基于线性回归的无线传感器网络分布式数据采集优化策略

宋 欣[1),2)] 王翠荣[1)]

[1)](东北大学信息科学与工程学院 沈阳 110819)

[2)](中国科学院自动化研究所复杂系统与智能科学重点实验室 北京 100190)

摘 要 事件监测是无线传感器网络中最重要的应用之一,部署在监测区域内的传感器节点通过对感知数据信息的采集、处理和传输等基本操作完成具体的监测任务,在各种操作中,节点之间的数据传输是最消耗能量的.为了减少节点之间的通信数据量,达到降低网络能耗和延长网络生命周期的目的,该文提出了一种能量高效的基于线性回归的无线传感器网络分布式数据采集优化策略,通过应用线性回归分析方法构建感知数据模型,保持感知数据的特征,使节点仅传输回归模型的参数信息,代替传输实际监测的感知数据信息.仿真实验结果表明,文中提出的数据采集优化策略能通过较小的通信量有效地实现事件监测区域感知数据的预测和估计,降低网络的总能量消耗,延长网络的生命周期.

关键词 无线传感器网络;数据采集;线性回归;能量高效;优化策略;绿色计算;物联网

中图法分类号 TP393 **DOI号**: 10.3724/SP.J.1016.2012.00568

Linear Regression Based Distributed Data Gathering Optimization Strategy for Wireless Sensor Networks

SONG Xin[1),2)] WANG Cui-Rong[1)]

[1)](*School of Information Science and Engineering, Northeastern University, Shenyang* 110819)

[2)](*Key Laboratory of Complex System and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

Abstract Event detection is one of the most critical applications in wireless sensor networks (WSN). For implementing the event detection task, the sensor nodes deployed in monitoring region are capable of gathering data, processing data, transmitting data to sink node and so on. For such sensors, transmission is much more energy consuming than computation. Therefore, the amount of sensory data communication overhead should be kept as low as possible, in order to prolong the lifetime of wireless sensor networks and reduce energy consumption. In this paper, an energy-efficient linear-regression-based distributed data gathering optimization strategy is proposed. The linear regression model can accurately represent the feature of the original monitoring data. Rather than transmitting measurements to another node, nodes transfer constraints on the model parameters, drastically reducing the communication required. The theoretical analysis and experimental results show that the proposed strategy is able to implement measurements prediction and estimate with lower communication cost. The designed algorithm achieves more energy savings and extends the wireless sensor networks lifetime.

Keywords wireless sensor networks; data gathering; linear regression; energy efficiency; optimization strategy; green computing; Internet of Things

收稿日期:2011-08-31;最终修改稿收到日期:2012-01-03.本课题得到国家自然科学基金(61070162,71071028)、中国科学院开放课题(20100106)资助.**宋 欣**,女,1978年生,博士研究生,讲师,主要研究方向为无线传感器网络、分布式数据处理、智能信息处理.E-mail:bravesong@163.com.**王翠荣**,女,1963年生,博士,教授,主要研究领域为下一代网络技术、无线传感器网络、数据中心.

1 引　言

无线传感器网络(Wireless Sensor Networks, WSN)的事件监测应用依赖于部署在事件区域内的资源有限的大量无线传感器节点对监测环境数据信息的采集、处理和上传到基站等一系列操作完成.执行监测任务时,传感器网络执行长时间的数据采集和查询任务,并要求将监测结果传送到基站,以供进一步的分析和决策.但是,对于监测系统来说,如果单纯地通过传感器节点的数据采集、传输操作获得环境区域内较完整的感知信息结构是不现实的,对于传感器的感知时间间隔设置是没有标准的.时间设置太长,则丢失重要数据信息的概率增大;设置太短,又会产生巨大的网络能耗和数据冗余,造成网络过早失效.例如,传感器节点以一定的时间间隔采集监测环境范围内的温度信息,这种情况下,采集的信息量可能非常大.由于传感器节点在能量、存储空间、计算能力以及无线传感器网络本身负载能力等方面的限制,数据采集和传输将带来巨大的网络流量压力,节点能量消耗大(尤其是节点发送和接收数据时的能量消耗),采集数据存在冗余,为监测系统进一步分析和处理感知数据信息,增加了不必要的时间、空间复杂度.在 Mobicom 2002 会议上,Deborah Estrin 在特邀报告中指出,传感器节点传输 1bit 信息 100 m 距离需要的能量大约相当于执行 3000 条计算指令消耗的能量.图 1 所示为传感器节点在各工作状态下的功耗情况,从中可以看出,节点的无线通信模块产生的功耗远大于节点感知和计算的功耗[1].另外,文献[2]中比较了通信和计算两种工作模式中 Mica2dot 节点的功耗,发现传输 1 bit 数据大约相当于节点微控制器运行 2090 次时钟周期,证明了节点的计算能耗远小于通信能耗.也就是说,影响无线传感器网络总能耗的主要因素是网络中的通信能耗.所以,为了减少频繁的数据通信,无线传感器网络采取了网内数据聚合策略,即节点在发送感知数据之前对数据进行必要的处理,从而减少数据的发送量,达到节约能量的目的.但是,如果仅仅采取一些简单的本地统计运算(如求平均值、最大值、最小值或直方图等),虽然在原有数据集的基础上根据应用的需求提取了部分代表性数据,减少了一定量的数据转发操作,却失去了数据本身的结构特征信息,监测系统分析的仅仅是比较粗糙的数据集.所以,寻找一种能够更平滑的表示较完整感知数据结构特征的、能较大程度地减少节点间通信量的数据采集策略是非常重要的.

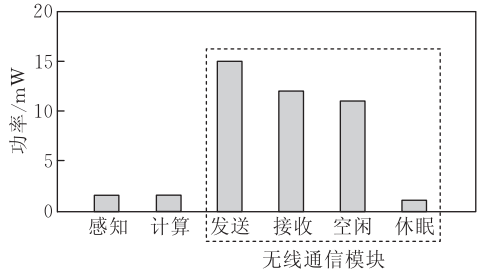

图 1　节点在各工作状态下的功耗

本文在分析已有重要研究成果的基础上,提出了一种能量高效的分布式线性回归数据采集优化策略,根据同一监测区域附近的传感器节点采集的数据测量值具有极大相关性这一特点,通过构建线性回归模型表示传感器的感知数据.这样在数据测量值抽取和表示时,节点只需要传输模型中定义的基函数的参数信息,就可以得到原始数据信息有效的和比较完整的结构表示,大大减少了传感器节点间频繁数据传输带来的通信开销,降低了节点的能量消耗,从而达到了延长网络生命周期的目的.

本文在第 2 节介绍已有数据传输优化策略的重要研究成果;第 3 节讨论通过线性回归模型进行数据采集优化的主要思想和具体实现;第 4 节通过实验,测试和分析算法的有效性以及对网络能耗的影响;最后给出本文的结论和进一步的研究设想.

2 相关工作

在网络应用规模不断扩大的今天,网络能耗一直是研究人员非常重视的问题.无论是在有线网络还是在无线网络的应用中,构建节能的绿色网络框架和行之有效的优化评价策略都关系着社会的可持续发展问题[3].在社会信息化、智能化过程中,无线传感器网络的应用逐渐扩大,但受到传感器节点自身资源(包括能量、存储、计算等)限制,设计能量高效的无线传感器网络系统仍然非常重要.特别是当无线传感器网络部署在危险或者人为很难更换电源的监测区域时,如何减少能耗问题的研究更是其他关键技术的前提和根本原则.通过最小化通信负载实现能量高效的事件监测应用已经取得了一些成果.一类常用的方法是在节点分簇基础上,利用 Slepian-Wolf 界限编码等分布式信源编码技术实现感知数据信息的压缩,优化簇内信息速率分配问题,以达到最小化通信能耗的目的[4-7].虽然通过分簇,簇

