软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(Suppl.(1)):93−102

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于相似度预测的 WSN 数据收集算法*

李 平[1], 阳 武[1], 吴佳英[1], 胡海罗[2]

[1](长沙理工大学 计算机与通信工程学院,湖南 长沙 410114)

[2](中南勘测设计研究院,湖南 长沙 410014)

通讯作者: 李平, E-mail: lping9188@163.com

摘 要: 针对基于拟合曲线的数据预测机制在 WSN 数据收集应用中区间敏感的问题,提出了基于时间周期的拟合曲线相似度序列,将基于预测的数据收集问题转化为一定精度下预测相似度的估计问题.基于特征相似度服从高斯分布的假定,研究准确预测感知相似度的最大概率,采用贪婪算法动态调整预测相似度.最后,采用 PSO 算法实现基于预测相似度的预测数据推断.仿真实验结果表明,该算法达到了预期效果,在能耗方面有较大的提高.

关键词: 曲线相似度;拟合;无线传感器网络;数据预测

中文引用格式: 李平,阳武,吴佳英,胡海罗.基于相似度预测的 WSN 数据收集算法.软件学报,2014,25(Suppl.(1)):93−102. http://www.jos.org.cn/1000-9825/14011.htm

英文引用格式: Li P, Yang W, Wu JY, Hu HL. Algorithm of WSN data collection based on similarity prediction. Ruan Jian Xue Bao/Journal of Software, 2014,25(Suppl.(1)):93−102 (in Chinese). http://www.jos.org.cn/1000-9825/14011.htm

Algorithm of WSN Data Collection Based on Similarity Prediction

LI Ping[1], YANG Wu[1], WU Jia-Ying[1], HU Hai-Luo[2]

[1](School of Computer and Telecommunications, Changsha University of Science and Technology, Changsha 410114, China)

[2](Mid-South Design and Research Institute, Changsha 410014, China)

Corresponding author: LI Ping, E-mail: lping9188@163.com

Abstract: To tackle the issue of interval sensitivity in the application of WSN data collection based on the prediction mechanism of fitting curve, this paper proposes fitting curve similarity sequence based on time periods to transform data collection problem based on prediction into the similarity estimation under certain accuracy. Feature-Based similarity is assumed to obey the Gaussian distribution. By studying the maximum probability of accurate prediction of perceived similarity, the proposed method uses Greedy algorithm to dynamically adjust the predicted similarity. Finally, it uses PSO algorithm to achieve inference of the predicted data based on predicted similarity. Simulation results show that this algorithm has achieved the desired results, and also provides great improvement in terms of energy consumption.

Key words: curve similarity; fitting; wireless sensor network (WSN); data prediction

近年来,WSN 技术广泛应用于军事侦察、目标跟踪、生态环境监测以及工农业等领域[1],然而,受节点电池能量有限且无法有效得到补充的影响,能量的节省成为 WSN 的一个重要研究方向[2,3].随着 WSN 降低能耗研究的深入,依赖时间相关性[4−7]的数据预测技术[8−10]已经有了较大的突破,对时间相关性的研究则促进了 WSN 数据预测技术的发展.基于预测机制的数据收集算法被视为弥补当前数据融合[11]缺陷的关键性技术.

在大多的无线传感器网络环境监测应用系统中,一段时间内传感器节点周围的环境变化不大,节点采集的数据序列也呈现一定的动态依存关系(时间相关性).因此,可以通过建立具有一定动态依存关系的数学模型实现未来感应数据的预测.若采集值落入预测值的某个范围内,则不用发送数据.如此,节点发送的数据量将大大

* 基金项目: 湖南省教育厅资助重点项目(14A004); 湖南省教育厅资助科研项目(13C1022))
收稿时间: 2014-05-10; 定稿时间: 2014-08-26

减少,从而节省节点能量,延长网络生命周期.目前,WSN 感知数据的时间相关性表现在两个方面:一方面是由于连续采样引起的数据相似相关性,WSN 数据的这一特点被广泛用于数据预测技术的研究;另一方面是相邻周期数据曲线形状的相似性(与单个数据相比,多个数据更能体现数据变化的特点),这可以作为数据预测的参考,提高预测精度,改善算法的准确性.基于此,本文提出了基于相似度预测的 WSN 数据收集算法.

1 研究进展

基于预测机制的数据收集算法首先通过预测算法实现数据预测,然后根据预测数据与实际数据的差值决定是否发送当前数据到基站,其关键在于如何设计基于时间相关性预测算法.对于大多数 WSN 应用,单个节点在时间上的连续观测值存在一定的相似或相关性.Kusuma 等人在 2001 年第一次对时间相关性进行定义[4].此后,Madden 等人提出了 TAG[5]模型.该模型的贡献在于:引入了时间一致性概念,减少数据传输量.TiNA[6]模型在每个采样周期对传感器节点当前时刻与上一时刻的采样数据进行比较,若差值小于用户预定义的阈值则不传送数据,基站将上一个收到的数据作为本次收集的数据.这种方式只有当相邻的两个采样数据的差值较小时才有效,而且容易造成基站采集值的偏差值累积.文献[7]则提出了一种较为成熟的双预测模型,它包含了两个完全相同的动态预测模型,一个部署于 WSN 节点,另一个部署于基站.当预测失败时,节点发送当前感知数据到基站,两个模型使用实际采样数据同步更新预测模型.这种双预测模型成为 WSN 中一种成熟的数据预测技术框架.

基于预测机制的数据收集算法的关键是预测机制的设计.目前主要的预测机制包括线性预测机制[12]、概率预测模型[7]、Kalma 滤波[13,14]模型等.此外,为了提高数据预测的性能,GEP 算法[9]和 PSO-BPNN 算法[10]被用来进行数据相关性挖掘和数据预测.这些算法虽然能够选择最佳属性并有效降低通信次数,但也大幅度增加了算法复杂度并且对预测精度的提高相对有限.

