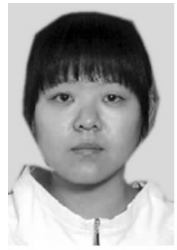
[13] Suciu D, Olteanu D, Ré C, et al. Probabilistic databases [J]. Synthesis Lectures on Data Management, 2011, 3(2): 1 – 180.

[14] Tang M, Li F, Phillips J M, et al. Efficient threshold monitoring for distributed probabilistic data [A]. 2012 28th IEEE International Conference on Data Engineering [C]. Piscataway, NJ: IEEE, 2012. 1120 – 1131.

[15] Keralapura R, Cormode G, Ramamirtham J. Communication – efficient distributed monitoring of thresholded counts [A]. Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data [C]. New York: ACM, 2006. 289 – 300.

[16] Meng S, Wang T, Liu L. Monitoring continuous state violation in datacenters: Exploring the time dimension [A]. 2010 26th IEEE International Conference on Data Engineering [C]. Piscataway, NJ: IEEE, 2010. 968 – 979.

作者简介



毕 冉 女, 1985 年出生, 博士研究生, 主要研究领域为海量数据计算, CPS, 无线传感器网络.
E-mail: biranhit@gmail.com

李建中(通信作者) 男, 1950 年出生, 教授、博士导师, 主要研究领域为海量数据计算, CPS, 无线传感器网络, 数据质量, 数据挖掘.
E-mail: lijzh@hit.edu.cn

高 宏 女, 1966 年出生, 教授、博士生导师, 主要研究领域为海量数据计算, CPS, 无线传感器网, 数据质量, 数据挖掘.