头节点承担了数据处理和上传到基站的任务，但是对于簇头节点的数据传输时间间隔，历史数据对后续测量数据的相关性等问题研究还不足. 文献[8]综合了预测模型和分簇技术的优越性，提出了一种基于集成自适应预测模型的无线传感器网络分层数据采集框架. 在该框架中，簇头实现簇内节点的数据采集和有效的预测分析，根据网络状态和性能分析结果平衡通信能耗和预测计算能耗，自适应选择是否执行预测模型，实现了能量高效的无线传感器网络网内数据聚合处理. 考虑到感知数据的时间相关性，文献[9]中利用数据预测传送机制提出了能量均衡的无线传感器网络数据汇聚协议. 如果从节点调度方面考虑，通过对节点活动时间槽的分配和优化，利用随机 Petri 网模型和设计合理的调度策略，寻找更符合应用需求的节点睡眠-唤醒机制[10-16]. 还有一些研究成果利用移动代理技术进行聚合分析，以减少数据冗余和节点能耗，由移动代理保存局部数据聚合结果迁移到下一个传感器节点继续执行聚合操作[17-18]. 另外，在特殊的应用环境中，也可以用随机投递的方式进行数据传输，这种方式牺牲了部分传输可靠性，利用带吸收态的有限状态马尔可夫链模型对规定期望投递概率的随机投递协议进行分析，对减少网络能量消耗同样有较佳的表现[19]. 节点利用自身计算能力构建预测模型，对感知数据进行逼近估计，也可以减少数据传输量[20]. 由于数据在时间和空间上具有相关性，而且计算的能量消耗远远小于节点通信的能量消耗，利用统计分析、维数约简、数据压缩等方法对数据进行抽取和表示，使得节点无需将环境感知数据逐一的传送到基站，这样不仅节省了节点能量，还可以较好地保持感知数据的结构特征，而且不影响决策分析[21-24]. 但是，无论采用何种算法和策略对感知数据进行聚合处理和传输控制，都应该将算法的可扩展性、简单高效性、是否能通过增量更新等作为重点解决的问题，因为这些问题关系着算法在实际应用中的可行性. 目前，利用统计分析方法构建预测模型对感知数据进行控制和聚合处理的成果还很少，研究适用于传感器网络应用的、简单高效的统计分析和预测模型是非常有价值的. 回归分析是确定两种或者两种以上变量间相互依赖的定量关系的统计分析方法，由于在事件监测传感器网络应用中感知数据在一定程度上具有时间和空间的相关性，为了减少大量数据传输带来的网络能耗，本文利用回归分析方法根据实测历史数据建立了线性回归模型，通过求解模型对应基函数的各个参数. 节点仅仅传输相关参数向量，减少了冗余数据，而且模型可通过简单增量更新的方法接收新的实际测量值，基函数的参数也可以在任何时间通过构建的线性回归模型进行求解.

3 基于线性回归的 WSN 分布式数据采集优化策略

无线传感器网络由一系列带有无线射频发射模块和数据感知模块的无线传感器节点组成，节点能实现短距离无线通信和应用环境信息的感知、存储、融合等操作. 本节详细论述了无线传感器网络中能量高效的分布式线性回归数据采集优化策略，有效的减少了网内数据的传输量，节省了节点的能量.

3.1 研究动机

无线传感器网络应用中，每个节点将产生大量感知数据，例如，系统中如果设置节点每1 min采集 1 次环境温度信息，则每小时将产生 60 个感知数据，而一天不间断的开启监测系统，一个节点就将产生 1440 个感知信息，那么对于较大规模部署的传感器网络(节点数量在 100 以上)，要求网络生命周期在 2 年左右，可想而知网络的数据量将非常大，这还只是考虑单一监控指标的情况. 如果是多维数据监测(包括湿度、光照强度、气压等)，那么数据量的剧增将导致节点能量耗尽，无法完成通信. 对于一般的农业监测而言，网络规模不大，而且对于短时间的温度变化要求不高，采用延长节点采样时间(如 10 min 或者 30 min 采样 1 次)，可以减少冗余的数据量，但是对于森林防火等实时监测应用而言，30 min 进行 1 次温度采样未免时间过长，一旦有火灾等险情发生，节点感知响应较慢，失去了实时预防和报警的功能. 所以，在无线传感器网络的应用中如果没有良好的数据处理和传输优化机制，将影响网络生命周期. 一种有效的方法是构建网内感知数据模型，使得节点间通过必要的模型参数传递，监测系统就能有效地提取数据而不失数据的一般特征，既能减少冗余的感知数据，又不影响系统的决策分析.

3.2 传感器网络分布式线性回归模型

可以根据网络的具体应用环境和传感器节点的存储空间、处理能力等性能指标，选取传感器节点一定时间间隔内最近的 m 个感知数据，假设为 $(t_1, y_1), (t_2, y_2), \cdots, (t_m, y_m)$，其中 t_i 和 $y_i (i \in [1, m])$ 表示采样时间点和可受到测量误差影响的实测值，对于这 m 个感知数据构建函数 $Y(t)$，满足逼近误差

$\delta_i = Y(t_i) - y_i$是很小的(在监测系统采集数据的置信区间范围内). 函数 $Y(t)$的形式依赖于具体问题,在此,可以将 $Y(t)$表示为等式(1)的形式:

$$Y(t) = \sum_{j}^{n} \lambda_j B_j(t) \tag{1}$$

其中和项的个数 n 和特定的基函数 B_j 取决于实际的问题,一般情况下,选择的基函数可以为 $B_j(t) = t^{j-1}$,则等式(1)可以表示为 t 的 $n-1$ 次多项式,即:

$$Y(t) = \lambda_1 + \lambda_2 t + \lambda_3 t^2 + \cdots + \lambda_n t^{n-1} \tag{2}$$

选择 $n = m$ 可以确切地计算出对应的 y_i 值,但计算高阶函数 Y 容易对数据产生干扰,对未预见的 t 预测其相应的 y 值时,将影响其精确性. 较好的做法是选择一个远小于 m 的 n 值,即 $n \ll m$,通过选择系数 λ_i 的值,获得函数 Y 对应于测量值 y 的估计值. 在无线传感器网络应用中,假设选择节点最近采集的 50 个温度测量值,则构建一个三阶多项式函数模型:$Y(t) = \lambda_1 + \lambda_2 t + \lambda_3 t^2 + \lambda_4 t^3$估计测量值 $y_i (i = 1, 2, \cdots, 50)$即可,节点不需要传输 50 个实际测量值,构建函数模型之后,仅仅需要在网内传输 4 个参数值,即 λ_1,λ_2,λ_3和 λ_4作为测量值的压缩表示,所以减少了网内的信息传输量. 若通过线性回归模型求得系数,需要将多项式表示模型转换为矩阵表示,这样,节点不必求解高阶多项式的解,只需维护相关矩阵即可. 设所求的系数 n 维向量为 $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_n)^{\mathrm{T}}$,实际测量值的 m 维向量为 $\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^{\mathrm{T}}$,相应时间采样点 t_i 的基函数矩阵

$$\boldsymbol{M} = \begin{bmatrix} B_1(t_1) & B_2(t_1) & \cdots & B_n(t_1) \\ B_1(t_2) & B_2(t_2) & \cdots & B_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(t_m) & B_2(t_m) & \cdots & B_n(t_m) \end{bmatrix},$$

其中矩阵元素 $m_{ij} = B_j(t_i)$,则等式(1)在 t_i 采样时间点的预测函数 m 维向量 $\boldsymbol{Y} = (Y(t_1), Y(t_2), \cdots, Y(t_m))^{\mathrm{T}}$表示为等式(3),即

$$\boldsymbol{Y} = \begin{bmatrix} Y(t_1) \\ Y(t_2) \\ \vdots \\ Y(t_m) \end{bmatrix} = \boldsymbol{M}\boldsymbol{\lambda} = \begin{bmatrix} B_1(t_1) & B_2(t_1) & \cdots & B_n(t_1) \\ B_1(t_2) & B_2(t_2) & \cdots & B_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(t_m) & B_2(t_m) & \cdots & B_n(t_m) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} \tag{3}$$

则逼近误差向量 $\boldsymbol{\delta}$ 可表示为等式(4),即

$$\boldsymbol{\delta} = \boldsymbol{M}\boldsymbol{\lambda} - \boldsymbol{y} \tag{4}$$

为了使估计值的逼近误差 δ 最小,选定使得逼近误差向量 $\boldsymbol{\delta}$ 的范数最小为优化目标,即可以得到

$$\min\left(\|\boldsymbol{\delta}\| = \left(\sum_{i=1}^{m} \delta_i^2 \right)^{1/2} \right) \tag{5}$$

结合等式(4)和优化目标式(5)可得

$$\min\left(\|\boldsymbol{\delta}\|^2 = \|\boldsymbol{M}\boldsymbol{\lambda} - \boldsymbol{y}\|^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} m_{ij}\lambda_j - y_i \right)^2 \right) \tag{6}$$