文献[15]提出了一种 Model-Aided 算法.该算法采用基于泰勒定理的多项式曲线拟合方法将时间区间内的感知数据拟合成多项式曲线(函数),然后根据此函数预测区间外的感知数据,最后,对比预测值和实测值决定是否将感知数据发送到基站.同时,基站也包含与感知节点相同的预测模型.若基站没有接收到数据,则直接通过模型预测数据.此方法在一定程度上减少了感知节点的数据发送量,但是,如图 1 所示,此方法存在以下问题:

(1) 通过文献[15]的方法拟合的曲线具有区间敏感的特点,尽管通过多项式方法能够准确地实现区间内(如图 1 中 $x\in[1,5]$) 的多项式曲线拟合,但是,利用此曲线直接进行数据预测(如图 1 中 $x=6,7...$) 的误差太大(一般区间外的数据与当前的多项式曲线的变化趋势不同).

(2) 此方法不能有效去除少量噪音数据的影响(突变的噪音数据一般与当前数据差异较大,而之后的数据与当前数据相似,传统的预测模型需要传输噪音数据到基站).

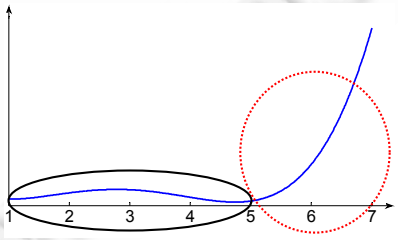

Fig.1 Data prediction model based on curve fitting

图 1 基于曲线拟合的数据预测模型

基于此,本文提出了基于时间周期的拟合曲线相似度序列,利用文献[15]的方法将区间内的数据拟合成多项式曲线,根据连续时间曲线之差的积分(曲线相似度)的变化特点,将基于预测的数据收集问题转化为一定精度下的预测相似度的估计问题.基于特征相似度服从高斯分布的假定,通过研究准确预测感知相似度的最大概率,采用贪婪算法动态调整预测相似度.最后,采用 PSO 算法实现基于预测相似度的预测数据推断.

本文首先介绍基于相似度预测的 WSN 数据收集算法的研究背景.第 1 节介绍基于预测机制的 WSN 数据

收集算法的研究现状.第 2 节介绍算法的条件假设及相关定义.第 3 节描述算法步骤并分析算法性能.第 4 节介绍算法的仿真实验分析.第 5 节对全文进行总结.

2 条件假设及相关定义

2.1 系统整体构架

本文研究的无线传感器网络系统采用分层的体系结构,如图 2 所示,整个系统由底层到上层分别为无线传感器网络、服务器以及客户端.其中,无线传感器网络负责获取信息,通过网关将数据传输到服务器.服务器存储从无线传感器网络获取的相关信息,向本地用户以及远程用户提供相应的服务.本地用户可以直接访问数据库服务器,远程用户可以通过 Internet 访问 Internet 服务器.同时,对于节点采集,发送数据过程做如下两个假设:(1) 网络中的节点以固定频率 f [15]周期性地采集数据.(2) 当感知节点发送数据包到基站时,整个通信过程是可靠的[16](不丢失数据包).

在 WSN 部署前,我们在传感器节点中预置一个前端预测模型,该模型以当前时间周期之前的特征相似度序列(在某个时间周期,若节点需要传输感知相似度到服务器,则特征相似度为感知相似度,否则特征相似度为预测相似度)作为输入,输出当前时间周期的预测相似度.在该时间周期末,节点首先将该时间周期采集的数据拟合成多项式曲线并计算感知相似度,然后计算感知相似度与预测相似度的差值,若差值大于某一阈值,则发送替代当前感知数据的感知相似度,否则不发送数据.最后,更新特征相似度序列,为下一个周期的预测做好准备.

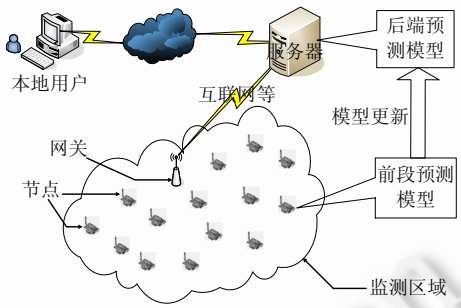

Fig.2 System architecture

图 2 系统架构

同样地,我们在服务器中预置一个后端预测模型,该模型同样以当前时间周期之前的特征相似度序列(在某个时间周期,若服务器接收到节点传输的感知相似度,则特征相似度为感知相似度,否则特征相似度为预测相似度)作为输入,输出当前周期的预测相似度.服务器等待一个周期后,根据是否接收到感知相似度确定当前的特征相似度.然后根据特征相似度以及前一周期的特征曲线反推特征曲线,计算特征数据集,并将特征数据集传输给用户.最后,更新特征相似度序列,使其与节点的预测模型保持同步.这种方法在满足用户对于数据准确性要求的前提下,有效地减少了无线传感器网络与服务器之间的数据传输量,同时具有实时监测数据变化并自适应调节的能力,使得该系统不会漏检突发事件.

2.2 相关定义及说明

定义 1(拟合周期序列 *t*). 假设节点的采样频率固定为 f,而曲线拟合的频率固定为 $f'=f/k=1/T$,依次定义 $\underbrace{[0,T]}_{1},\underbrace{[T,2T]}_{2}$... 为拟合周期序列 $t(1,2,...)$.其中,k 为拟合周期内的采样数据数量.