可以通过对每一个 $\lambda_k (k = 1, 2, \cdots, n)$求 $\|\boldsymbol{\delta}\|^2$ 的微分并令结果为 0,以求得 $\|\boldsymbol{\delta}\|$ 的最小值,即

$$\frac{\mathrm{d}\|\boldsymbol{\delta}\|^2}{\mathrm{d}\lambda_k} = \sum_{i=1}^{m} 2\left(\sum_{j=1}^{n} m_{ij}\lambda_j - y_i \right) m_{ik} = 0, k = [1, n] \tag{7}$$

根据等式(4)可推导出如下与等式(7)等价的矩阵方程,即

$$(\boldsymbol{M}\boldsymbol{\lambda} - \boldsymbol{y})^{\mathrm{T}}\boldsymbol{M} = \boldsymbol{0} \tag{8}$$

$$\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{M}\boldsymbol{\lambda} - \boldsymbol{y}) = \boldsymbol{0} \tag{9}$$

$$\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\lambda} = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{y} \tag{10}$$

因为定义的基函数为 $B_j(t) = t^{j-1}$,所以基函数矩阵 $\boldsymbol{M}$ 为列满秩矩阵,对任意列满秩矩阵 $\boldsymbol{M}$,可以得到 $\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}$ 是正定的,所以 $(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})^{-1}$ 存在,根据等式(10)得到系数向量 $\boldsymbol{\lambda}$ 的解为

$$\boldsymbol{\lambda} = (\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})^{-1}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{y} \tag{11}$$

令

$$\boldsymbol{A} = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{M} = \begin{bmatrix} \langle B_1 \cdot B_1 \rangle & \langle B_1 \cdot B_2 \rangle & \cdots & \langle B_1 \cdot B_n \rangle \\ \langle B_2 \cdot B_1 \rangle & \langle B_2 \cdot B_2 \rangle & \cdots & \langle B_2 \cdot B_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle B_n \cdot B_1 \rangle & \langle B_n \cdot B_2 \rangle & \cdots & \langle B_n \cdot B_n \rangle \end{bmatrix} \tag{12}$$

$$\boldsymbol{z} = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{y} = \begin{bmatrix} \langle B_1 \boldsymbol{y} \rangle \\ \langle B_2 \boldsymbol{y} \rangle \\ \vdots \\ \langle B_n \boldsymbol{y} \rangle \end{bmatrix} \tag{13}$$

则根据等式(12)、(13),等式(11)可写成 $\boldsymbol{\lambda} = \boldsymbol{A}^{-1}\boldsymbol{z}$,即

$$\boldsymbol{A}\boldsymbol{\lambda} = \boldsymbol{z} \tag{14}$$

其中:$\boldsymbol{A}$ 是基函数的数量积矩阵;$\boldsymbol{z}$ 是测量值向量的基函数投影. 至此,已知实测值和基函数,就可以通过求解等式(14)这一典型的线性系统,得到优化回归系数.

3.3　回归模型的参数更新方法

无线传感器网络事件监测应用中,随着监测时间延长,传感器节点采集的环境数据量也在不断增加,由于受到传感器节点本身能量、存储和处理能力

的限制，节点只能存储一定时间段内的采样数据，当利用线性回归模型计算数据表示系数时，模型的更新操作，可采取如下增量方式计算.

假设已经计算得到 t_1 到 t_{m-1} 采样时间段内基函数数量积矩阵 $\boldsymbol{A}$ 和投影向量 $\boldsymbol{z}$，那么对于在 t_m 时间的新测量值有

$$\boldsymbol{A}(t_m)=\begin{bmatrix}\langle B_1(t_m)\cdot B_1(t_m)\rangle & \cdots & \langle B_1(t_m)\cdot B_n(t_m)\rangle\\ \langle B_2(t_m)\cdot B_1(t_m)\rangle & \cdots & \langle B_2(t_m)\cdot B_n(t_m)\rangle\\ \vdots & \vdots & \vdots\\ \langle B_n(t_m)\cdot B_1(t_m)\rangle & \cdots & \langle B_n(t_m)\cdot B_n(t_m)\rangle\end{bmatrix},$$

$$\boldsymbol{z}(t_m)=\begin{bmatrix}\langle B_1(t_m)\boldsymbol{y}(t_m)\rangle\\ \langle B_2(t_m)\boldsymbol{y}(t_m)\rangle\\ \vdots\\ \langle B_n(t_m)\boldsymbol{y}(t_m)\rangle\end{bmatrix},$$

则新的采样时间段内的基函数数量积矩阵 $\boldsymbol{A}$ 和投影向量 $\boldsymbol{z}$ 为

$$\boldsymbol{A}\leftarrow\boldsymbol{A}+\boldsymbol{A}(t_m);\boldsymbol{z}\leftarrow\boldsymbol{z}+\boldsymbol{z}(t_m)\qquad(15)$$

对于矩阵 $\boldsymbol{A}$ 和向量 $\boldsymbol{z}$ 的规模控制采用了滑动窗口机制，系统综合节点的计算、存储能力以及应用要求设置滑动窗口大小，随着矩阵 $\boldsymbol{A}$ 和向量 $\boldsymbol{z}$ 规模的不断扩大，当 t_1 时间的数据超出滑动窗口的设置后，可根据式(16)计算更新后的 $\boldsymbol{A}$ 和 $\boldsymbol{z}$.

$$\boldsymbol{A}\leftarrow\boldsymbol{A}-\boldsymbol{A}(t_1);\boldsymbol{z}\leftarrow\boldsymbol{z}-\boldsymbol{z}(t_1)\qquad(16)$$

综上，节点可通过计算线性系统 $\boldsymbol{A\lambda}=\boldsymbol{z}$ 提取回归系数，而且可以采用增量方式更新线性回归系统模型的矩阵和向量参数.

3.4 网络模型

本文假设无线传感器网络由随机均匀散布在一个 $K\times K$ 的二维正方形区域中的 N 个无线传感器节点组成，网络具有如下性质：

(1) 节点部署之后就不再移动，静止不动地获取监测事件采样数据.

(2) 系统中设置唯一的部署在固定位置的基站(BS)，具有较强的通信和计算能力，并且能量不受限制.

(3) 节点不能获知各自的位置信息，需要借助GPS、定位算法等才能获取.

(4) 节点通信具备各向同性传播模式.

(5) 节点的无线发射功率是可控的，节点能通过接收者的距离调整传输功率.

(6) 节点能通过接收信号强度和给定的传输功率估计节点间的近似距离.

(7) 为了简化理论分析，节点采用相同的无线通信模块，所有无线通信信道是对称的，即在给定信噪比的情况下，无论是从 i 节点到 j 节点，还是从 j 节点到 i 节点传输 m 比特的信息产生的能量消耗是一致的.

(8) Sink 节点(即汇聚点)通过基于分簇的无线传感器网络收集感知数据. 采用分布式的簇形成算法已经被证明能有效地节省网络能量和提高网络的可扩展性.

(9) 为平衡簇头节点的能量负载，簇头尽可能均匀分布在整个网络中，设置节点通信半径 R_{node} 大于 2 倍的簇半径 R_{cluster}，保证了被选举出的簇头节点之间能直接通信.

系统根据合理的睡眠调度机制，簇内节点采集监测环境中的感知信息，并传输采样数据到簇头节点(CH)，簇头节点构建线性回归模型估计采样数据，并根据查询统计需求上传表示数据特征的模型参数到基站. 簇头节点比对线性回归预测模型计算出的数据与实际采集数据的误差. 如果没有超出阈值 T_{error}，则不更新回归模型，如果超出 T_{error} 则通过容错策略判断是否更新回归模型，重新计算参数. 分布式数据采集如过程 1 所示，基于分簇的无线传感器网络数据采集逻辑结构如图 2 所示.

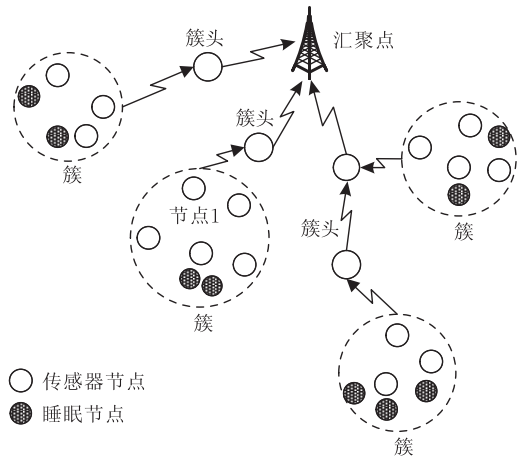

图 2 基于分簇的无线传感器网络数据采集逻辑结构

过程 1. 分布式数据采集过程.