定义 2(特征数据集). 假设感知节点 a_i 在拟合周期序列 t 的感知数据集和预测数据集分别为 $ZS(a_i)_t=\{zs_1(a_i)_t,zs_2(a_i)_t,...,zs_k(a_i)_t\}$,$ZP(a_i)_t=\{zp_1(a_i)_t,zp_2(a_i)_t,...,zp_k(a_i)_t\}$,定义 $ZC(a_i)_t$ 为拟合周期序列 t 的特征数据集.

$$ZC(a_i)_t=\begin{cases}ZS(a_i)_t, & ZS(a_i)_t\text{满足发送要求}\\ ZP(a_i)_t, & ZS(a_i)_t\text{不满足发送要求}\end{cases} \tag{1}$$

最小二乘多项式曲线拟合. 如果 $f(x)$ 在开区间 (a,b) 内含有连续的 $n+1$ 阶导数,根据泰勒定理(如式(2)所示),$f(x)$ 可以用关于 $(x-x_0)$ 的多项式之和逼近.

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+f''(x_0)\frac{(x-x_0)^2}{2!}+...+f^n(x_0)\frac{(x-x_0)^n}{n!}+R_n(x) \tag{2}$$

若给定的一组目标函数为 $f(x)$ 的二维数据 $(x_i,y_i),i=1,2,...,n,x_i\neq x_j$, 则采用最小二乘方法(σ 取最小值)可以确定 $\lambda_0,\lambda_1,...,\lambda_{k-1}$ 的取值,从而实现二维数据的曲线拟合.

$$f(x)=\lambda_0+\lambda_1x+\lambda_2x^2+\lambda_3x^3+...+\lambda_{k-1}x^{k-1},x\in[a,b] \tag{3}$$

$$\sigma=\sum_{i=1}^{n}(f(x_i)-y_i)^2 \tag{4}$$

定义 3(特征曲线). 给定 $ZS(a_i)_t,ZP(a_i)_t$ 以及 $ZC(a_i)_t$, 根据最小二乘曲线拟合理论,定义 $ZS(a_i)_t$ 拟合的曲线 $f_{s,a_i}^t(x)(x\in[a,b])$ 为感知曲线; $ZP(a_i)_t$ 拟合的曲线 $f_{p,a_i}^t(x)(x\in[a,b])$ 为预测曲线; $ZC(a_i)_t$ 拟合的曲线 $f_{c,a_i}^t(x)(x\in[a,b])$ 为特征曲线.同时,我们称 $f_{c,a_i}^1(x),f_{c,a_i}^2(x),...,f_{c,a_i}^t(x)$ 为特征曲线序列.

定义 4(曲线相似度). 给定曲线 $F_i(x),F_j(x),x\in[x_1,x_2]$, 定义 $s(F_i(x),F_j(x))$ 为 $F_i(x)$ 与 $F_j(x)$ 的曲线相似度.

$$s(F_i(x),F_j(x))=\int_{x_1}^{x_2}[F_i(x)-F_j(x)]\,\mathrm{d}x \tag{5}$$

定义 5(特征相似度). 给定 $f_{c,a_i}^{t-1}(x),f_{s,a_i}^t(x),f_{p,a_i}^t(x),x\in[a,b]$, 定义 $s_{s,a_i}(t-1)=s(f_{c,a_i}^{t-1}(x),f_{s,a_i}^t(x))$ 为拟合周期序列 t 的感知相似度, $s_{p,a_i}(t-1)=s(f_{c,a_i}^{t-1}(x),f_{p,a_i}^t(x))$ 为拟合周期序列 t 的预测相似度, $s_{c,a_i}(t-1)=s(f_{c,a_i}^{t-1}(x),f_{c,a_i}^t(x))$ 为拟合周期序列 t 的特征相似度.同时,我们称 $s_{c,a_i}(1),s_{c,a_i}(2),...,s_{c,a_i}(t-1)$ 为拟合周期序列 t 的特征相似度序列.

发送条件假设. 对于拟合周期序列 t,给定 $s_{s,a_i}(t-1),s_{p,a_i}(t-1)$ 若 $|s_{s,a_i}(t-1)-s_{p,a_i}(t-1)|\leqslant\varepsilon$, 则节点 a_i 不发送当前的感知数据,若 $|s_{s,a_i}(t-1)-s_{p,a_i}(t-1)|>\varepsilon$, 则节点发送 $s_{s,a_i}(t-1)$ 到服务器. ε 为是否发送数据的阈值.

3 算法描述及性能分析

本文提出的基于相似度预测的 WSN 数据收集算法分两个部分,第 1 部分为数据收集算法,第 2 部分为数据预测机制和性能分析.

3.1 数据收集算法

假设 WSN 部署前,节点和服务器都布置了初始的特征数据集、特征曲线,以及特征相似度序列.图 3 和图 4 分别为基于相似度预测的 WSN 数据收集算法的节点工作流程和服务器工作流程,其主要步骤如下:

(1) 对于传感器节点 a_i, 初始化 k,ε, 根据拟合周期序列 t 之前的感知数据确定特征数据集 $ZC(a_i)_1$, $ZC(a_i)_2,...,ZC(a_i)_{t-1}$, 计算特征曲线序列 $f_{c,a_i}^1(x),f_{c,a_i}^2(x),...,f_{c,a_i}^{t-1}(x)$ 以及特征相似度序列 $s_{c,a_i}(1),s_{c,a_i}(2),...,s_{c,a_i}(t-2)$.

(2) 当检测到拟合周期序列 t 的感知数据集 $ZS(a_i)_t$ 后做如下处理:拟合感知曲线 $f_{s,a_i}^t(x)$;计算感知相似度 $s_{s,a_i}(t-1)$;求解拟合周期序列 t 的预测相似度 $s_{p,a_i}(t-1)$,具体求解方法详见下文.

(3) 确定是否发送当前的感知相似度到服务器并求出当前的特征曲线,特征数据集.若 $|s_{s,a_i}(t-1)-|s_{p,a_i}(t-1)|>\varepsilon$, 则将 $s_{c,a_i}(t-1)$ 发送到服务器,此时 $s_{c,a_i}(t-1)=s_{s,a_i}(t-1)$. 若 $|s_{s,a_i}(t-1)-s_{p,a_i}(t-1)|\leqslant\varepsilon$, 则不发送当前的特征相似度,此时 $s_{c,a_i}(t-1)=s_{p,a_i}(t-1)$.

(4) 将 $s_{c,a_i}(t-1)$ 添加到 $s_{c,a_i}(1),s_{c,a_i}(2),...,s_{c,a_i}(t-2)$ 形成新的特征相似度序列 $s_{c,a_i}(1),s_{c,a_i}(2),...,s_{c,a_i}(t-1)$.

(5) 更新感知节点的预测模型,回到步骤(2),等待下一拟合周期序列的感知数据.

同样地,对于节点 a_i 在服务器存在一个后端预测模型,具体工作流程的步骤如下:

(1) 初始化 k ,根据拟合周期序列 t 之前的特征相似度序列 $s_{c,a_i}(1),s_{c,a_i}(2),...,s_{c,a_i}(t-2)$ 求解拟合周期序列 t 的预测相似度 $s_{p,a_i}(t-1)$, 具体求解方法详见下文.

(2) 根据 $s_{p,a_i}(t-1)$ 以及 $f_{c,a_i}^{t-1}(x)$ 确定拟合周期序列 t 的预测曲线 $f_{p,a_i}^t(x)$ 和预测数据集 $ZC(a_i)_t$, 具体确定方法详见下文.

(3) 等待一个拟合周期,监测是否接收到拟合周期序列 t 传输的感知相似度,若接收到感知相似度,则根据步骤(2)的方法确定感知曲线 $f_{s,a_i}^t(x)$ 和感知数据集 $ZS(a_i)_t$.

(4) 根据定义 2、定义 3、定义 5,确定特征相似度 $s_{c,a_i}(t-1)$,特征曲线 $f^t_{c,a_i}(x)$,特征数据集 $ZC(a_i)_t$ 并将特征数据集发送到用户.

(5) 以与节点工作流程中步骤(4)和步骤(5)相同的方式更新模型,回到步骤(2),等待下一拟合周期序列是否接收到数据.

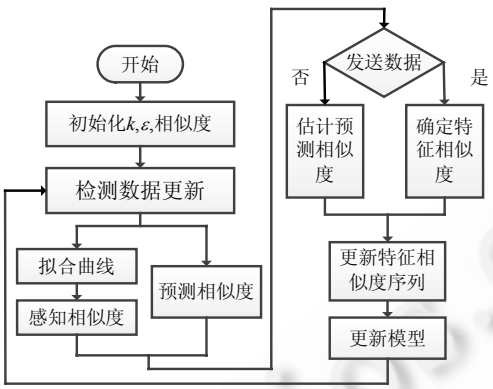

Fig.3 WSN nodes running flowchart

图 3 WSN 节点工作流程图

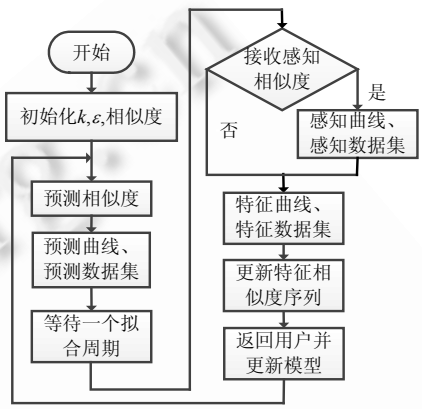

Fig.4 Server running flowchart

图 4 服务器工作流程图

3.2 预测机制及性能分析

WSN 常被用于监测矿井、森林等环境的异常,由于环境变化的随机性,文献[17]假设 WSN 节点监测的感知数据服从高斯分布.根据拉格朗日中值定理,在多项式曲线的拟合区间内,一定可以找到一点的函数取值乘以拟合区间跨度等于多项式曲线的定积分(面积).

由于相似度是由连续两段曲线的差的积分构成,所以,在时间尺度较小的情况下,我们可以认为相似度是由两个独立的服从正态分布的随机变量的差值,而此差值同样服从正态分布,基于此,我们可以近似地认为相似度服从高斯分布.

第 3.1 节数据收集算法的关键是前端预测模型和后端预测模型预测相似度的求解,以及后端预测模型中基于特征相似度的数据推断.对于节点 a_i 和服务器在拟合周期序列 t,假设以 $s_{s,a_i}(1),s_{s,a_i}(2),...,s_{s,a_i}(t-2)$ 为样本观测值的随机变量 $S_s \sim N(\mu,\sigma^2)$,由于 S_c 总是"努力"逼近 S_s,所以我们近似地认为以 $s_{c,a_i}(1),s_{c,a_i}(2),...,s_{c,a_i}(t-2)$ 为样本观测值的随机变量 $S_c \sim N(\mu,\sigma^2)$.由于节点 a_i 和服务器都拥有 S_c 的样本观测值,我们将 S_c 的极大似然估计值 $\hat{\mu}$ 初始化预测相似度 $s_{p,a_i}(t-1)$,然后对比感知相似度在区间 $[\hat{\mu}-\varepsilon,\hat{\mu}+\varepsilon]$ 与 $[\hat{\mu}+\alpha-\varepsilon,\hat{\mu}+\alpha+\varepsilon]$ 的概率调整 $s_{p,a_i}(t-1)$ 的取值,使得成功预测感知相似度的概率最大化,同时,对于后端服务器,研究基于 PSO 的数据预测机制,通过特征相似度推断特征数据并传输到用户.

性质 1. 给定拟合周期序列 t 的特征相似度序列 $s_{c,a_i}(1),s_{c,a_i}(2),...,s_{c,a_i}(t-1)$,则拟合周期序列 $t+1$ 的预测相似度初始值 $s_{p,a_i}(t)=\hat{\mu}=\overline{S_c}=\dfrac{\sum_{j=1}^{t-1}s_{c,a_i}(j)}{t-1}$.