1. 执行 initialize()函数，初始化节点各参数

2. 簇结构建立

if ($alive\_node$=1)//如果节点存活

broadcast($nodei$,$energy\_currenti$)

//广播节点 ID 号和当前能量

Distributed_cluster_formation();//执行簇建立函数

3. 簇头接收数据，执行回归模型，上传模型参数到 Sink 节点

for each CH node do

{wait for receiving messages;

if(message from BS) //BS 为基站

update processing parameters;

else

if(message from in_cluster node)

{gather sampling data and predict data;

```
        sensor_Distri_Regression();
        if (error>threshold) fault_tolerant();
        forwarding the model parameters to next CH;}
    else
    farwarding data to next CH;
}
4. 簇内节点执行采样任务并上传到簇头节点
for each in_cluster node do
{if(node_state=sleep)
    wait for wake_up messages;
    else
      if(energy_currenti=1)
      sendmessage(nodei,CHi);
      else
      energy_currenti=0;
}
5. 基站接收簇头节点发送的模型参数信息
for BS node do
{ while(1)
{ wait for data from CHi;
  if(received data ratio>threshold)
  {send parameter to CHi;
  continue;}
  else
  save received data;
}
```

3.5　分布式数据处理过程

假设簇内 $node1$ 节点在 t_1 到 t_6 时刻采集的温度信息 y_1 到 y_6 的向量表示为 $(t_1,y_1),(t_2,y_2),\cdots,(t_6,y_6)$，对应的感知数据为 (1,26.7)，(2,28.6)，(3,27.3)，(4,25.8)，(5,29.4)，(6,27.9)，以此为例说明回归模型的建立和参数求解过程. 这 6 个数据点在平面上的分布如图 3 所示，三次曲线 S 表示在较理想的历史数据采样规模下，回归模型计算得出的线性回归预测结果. 为简单起见，用三次多项式 $Y(t)=\lambda_1+\lambda_2 t+\lambda_3 t^2+\lambda_4 t^3$ 对这 6 个感知数据进行拟合，首先构造基函数值的矩阵

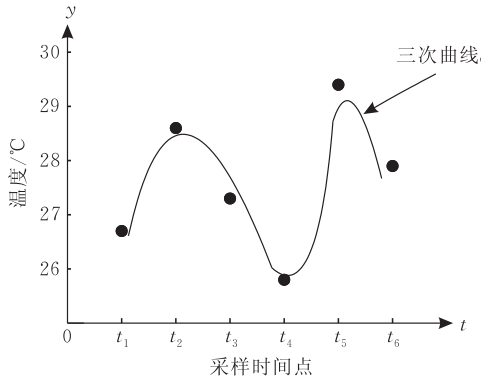

图 3　温度值理想回归曲线

$$\boldsymbol{M}=\begin{bmatrix}1 & t_1 & {t_1}^2 & {t_1}^3\\ 1 & t_2 & {t_2}^2 & {t_2}^3\\ 1 & t_3 & {t_3}^2 & {t_3}^3\\ 1 & t_4 & {t_4}^2 & {t_4}^3\\ 1 & t_5 & {t_5}^2 & {t_5}^3\\ 1 & t_6 & {t_6}^2 & {t_6}^3\end{bmatrix}=\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 2 & 4 & 8\\ 1 & 3 & 9 & 27\\ 1 & 4 & 16 & 64\\ 1 & 5 & 25 & 125\\ 1 & 6 & 36 & 216\end{bmatrix} \tag{17}$$

则根据等式(14)，可以得到回归模型系数向量 $\boldsymbol{\lambda}=(25.8667,1.7507,-0.5758,0.0593)$，所以，由线性回归产生的三阶方程式为

$$Y(t)=25.8667+1.7507t-0.5758t^2+0.0593t^3 \tag{18}$$

仅由 6 个历史采样数据得到的回归模型预测结果如图 4 所示. 得到的 $Y(t)$ 温度值分别是：27.101，27.539，27.537，27.449，27.633 和 28.442. 由于只有 6 个历史数据参与预测，所以回归模型预测的精度不高，误差稍大. 回归模型构建时参与的历史采样数据的多少，可通过计算误差、设置回归模型的置信区间 α 进行控制. 根据实际的应用需要，设置合适的回归模型采样时间窗口，即多少个采样时间点的监测值参与回归估计，在不影响预测精度的情况下，降低算法的时间和空间复杂度.

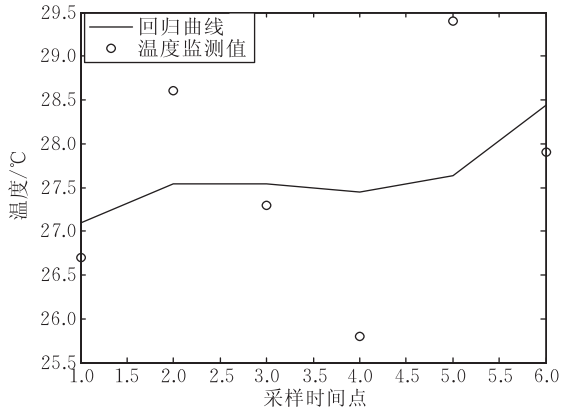

图 4　线性回归曲线(6 个采样点)

传感器网络节点维护基函数的数量积矩阵 $\boldsymbol{A}$ 和测量值向量的基函数投影向量 $\boldsymbol{z}$，线性回归模型参数求解过程如算法 1 所示.

算法 1.　求解线性回归模型参数.

输入：(1) 采样时间滑动窗口大小

Max_Time_W←size of time window;

(2) 估计值置信区间

α←confidence interval;

(3) 消息发送间隔

Message_itv←timer interval to send messages;

(4) 节点所在簇的 ID 号

Cluster_ID←the cluster ID number of sensor nodes;
(5) 初始化基函数矩阵
M←the basis functions matrix;
输出:线性回归模型参数向量
1. 初始化矩阵 $\boldsymbol{A}$,向量 z 和向量 $\boldsymbol{\lambda}$
For(each node i) do
{$\boldsymbol{A}$←0; z←0; $\boldsymbol{\lambda}$←0;}
2. 建立线性回归模型,求解模型参数
For (each node i) do {
if ($T<Max\_Time\_W$)
{$\boldsymbol{A}=\boldsymbol{A}+\boldsymbol{A}(T)$; $z=z+z(T)$; }
else
{ $\boldsymbol{A}=\boldsymbol{A}-\boldsymbol{A}(T1)$; $z=z-z(T1)$;
$\boldsymbol{A}=\boldsymbol{A}+\boldsymbol{A}(T)$; $z=z+z(T)$;}
$T++$;
For (each *Message_itv*)
{ $\boldsymbol{\lambda}=\boldsymbol{A}^{-1}z$;
Send_Message(*CH_ID*, $\boldsymbol{\lambda}$); }
}
3. 簇头节点进行预测控制
For (each CH node j) do
{if ($Y(T)\in\alpha$) *Received_Message*=true;
Else
{*Received_Message*=false;
Send_alarm_Record(*Sink*, *CH_ID*, *Node_ID*);
fault_tolerant( );
}

3.6 算法复杂度分析

无线传感器网络节点执行线性回归模型算法时,在每个消息发送周期,算法主要涉及的运算包括矩阵的加法、减法、乘法和求逆操作.对于 $\boldsymbol{A}$ 矩阵的构成,矩阵的乘法操作可以通过 Strassen 算法将矩阵的乘法运算时间复杂度由 $O(n^3)$ 降低为 $O(n^{2.81})$.对于线性方程 $\boldsymbol{A\lambda}=z$ 的求解,最简单的方法是采用矩阵求逆的方法,但是运用 LUP 分解来求解,更具有数值稳定性和加快求解速度,LUP 分解的基本算法思想是找出 3 个矩阵 $\boldsymbol{L}$、$\boldsymbol{U}$ 和 $\boldsymbol{P}$,满足 $\boldsymbol{PA}=\boldsymbol{LU}$,其中 $\boldsymbol{L}$ 是一个单位下三角矩阵,$\boldsymbol{U}$ 是一个上三角矩阵,$\boldsymbol{P}$ 是一个置换矩阵,根据矩阵的性质可以证明每一个非奇异矩阵 $\boldsymbol{A}$ 都有一个 LUP 分解.由于 $\boldsymbol{A}$ 是对称的正定矩阵,可以根据对称正定矩阵的性质证明矩阵求逆运算并不比矩阵乘法运算更困难.即:如果能在 $T(n)$ 的时间内计算出两个 $n\times n$ 实矩阵的乘积,其中 $T(n)=\Omega(n^2)$ 且 $T(n)$ 满足两个正则条件:对任意的 $0\leq k\leq n$ 有 $T(n+k)=O(T(n))$ 以及对某个常数 $C<1/2$ 有 $T(n/2)\leq C\times T(n)$,则可以在 $O(T(n))$ 时间内求出任何一个 $n\times n$ 非奇异实矩阵的逆矩阵.