证明:由 $\overline{S_c} \sim N(\mu,\sigma^2)$,参数 μ,σ^2 的取值均未知,且 $s_{p,a_i}(t)$ 取 μ 的极大似然估计值,我们将 $s_{c,a_i}(1),s_{c,a_i}(2),...,s_{c,a_i}(t-1)$ 作为样本取值,取似然函数为 $L(\mu,\sigma^2)$:

$$L(\mu,\sigma^2)=\prod_{j=1}^{t-1}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(s_{c,a_i}(j)-\mu)^2}{2\sigma^2}}=(2\pi\sigma^2)^{-\frac{t-1}{2}}e^{-\frac{\sum_{j=1}^{t-1}(s_{c,a_i}(j)-\mu)^2}{2\sigma^2}} \tag{6}$$

两边取对数得到:

$$l(\mu,\sigma^2)=-\frac{t-1}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{j=1}^{t-1}(s_{c,a_i}(j)-\mu)^2 \tag{7}$$

将 $l(\mu,\sigma^2)$ 分别对 μ,σ^2 求偏导,并令它们都为 0,得到似然方程组:

$$\begin{cases}\dfrac{\partial l((\mu,\sigma^2)}{\partial\mu}=\dfrac{1}{\sigma^2}\sum_{j=1}^{t-1}(s_{c,a_i}(j)-\mu)^2=0\\ \dfrac{\partial l((\mu,\sigma^2)}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{\sigma^4}\sum_{j=1}^{t-1}(s_{c,a_i}(j)-\mu)^2=0\end{cases} \tag{8}$$

解方程组得:

$$s_{p,a_i}(t)=\hat{\mu}=\overline{S_c}=\sum_{j=1}^{t-1}s_{c,a_i}(j)\Big/t-1 \tag{9}$$

证毕. □

性质 2. 假设拟合周期序列 t,a_i 不发送数据的概率为 $P\left(\left|s_{s,a_i}(t-1)-s_{p,a_i}(t-1)\right|\leqslant\varepsilon\right)$, 则当 $s_{c,a_i}(t-1)=\mu$ 时, $P\left(\left|s_{s,a_i}(t-1)-s_{p,a_i}(t-1)\right|\leqslant\varepsilon\right)=\max\left\{P\left(\left|s_{s,a_i}(t-1)-s_{p,a_i}(t-1)\right|\leqslant\varepsilon\right)\right\}=2\Phi\left(\dfrac{\varepsilon}{\sigma}\right)-1$.

证明:由于 $\left|s_{s,a_i}(t-1)-s_{p,a_i}(t-1)\right|\leqslant\varepsilon\Leftrightarrow s_{p,a_i}(t-1)-\varepsilon\leqslant s_{s,a_i}(t-1)\leqslant s_{p,a_i}(t-1)+\varepsilon$ 且 $S_s\sim N(\mu,\sigma^2)$, 所以,

$$P(s_{p,a_i}(t-1)-\varepsilon\leqslant s_{s,a_i}(t-1)\leqslant s_{p,a_i}(t-1)+\varepsilon)=\Phi\left(\frac{s_{p,a_i}(t-1)+\varepsilon-\mu}{\sigma}\right)-\Phi\left(\frac{s_{p,a_i}(t-1)-\varepsilon-\mu}{\sigma}\right).$$

由于 $S_{s,a_i}(t)$ 的概率密度函数为偶函数,且在区间 $[\mu,+\infty)$ 单调递减,很明显,当 $\dfrac{s_{p,a_i}(t-1)-\mu}{\sigma}=0$ 时, $P(s_{p,a_i}(t-1)-\varepsilon\leqslant s_{s,a_i}(t-1)\leqslant s_{p,a_i}(t-1)+\varepsilon)$ 取最大值,此时, $s_{p,a_i}(t-1)=\mu$. □

特征相似度分析. 如图 5 所示,对于基站没有接收到感知数据的情况,由于 S_c 的估计 $\hat{\mu}$ 不准确,使得感知相似度落在 $\hat{\mu}$ 的 ε 邻域内的概率没有达到最高,即 $P\{x\in[\hat{\mu}-\varepsilon,\hat{\mu}+\varepsilon]\}\leqslant P\{x\in[\mu-\varepsilon,\mu+\varepsilon]\}$, 基于此,我们采用贪婪算法调整预测相似度 $s_{c,a_i}(t)$. 若 α 满足如下条件,则取 $\hat{\mu}+\alpha$ 作为当前的预测相似度.

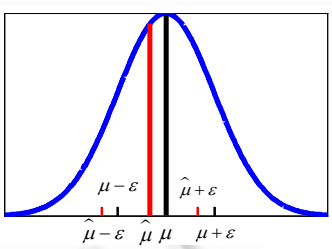

Fig.5 Prediction similarity analysis

图 5 预测相似度分析

对于任意的 α, 当满足 $\left|P(S_c\in[\hat{\mu}+\alpha-\beta,\hat{\mu}+\alpha])-P(S_c\in[\hat{\mu}+\alpha,\hat{\mu}+\alpha+\beta])\right|\leqslant\eta$ 时, $\hat{\mu}+\alpha$ 的取值为预测相似度 $s_{c,a_i}(t)$ 的优化值.其中, α 表示特征相似度 S_c 的估计值 $\hat{\mu}$ 动态调整的幅度.以 $\hat{\mu}+\alpha$ 为中心,当满足置信度差值 $\left|P(S_c\in[\hat{\mu}+\alpha-\beta,\hat{\mu}+\alpha])-P(S_c\in[\hat{\mu}+\alpha,\hat{\mu}+\alpha+\beta])\right|$ 尽量最大化时即为 β 的最优取值. η 表示当 $\hat{\mu}$ 偏离 μ 时,基站对置信度差值 $\left|P(S_c\in[\hat{\mu}+\alpha-\beta,\hat{\mu}+\alpha])-P(S_c\in[\hat{\mu}+\alpha,\hat{\mu}+\alpha+\beta])\right|$ 的最大容忍误差.