为了进一步提高预测精度,如果在路由树中某个中间节点执行回归模型参数运算,那么其它节点之间传输的信息则是矩阵 $\boldsymbol{A}$ 和向量 z,算法的复杂度与回归模型计算时的采样个数(采样时间窗口)设置大小有关.假设传感器网络中采样个数设置规模为 S,在最坏的情况下,节点传输的信息量为 S^2+S(矩阵 $\boldsymbol{A}$ 和向量 z 的信息量).如果网络由 N 个节点组成,节点之间最坏的情况下,互传信息量为 $2N(S^2+S)$.如果节点向 Sink 节点传送模型参数信息,则还需要额外传输 $deep\times num$ 大小的信息量,其中 $deep$ 为传输节点到 Sink 节点路由树的深度,num 为模型参数个数.

4 实验测试和性能分析

为测试提出的分布式数据采集优化策略,实验分算法有效性测试和网络能耗分析两个方面.算法有效性测试分析在架构的小型无线传感器网络监测系统中实现,分别取上午、中午和晚上 3 个时间段采样.测试算法对具有时间相关性的温度采样数据信息的预测估计精度.网络能耗优化测试利用了 NS2 网络仿真工具构建无线传感器网络场景,将算法加入到改进的 LEACH 协议中,并与典型的 LEACH、LEACH-C 协议[25]进行数据采集比较,分析了网络生命周期和能耗方面的优化效果.

4.1 算法有效性测试分析

为了测试算法的有效性,在某体育中心足球场部署 6 个 MICAz 传感器节点,形成一个小型的无线传感器网络,用以监控场地的温度、湿度信息,实现足球场地草坪的灌溉控制和提高维护保养质量.以温度信息采集中的 3 个时间段为例进行分析,即上午(6:00~9:00)、中午(11:00~14:00)和晚上(20:00~23:00),网络拓扑结构如图 5 所示.选取节点 1 在每个时间段中 20 个采样时间点(每 9 min 执行 1 次温度采样)的温度监测值为 y 值,各时间段采样时间点与具体采样时间对应情况如表 1 所示.设 $\boldsymbol{\lambda}_s$,$\boldsymbol{\lambda}_z$,$\boldsymbol{\lambda}_w$ 分别表示上午、中午和晚上 3 个时间段取 20 个采样时间点时节点计算得到的回归模型系数向量,根据算法 2 计算得到的结果如表 2 所示,各系数向量下标 0,1,2,3 对应的基函数向量为 $\boldsymbol{t}=[t^3,t^2,t,1]$.三个时间段的回归曲线如图 6 所示,从图中可以看出,取 20 个温度采样时间点时所得到的温度值回归曲线就可以有效地体现各个时间段的温度变化趋势,利用了无线传感器网络数据采集在某一时间

段采样数值具有的时间相关特性，节点 1 通过传输 4 个回归参数即可得到误差允许范围之内的温度估计值，无需每个采样时间点都传输温度值，减少了数据通信量，从而节约了节点的能量消耗. 在 3 个时间段各采样时间点，节点 1 的温度监测值与回归模型估计的温度值误差绝对值如图 7 所示，从图中可以看出，取 20 个采样时间点时最大的误差绝对值为上午时间段的第 8 个采样时间点的 0.8℃，没有超出 1℃，一般的应用中，对于环境温度的监测中，误差阈值范围可设置在 2～3℃，所以，算法在被测量（温度）稳定或呈线性变化时对误差的控制是比较理想的.

表 1　各采样时间段 20 个采样时间点对应的采样时间

采样时间点	上午时间	中午时间	晚上时间
1	6:00	11:00	20:00
2	6:09	11:09	20:09
3	6:18	11:18	20:18
…	…	…	…
20	9:00	14:00	23:00

表 2　取 20 个采样时间点时回归模型的系数向量值

系数下标	λ_s	λ_z	λ_w
0	−0.0014595	−0.00076658	0.0010413
1	0.041379	0.017232	−0.029366
2	0.064083	0.060252	−0.10359
3	20.383	28.005	24.621

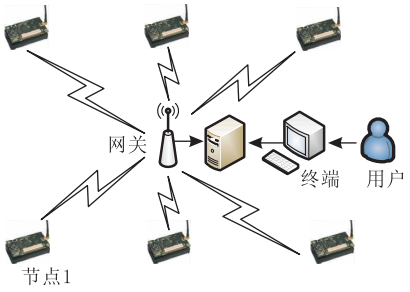

图 5　网络拓扑结构

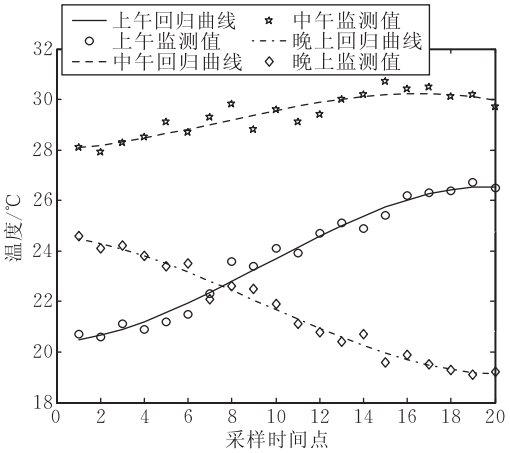

图 6　取 20 个采样时间点对应的温度回归曲线

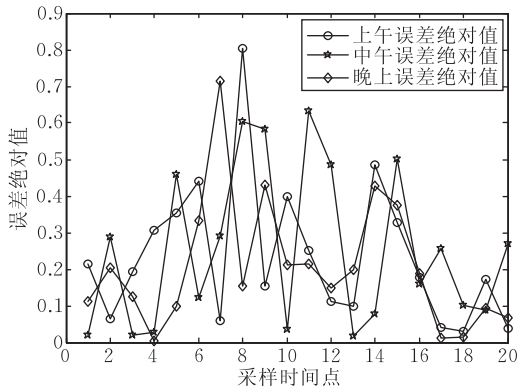

图 7　各采样时间点的温度误差绝对值

如果被测量（温度）在某一采样时间点发生突变，与原有历史数据构建的线性回归模型将产生较大误差. 假设在上午时间段的第 20 个采样时间点，网络中的某一簇头节点接收到簇内节点发送的温度突变值为 223℃，更新模型后形成的回归曲线如图 8 所示，和温度采样正常时产生的误差对比如图 9 所示，仿真结果显示，当温度采样值发生突变时回归模型估计值与实测值存在较大误差，第 20 个采样点的误差值达到近 90℃，远远超出了设定的误差阈值. 此时，系统可采取容错机制进行处理，有两种情况可造成被测量的突变：一是监测环境中被测量的

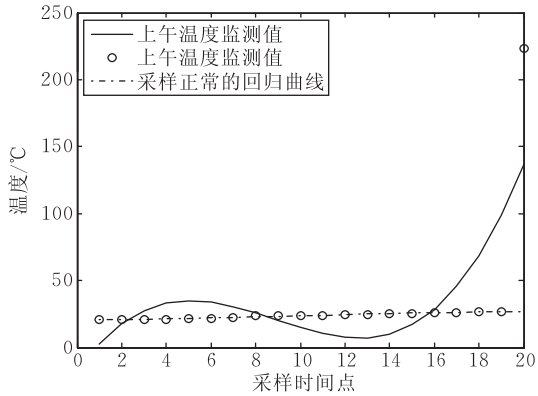

图 8　温度突变时和正常时的回归曲线比较

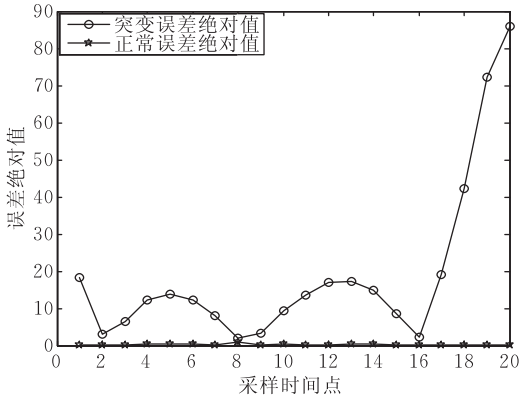

图 9　温度突变时和正常时的误差比较

真实变化(如突发火灾),另一种情况是传感器节点故障造成的误判.如果节点的多数邻居节点也监测到了该被测量突变事件发生,并符合预设的统计假设条件,则可识别为环境监测量的真实变化,簇头节点将转发被测量并利用新的历史数据更新线性回归模型.如果节点的多数邻居节点没有监测到突变事件的发生,并且不符合统计假设条件,则认为该节点出现了误判,标记节点发生错误,簇头将不接收该节点的监测数据.