基于 PSO 的数据预测机制. 与单个数据的预测分析相比,相似度能够更好地反映数据的变化规律,所以我们将数据预测转化为一定精度下的相似度预测问题.但是,对于后端服务器,利用相似度不能唯一精确地确定预测曲线,求解预测数据集,基于此,我们采用 PSO 算法优化预测多项式曲线的系数,使得一定预测相似度条件下的预测数据取得理想的解集.

由 $s_{p,a_i}(t-1)=\int_a^b[f_{c,a_i}^{t-1}(x)-f_{p,a_i}^{t}(x)]\,\mathrm{d}x$, 且 $s_{p,a_i}(t-1)$ 为唯一确定的常数, $f_{c,a_i}^{t-1}(x)$ 在上一个拟合周期已经确定,此时, $f_{p,a_i}^{t}(x)$ 的求解变成了一个给定曲线积分取值求多项式曲线的问题.考虑到 $f_{p,a_i}^{t}(x),f_{c,a_i}^{t-1}(x)$ 都是 $k-1$ 次多项式函数,假设:

$$f_{c,a_i}^{t-1}(x)-f_{p,a_i}^{t}(x)=\lambda_0+\lambda_1x+\lambda_2x^2+\lambda_3x^3+...+\lambda_{k-1}x^{k-1},\ x\in[a,b] \tag{10}$$

$$s_{p,a_i}(t-1)=\int_a^b[\lambda_0+\lambda_1x+\lambda_2x^2+\lambda_3x^3+...+\lambda_{k-1}x^{k-1}]\,\mathrm{d}x \tag{11}$$

由于 $f_{c,a_i}^{t-1}(x)-f_{p,a_i}^{t}(x)$ 体现了感知数据与预测数据的差值,很明显 $\lambda_0,\lambda_1,...,\lambda_{k-1}$ 的解不唯一,且不同的 λ_0, $\lambda_1,...,\lambda_{k-1}$ 会改变感知数据与预测数据的差值影响预测精度.如何找到合适的 $\lambda_0,\lambda_1,...,\lambda_{k-1}$ 是求解预测数据的关键.假设当发送数据的相似度阈值为 ε 时,感知数据与预测数据的最大误差为 ε'(一般的$\varepsilon'=\varepsilon/k$), 那么优化 λ_0, $\lambda_1,...,\lambda_{k-1}$ 的目标是 $\max\{f_{c,a_i}^{t-1}(x)-f_{p,a_i}^{t}(x)\}\leqslant\varepsilon',x\in[a,b]$, 且尽量使得 $\max\{f_{c,a_i}^{t-1}(x)-f_{p,a_i}^{t}(x)\},x\in[a,b]$ 最小化.

考虑到 PSO 算法对参数优化具有稳定高效的收敛性能,我们采用 PSO 算法优化 $\lambda_0,\lambda_1,...,\lambda_{k-1}$ 的取值.文献[18]提出了一种优化参数的 PSO 算法,我们主要给出 PSO 算法的一些关键条件.

根据文献[18],首先确定种群数量 N, 假设 N 个粒子依次为 $p_1,p_2,...,p_N$, 由于待优化的解的参数为 k 个,所以粒子的维度可表示为 $v_i^k=[v_i^1,v_i^2,...,v_i^k]$, 位置向量可表示为 $x_i=[\lambda_0,\lambda_1,...,\lambda_{k-1}]=[x_i^1,x_i^2,...,x_i^k]$. 某个粒子 $p_i(i=1,...,N)$ 自身的历史最优用 $pBest_i$ 表示,所有粒子全局的历史最优解用 $gBest$ 表示.某个粒子 p_i 的适应度函数如(12)所示:

$$f(p_i)=\lambda_0+\lambda_1x+\lambda_2x^2+\lambda_3x^3+...+\lambda_{k-1}x^{k-1},\ x\in[a,b] \tag{12}$$

速度和位置更新方式如(13)所示,其中 ω 为惯量权重, $c_1,...,c_k$ 为加速系数,一般的取值 0.5, $r_i^d(i=1,...,k)$ 为[0,1]的随机数.

$$v_i^d=\omega\times v_i^d+c_1\times r_1^d\times(pBest_i^d-x_i^d)+...+c_k\times r_k^d\times(gBest^d-x_i^d),\ x_i^d=x_i^d+v_i^d \tag{13}$$

单个粒子的局部最优和所有粒子的全局最优分别如式(14)和式(15)所示:

$$f(pBest_i)=\min\{\lambda_0+\lambda_1x+\lambda_2x^2+\lambda_3x^3+...+\lambda_{k-1}x^{k-1}\},\ x\in[a,b] \tag{14}$$

$$f(gBest_i)=\min\{f(pBest_i)\},\ i=1,...,N \tag{15}$$

当满足 $f(gBest_i)\leqslant\varepsilon'$ 时,可以确定 $gBest_i$, 从而确定 $\lambda_0,\lambda_1,...,\lambda_{k-1}$ 的取值.由式(10)可以唯一确定预测曲线,求解预测数据集.

4 实验与仿真

本文采用 Matlab 作为实验平台,进行算法仿真实验.实验采用 2004-2-29~2004-3-31 英特尔伯克利研究实验室[19]采集的的固定采样周期为 30s 的真实数据进行算法仿真实验.数据库中包括温度、湿度、光强度和电压这 4 种传感器数据.本文重点对温度数据进行分析.由于任意一天的数据具有一定的周期性,我们重点对 2004-2-29, 00:00:00~23:59:59 的数据进行仿真实验.本实验主要包含两部分:有效性验证和能耗分析.第 1 部分分析预测相似度优化过程中的参数 α,β,η, 曲线拟合次数 k 等对算法性能的影响,验证算法的有效性以及估计误差.第 2 部分对比本文算法与 Model-Aided 算法在不同预测误差下预测数据的成功率,证明本算法在能耗等方面的优势.

4.1 有效性验证

本实验分析不同 β 取值时, η 对预测相似度调整的关键参数 α 的平均取值的影响,据根不同 k 的取值分析 k 对两种算法平均估计误差的影响,然后给出两种算法的预测数据分布样例.

图 6 描绘了用于调整的预测相似度取值的参数 β,η 对 $|\overline{\alpha}|$ 的影响,图中包含 $\beta=0.25,0.5,0.75,1$ 这 4 条折线,从图中可以看出,对于任意一条折线,随着 η 的增大, $|\overline{\alpha}|$ 的取值明显减小.同时,给定 η 的取值,随着 β 的增加, $|\overline{\alpha}|$ 的取值增加明显,但是随着 η 的增大, β 对 $|\overline{\alpha}|$ 的影响越来越小.

图 7 是拟合周期内样本数据数量 k 对预测数据的影响.从图 7 中可以看出,对于本文算法,当 $k(k<5)$较小时,

随着 k 的增加,预测误差增大,当 k=5 时,平均估计误差较低,而随着 k 的持续增大,预测性能急剧下降.对于 Model-Aided 算法,随着 k 的增加,估计误差直线增加.当 k=2 时,本文的预测误差与 Model-Aided 算法的预测误差相同,因为 Model-Aided 算法计算斜率与本文算法通过直线段之间的差的面积积分预测效果相同,但是随着 k 取值的增加,本文算法明显更优,这是由于本文的预测机制受到参与曲线拟合的样本数量的影响较小的缘故.

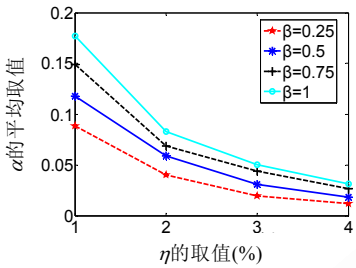

Fig.6 Relationship of key parameters

图 6 关键参数的关系

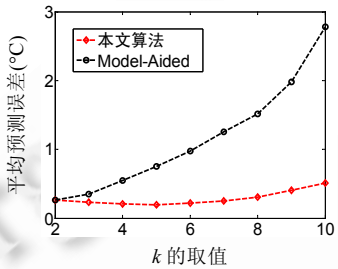

Fig.7 Effect of k on the average prediction error

图 7 k 对平均预测误差的影响

通过图 7 的分析,为了使得实际的算法效果可视化,如图 8 所示,我们随机选取 k=5(k=[2,3,…,10]),将本文算法与 Model-Aided 算法的估计值与实际采样值的在时间-温度的坐标图上进行对比.图 8 中方形标记为 Model-Aided 算法的预测值,菱形标记为本文算法的预测值,黑色实线为实际的节点采样值,从图 8 中可知,菱形标记整体明显更加靠近黑实线.与之对应的是 Model-Aided 算法的平均差值为 0.981 7,本文算法为 0.648,能够有效说明本文算法的精度相对于 Model-Aided 算法有一定的提高.

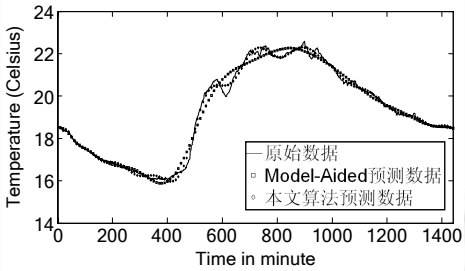

Fig.8 Raw data and forecast data

图 8 原始数据及预测数据

4.2 能耗分析

节点的能耗是体现本文算法以及 Model-Aided 算法的关键性能,为了增强本文与文献[15]的可对比性,本文与文献[15]采用同样的环境开展仿真实验工作.另外,本文直接对固定采样周期为 30s 的数据[19]进行算法仿真.相比于传输数据的能耗,算法计算的能耗可以忽略,所以我们重点考虑节点传输数据的能耗.与文献[20]相同,假设每一个测量数据或者每一个模型都存储在一个固定大小为 120 bit 的数据包中.每次数据传输或者模型更新的能耗为传输 120 bit 数据包的能耗.基于此,我们认为节点的能耗近似线性等价于节点发送数据包的数量.所以,我们重点对节点传输包的数量进行了分析.

通过上面的分析,我们得到了 $\eta=4\%,\beta=1$ 的最优取值,为了契合图 6 的数据,我们将 k=5 时本文算法与 Model-Aided 算法的发包数量进行对比.由于 Model-Aided 算法对单个感知数据进行预测,而误差阈值设计分别为[0.05,0.1,0.15,0.2],为了更好地对比两种算法的性能,本文选择 ε' 的取值分别为[0.05,0.1,0.15,0.2].本文算法对多个数据同时进行估计,通过上面的验证,当 $k=5$ 且拟合次数为 4 次多项式时,算法性能较好.由于 $\varepsilon=k\varepsilon'$,我们通过假设相似度误差分别为[0.05,0.1,0.15,0.2],然后对一天的数据预测性能进行分析.图 9~图 12 显示了误差阈值从 0.05℃增加到 0.2℃时,两种预测算法的发包数量的变化.