为了测试不同采样时间点取值对回归模型估计精度的影响,定义温度监测值与回归估计值的均方根误差为:

$$\omega=\sqrt{\left(\sum_{j=1}^{m}(Y(t_j)-y_{t_j})^2\right)/m} \tag{19}$$

在上午时间段,分别取温度采样时间点个数为10,20,30,40和50,表示在上午6:00～9:00这3个小时中,每间隔18 min、9 min、6 min、4 min 30 s、3 min 30 s执行1次温度采样.构建线性回归模型,计算得到的回归曲线如图10所示,根据等式(19)计算得到不同采样点个数的均方根误差如图11所示,误差值范围在0～0.4之间,在预设的误差阈值范围之内,说明了回归模型有效性.

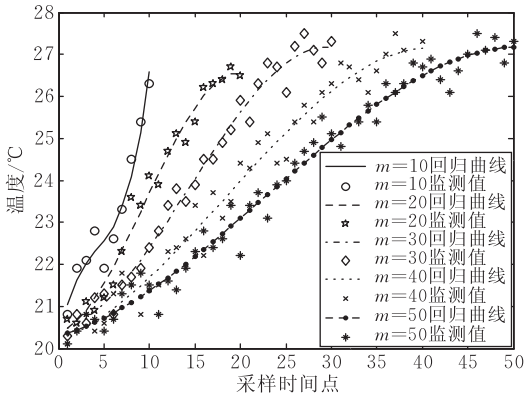

图10 不同采样点个数的温度回归曲线变化情况

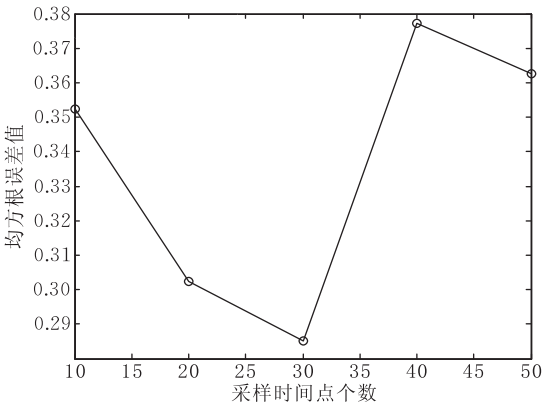

图11 不同采样点个数的温度均方根误差值

从图11中可以看出,本次实验中当采样时间点个数为30时,其均方根误差值最小,即表示其估计精度最高,图11中显示采样时间点个数的选择不一定是越大越好.因为在一定的时间段内,如果采样个数过多,感知数据量就越大,有可能使得数据的离散程度加大;但是,如果采样个数取值过小,同样使得采样时间间隔加大,温度的变化明显,也增大了回归模型计算的估计值与实际监测值之间的误差.所以,可以综合考虑回归模型估计精度和计算复杂度,选取合适的采样个数.

4.2 算法对网络能耗优化的测试分析

为了测试算法在基于分簇的无线传感器网络中进行数据采集时对网络生命周期和整体的能量消耗的影响,选用美国加州大学伯克利分校研发的可扩展、易配置、易编程的网络仿真工具NS2作为仿真平台.LEACH协议是典型的分布式分簇路由协议,其分层的数据转发机制所产生的网络能耗远小于平面路由协议.实验中将线性回归优化处理过程加入到LEACH协议中,网络中簇头节点接收簇内节点的感知信息的同时,还进行线性回归模型计算,用传输回归模型参数信息代替传输感知的原始数据.簇头实现感知数据的估计、预测和容错处理,协议仍保留LEACH的簇头选取方法,但是将簇头节点与Sink节点直接通信方式改为以簇头间多跳方式传送数据.实验中,通过在NS2仿真环境的mit/uAMPS/sims路径下,运行ns genscen命令生成场景文件new100nodes.txt,节点分布如图12所示,即在100 m×100 m的平面区域内随机部署100个传感器节点,基站的位置坐标设置为(50,80),每个节点的初始能量为2J,采用的能量衰减模型如图13所示,根据无线通信原理,发射功率随传输距离的增大呈指数衰减.设发送节点和接收节点之间的距离为d,当d小于设定的常数阈值d_{Thres}时,发射功率呈d^2衰减,即自由空间衰减模型(free space channel model);当d大于d_{Thres}时,发射功率呈d^4衰减,即多路径衰减模型(multipath fading channel model).则发送k比特数据产生的能量消耗$E_{\text{T}}(k,d)$由发射电路能耗$E_{\text{T-elec}}(k)$和功率放大器能耗$E_{\text{T-am}}(k,d)$两部分组成,如式(20)所示.接收k比特数据产生的能量消耗$E_{\text{R}}(k)$仅由电路能耗引起,如式(21)所示.

$$\begin{aligned}E_{\text{T}}(k,d)&=E_{\text{T-elec}}(k)+E_{\text{T-am}}(k,d)\\&=\begin{cases}k\times E_{\text{elec}}+k\times\varepsilon_s\times d^2 & d<d_{\text{Thres}}\\ k\times E_{\text{elec}}+k\times\varepsilon_m\times d^4 & d\geq d_{\text{Thres}}\end{cases}\end{aligned} \tag{20}$$

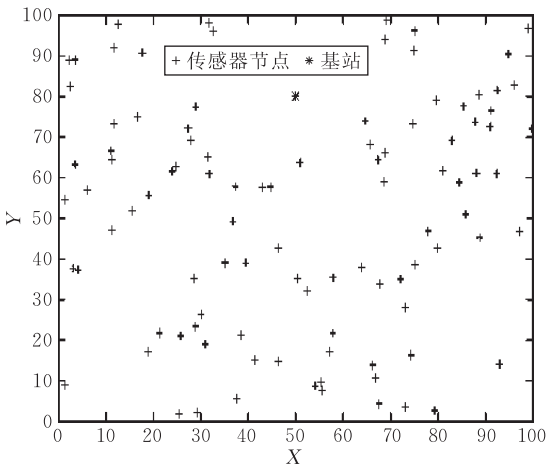

图 12　传感器节点分布图

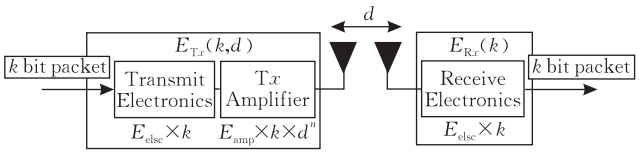

图 13　无线通信能量衰减模型

$$E_R(k) = E_{R\text{-elec}}(k) = k \times E_{\text{elec}} \tag{21}$$

其中：E_{elec}为发射(接收)1 bit 数据的能量消耗；ε_s和ε_m分别表示自由空间衰减模型和多路径衰减模型下发射 1 bit 数据功率放大器所需能量. 另外，对于簇头节点计算 1 次回归模型产生的能量消耗为 $E_{re} = n_{\text{CH}} \times E_{\text{com}}$，其中 n_{CH}表示簇头节点个数，E_{com}表示簇头节点计算 1 次回归模型所耗费的能量.

仿真实验中，设 $n_{CH} = 5$，$E_{\text{elec}} = 50$ nJ/bit，$\varepsilon_s = 10$ pJ/(bit·m^2)，$\varepsilon_m = 0.0013$ pJ/(bit·m^4)，$E_{\text{com}} = 5$ nJ/bit，带宽为 1 Mbps，消息长度为 500 Bytes，发送和接收的时延为 25 μs，仿真时间为 500 s，每轮簇头选举的时间间隔为 20 s，线性回归模型数据采样个数为 20，回归模型更新周期(Regression Period)为 60 s. 图 14 显示了仿真时间每间隔 20 s 时 LEACH 协议和线性回归策略(Regression)的簇头节点工作总能耗. 从实验结果可以看出，LEACH 协议中每轮簇头节点的总能耗在 4J～5J 之间，线性回归策略中每轮簇头节点的总能耗在 1.5J～2.5J 之间，明显低于 LEACH 协议. 虽然在仿真时间到 380 s 之后，LEACH 协议的簇头节点能耗出现下降，但并不是实际总能耗的减少，而是由于随着仿真时间的增加，网络中节点的能耗已经接近于节点的初始能量，被选举出的部分簇头节点在工作中能耗达到 2J，中途死亡. 实验计算的是簇头节点在正常工作时的总能耗. 另外，从图中结果看出，仿真时间每间隔 60 s，回归策略的簇头总能耗就有所增加，原因是由于簇头节点在每个回归模型的更新周期都要重新计算回归模型参数，并发送更新后的参数到基站，这样，就产生了比其他仿真时间更多的计算和通信能耗.

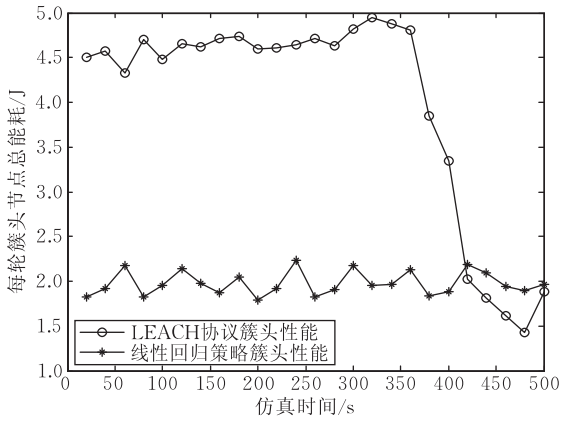

图 14　簇头总能耗随仿真时间变化情况

为了在仿真环境中模拟实际被测量变化对算法能耗的影响，将线性回归模型的更新周期分别设置为 10 s，30 s，60 s，则每轮簇头节点的总能耗如图 15 所示. 从实验结果中可以看出，回归更新周期设置越短，表示被测量的突变频率越高，簇头节点计算和重传回归模型参数的次数越多，所产生的总能耗越大. 实际环境监测应用中，被测量在大部分情况处于线性变化时，回归模型参数更新和重传的频率将很低，那么，网络的总能耗也将随之减少.

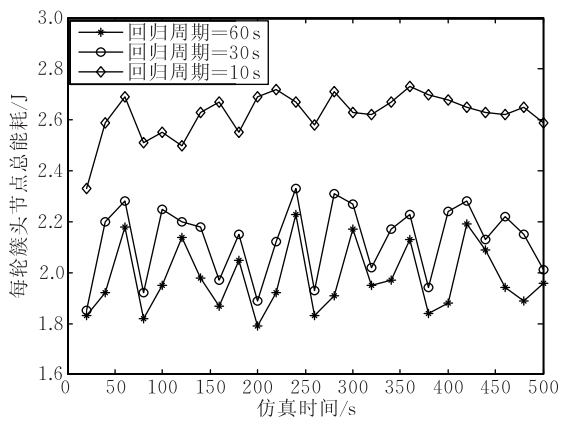

图 15　回归周期为 10 s，30 s，60 s 时簇头总能耗变化情况

在 500 s 的仿真时间中，更新周期为 60 s 时，线性回归策略的节点总能耗情况如图 16 所示，对网络生命周期的影响如图 17 所示. 从图 16 和图 17 可以看出，在相同待传输数据量情况下，与典型的 LEACH 协议和 LEACH-C 协议相比，线性回归策略的加入使得网络的整体能量消耗降低，节点的存活时间延长. 加入回归模型后，第 1 个节点的死亡时间为 420 s 之后，而 LEACH 协议的第 1 个死亡节点

出现在 210 s. 仿真时间到 500 s 时，网络的能量消耗比 LEACH 和 LEACH-C 协议大约节省了 100 J. 因为回归策略在节点传输采样数据时，考虑了采样数据在时间上的相关性，根据已有历史采样数据，构建了线性回归模型，节点在监测值误差阈值范围内，通过上传回归模型参数表示实际感知数据. 虽然增加了一定的节点计算能耗，但却大大降低了节点间的通信能耗.

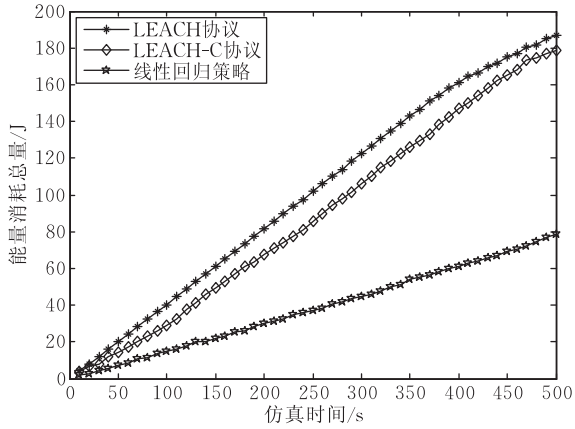

图 16 节点总能耗随仿真时间的变化情况

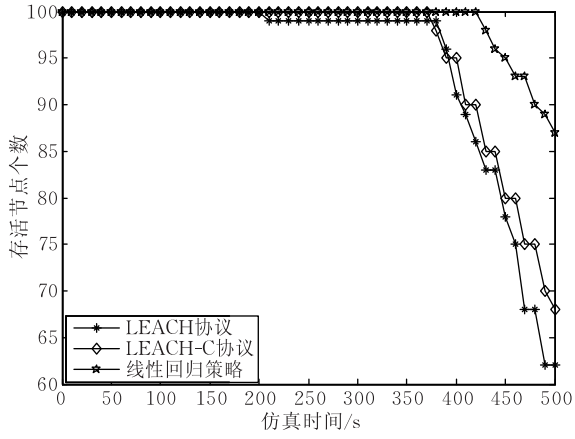

图 17 节点存活数量随仿真时间的变化情况

综合分析图 14 和图 16 所示的实验结果，在 380 s 之后，由于部分簇头节点中途死亡，造成了网络总能耗的增量逐渐减缓，尤其是 LEACH 和 LEACH-C 协议. 为了测试网络在簇头节点不死亡的情况下，产生的额外能耗总量，假设在下一轮簇头选举开始之前，簇头不受初始能量限制，继续完成当前数据传输任务，实验结果如图 18 所示，在仿真时间到达 500 s 时，LEACH 和 LEACH-C 协议产生的额外能耗总量在 80 J 左右，线性回归策略所需的额外能耗在 1 J 左右，明显低于 LEACH 和 LEACH-C 协议，原因在于线性回归策略在仿真过程中，节点的总能耗较低，大部分被选举出的簇头节点在 2 J 的初始能量耗尽之前能完成本轮任务.

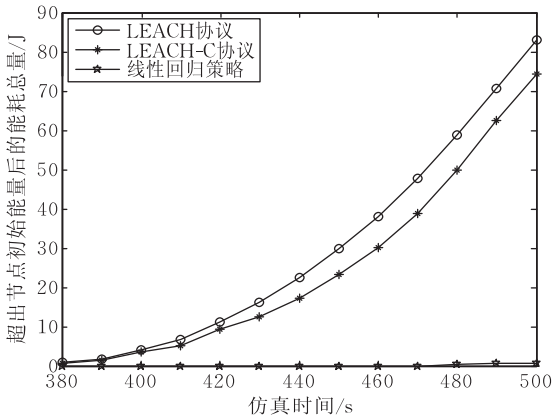

图 18 簇头不死亡情况下产生的额外能耗总量

5 结论和未来工作

本文提出了一种能量高效的基于线性回归的数据采集优化方法，利用了数据在一定采样时间内的相关性，构建了线性回归模型，在分簇的无线传感器网络拓扑结构下用传输回归模型参数实现采样数据预测，不失数据的一般特征. 通过构建足球场地小型无线传感器网络监测系统采集环境温度信息验证了算法的有效性，实现了对具有时间相关特性的感知数据的预测估计. 在随机部署的无线传感器网络仿真环境中，提出的数据采集优化策略在延长网络生命周期和减少网络总能耗方面也有良好表现，体现了算法的可行性和能量高效性.

下一步工作中，将继续深入研究线性回归模型的优化策略，通过引入核函数，有效的减少模型中矩阵的运算量，针对采样数据的不断变化，研究自适应的选取采样时间点个数的方法，达到降低回归估计值与实际监测值之间的误差以及当采样数据超出回归估计值的置信区间时的处理策略等等.

参 考 文 献

[1] Estrin D. Wireless sensor networks tutorial part Ⅳ: Sensor network protocols//Proceedings of the Invited Speech of International Conference on Mobile Computing and Networking (Mobicom). Atlanta, USA, 2005

[2] Wander A S, Gura N, Eberle H et al. Energy analysis of public-key cryptography for wireless sensor networks//Proceeding of the 3rd IEEE International Conference on Computing and Communications. Seattle, USA, 2005: 324-328

[3] Lin C, Tian Y, Yao M. Green network and green evaluation: mechanism, moodeling and evaluation. Chinese Journal of Computers, 2011, 34(4): 593-612(in Chinese)

(林闯，田源，姚敏. 绿色网络和绿色评价：节能机制、模型和评价. 计算机学报，2011，34(4)：593-612)

[4] Zheng J，Wang P，Li C . Distributed data aggregation using slepian—wolf coding in cluster-based wireless sensor networks. IEEE Transactions on Vehicular Technology，2010，59(5)：2564-2574

[5] Chen H F，Mineno H，Mizuno T. Adaptive data aggregation scheme in clustered wireless sensor networks. Computer Communications，2008，31(15)：3579-3585

[6] Nurdin H I，Mazumdar R R，Bagchi A. Reduced-dimension linear transform coding of distributed correlated signals with incomplete observations. IEEE Transactions on Information Theory，2009，55(6)：2848-2058

[7] Oka A，Lampe L. Energy efficient distributed filtering with wireless sensor networks. IEEE Transactions on Signal Process，2008，56(5)：2062-2075

[8] Jiang H B，Jin S D，Wang C G. Prediction or not? An energy-efficient Framework for clustering-based data collection in wireless sensor networks. IEEE Transactions on Parallel and Distributed Systems，2011，22(6)：1064-1071

[9] Yang J，Zhang D Y，Zhang Y Y，Wang Y. Cluster-based data aggregation and transmission protocol for wireless sensor networks. Journal of Software，2010，21(5)：1127-1137(in Chinese)
(杨军，张德运，张云翼，王毅. 基于分簇的无线传感器网络数据汇聚传送协议. 软件学报，2010，21(5)：1127-1137)

[10] Wu Y W，Li X Y，Liu Y H，Lou W. Energy-efficient wake-up scheduling for data collection and aggregation. IEEE Transactions on Parallel and Distributed Systems，2010，21(2)：275-287

[11] Xu X H，Li X Y，Mao X F et al. A delay-efficient algorithm for data aggregation in multihop wireless sensor networks. IEEE Transactions on Parallel and Distributed Systems，2011，22(1)：163-175

[12] Keshavarzian A，Lee H，Venkatraman L. Wakeup scheduling in wireless sensor networks//Proceedings of the 7th ACM International Symposium on Mobile ad Hoc Networking and Computing. Florence，Italy，2006：322-333

[13] Zhu J M，Hu X D. Improved algorithm for minimum data aggregation time problem in wireless sensor networks. Journal of Systems Science and Complexity，2008，21(4)：618-628

[14] Lei L，Lin C，Cai J. Performance analysis of wireless opportunistic schedulers using stochastic petri nets. IEEE Transactions on Wireless Communications，2009，8(4)：2076-2087

[15] Liu B，Ren F Y，Lin C，Jiang X. Performance analysis of sleep scheduling schemes in sensor networks using stochastic Petri net//Proceedings of the IEEE International Conference on Communications. Beijing，China，2008：4278-4283

[16] Ren Q Q，Li J Z，Gao H，Cheng S Y. A two-phase sleep scheduling based protocol for target tracking in sensor network. Chinese Journal of Computers，2009，32(10)：1971-1979(in Chinese)
(任倩倩，李建中，高宏，程思瑶. 传感器网络中一种基于两阶段睡眠调度的目标跟踪协议. 计算机学报，2009，32(10)：1971-1979)

[17] Charalampos K，Aristides P，Damianos G，Grammati P. Effective determination of mobile agent itineraries for data aggregationon sensor networks. IEEE Transactions on Knowledge and Data Engineering，2010，22(12)：1679-1693

[18] Biswas P，Qi H，Xu Y. Mobile agent-based collaborative sensor fusion. Information Fusion，2008，9(3)：399-411

[19] Xiong B B，Lin C，Ren F Y. Performance analysis of stochastic delivery transport protocols in WSNs. Journal of Software，2009，20(4)：942-953(in Chinese)
(熊斌斌，林闯，任丰原. 无线传感器网络随机投递传输协议性能分析. 软件学报，2009，20(4)：942-953)

[20] Jiang H，Jin S. Leap：Localized energy-aware prediction for data collection in wireless sensor networks//Proceedings of the 5th IEEE International Conference on Mobile Ad Hoc and Sensor Systems. Atlanta，USA，2008：491-496

[21] Guestrint C，Bodikz P，Thibauxt R et al. Distributed regression：an efficient framework for modeling sensor network data//Proceedings of the ACM International Conference on Sensor Networks. Berkeley，California，USA，2004：1-10

[22] Deligiannakis A，Kotidis Y，Roussopoulos N. Compressing historical information in sensor networks//Proceedings of the ACM SIGMOD International Conference on Management of data. Paris，France，2004：527-538

[23] Jiang H B，Jin S D，and Wang C G. Parameter-based data aggregation for statistical information extraction in wireless sensor networks. IEEE Transactions on Vehicular Technology，2010，59(8)：3992-4001

[24] Hao Z，Ioannis D Schizas，Georgios B Giannakis. Power-efficient dimensionality reduction for distributed channel-aware kalman tracking using WSNs. IEEE Transactions on Signal Processing，2009，57(8)：3193-3207

[25] Heinzelman W，Chandrakasan A，Balakrishnan H. An application- specific protocol architecture for wireless microsensor networks. IEEE Transactions on Wireless Communications，2002，1(4)：660-670

SONG Xin，born in 1978，Ph. D. candidate. Her research interests include wireless sensor networks，distributed computing and intelligent information processing.

WANG Cui-Rong，born in 1963，Ph. D. professor. Her research interests include wireless sensor networks，next generation network technology and data center.

Background

For most application in wireless sensor networks, users may want to continuously extract data from the networks for further analysis the monitoring region measurements. However, accurate data extraction is very difficult and too costly to obtain all sensory data. Some monitoring systems are typically used in one of two modes of operation: either the data from the network sensors is extracted and analyzed off-line or the information obtained from the sensors is aggregated using local computing operations. The reduction of communication through the latter technology is attractive since extraction of complete data sets requiring large amounts of communication that drains the limited energy of sensor nodes. One strategy so-called prediction in recent years has been introduced for in-network data aggregation. The existence of such technology capability implies that the sensor nodes do not need to transmit all sensory data if they differ from a predicted data by less than a pre-specified threshold, or error bound. In many previous researches, most prediction operations are carried out by the base station only, but not the sensor nodes. The disadvantages are potential high latency time and bandwidth for transmitting the raw data to the base station. The paper focuses on the sensor nodes (or cluster head nodes) distributed compute a linear regression model according to their local measurements for extracting much more complete information about the feature of sensor data while still use much less communication than methods that retrieve all reading from all sensor. The strategy not only retains valuable statistical information of the measurements based on the coefficients of the basic functions of linear regression model but also greatly reduces the communication cost.

The research work was supported by the National Natural Science Foundation of China under Grant Nos. 61070162 and 71071028, and Open Research Fund of Key Laboratory of Complex System and Intelligence Science, Institute of Automation, Chinese Academy of Sciences under Grant No. 20100106.