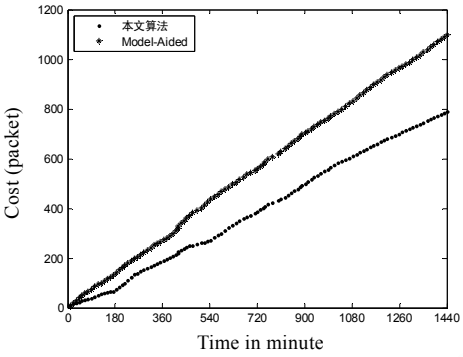

Fig.9 Analysis of energy consumption when $\varepsilon' = 0.05$

图 9 $\varepsilon' = 0.05$ 时的能耗分析

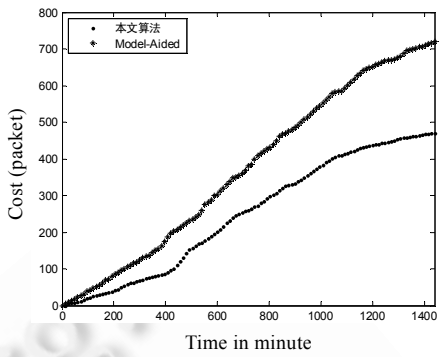

Fig.10 Analysis of energy consumption when $\varepsilon' = 0.1$

图 10 $\varepsilon' = 0.1$ 时的能耗分析

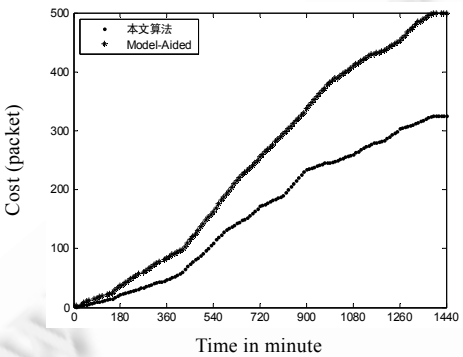

Fig.11 Analysis of energy consumption when $\varepsilon' = 0.15$

图 11 $\varepsilon' = 0.15$ 时的能耗分析

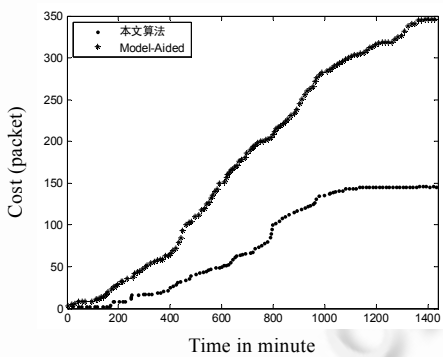

Fig.12 Analysis of energy consumption when $\varepsilon' = 0.2$

图 12 $\varepsilon' = 0.2$ 时的能耗分析

从图 9~图 12 中可以看出,随着误差阈值的增大,两种预测算法的发包数量呈减少的趋势.同时,我们可以明显地看出,在相同的误差阈值条件下,本文算法的发包数量明显少于 Model-Aided 算法.同时,当误差阈值分别为 0.05℃,0.1℃,0.15℃,0.2℃时,本文算法的预测成功率分别为 44.4%,67.4%,77.1%,90%,而 Model-Aided 算法的预测成功率分别为 21%,50%,65.3%,76%,体现出了本文算法性能的优势.

5 结束语

本文提出了一种基于相似度预测的 WSN 数据收集算法,将基于预测的数据收集问题转化为一定精度下的预测相似度的估计问题,采用贪婪算法动态地调整预测相似度.实验结果表明,该机制能够有效地提高准确预测数据的概率.同时,采用 PSO 算法实现基于预测相似度的预测数据推断.仿真实验结果表明,本文算法与 Model-Aided 等算法相比能够有效减少发包数量,提高能量的有效性.

References:

[1] Li JZ, GAO H. Survey on sensor network research. Journal of Computer Research and Development, 2008,45(1):1−15 (in Chinese with English abstract).

[2] Zheng J, Wang P, Li C. Distributed data aggregation using Slepian-Wolf coding in cluster-based wireless sensor networks. In: Proc. of the ICC 2007. IEEE Communications Society, 2007. 3616−3622.

[3] Al-Karaki JN, UI-Mustafa R, Kamal AE. Data aggregation and routing in wireless sensor networks: Optimal and heuristic algorithms. Computer Networks, 2009,53(7):945−960.

[4] Vaswani N, Lu W. Modified-CS: Modifying compressive sensing for problems with partially known support. IEEE Trans. on Signal Process, 2010,58(9):4595−4607.
[5] Angelosante D, Bazerque J, Giannakis G. Online adaptive estimation of sparse signals: Where RLS meets the l_1-norm. IEEE Trans. on Signal Process, 2010,58(7):3436−3447.
[6] Kopsinis Y, Slavakis K, Theodoridis S. Online sparse system identification and signal reconstruction using projections onto weighted l_1 balls. IEEE Trans. on Signal Process, 2011,59(3):936−952.
[7] Chu D, Deshpande A. Approximate data collection in sensor networks using probabilistic models. In: Proc. of the 22nd Int'l Conf. on Data Engineering. 2006. 48−53.
[8] Xiang M, Shi WR, Jiang CJ, Luo ZY. Energy saving algorithm based on cluster head prediction for wireless sensor networks. Computer Engineering, 2008,34(18):27−29 (in Chinese with English abstract).
[9] Luo C, Wu F, Sun J, Chen CW. Efficient measurement generation and pervasive sparsity for compressive data gathering. IEEE Trans. on Wireless Communications, 2010,9(12):3728−3738.
[10] Guo WZ, Xiong NX, Athanasios VV. Multi-Source temporal data aggregation in wireless sensor networks. Wireless Personal Communications, 2011,56(3):359−370.
[11] Rajagopalan R, Varshney P. Data aggregation techniques in sensor networks: A survey. IEEE Communications Surveys, 2006,8(4): 48−63.
[12] Kim H, Park J, Cho G. Statistical data aggregation protocol based on data correlation in wireless sensor networks. In: Proc. of the Int'l Symp. on Information Technology Convergence. 2007. 130−134.
[13] Min JK, Chung CW. EDGES: Efficient data gathering in sensor networks using temporal and spatial correlations. Journal of System and Software, 2010,83(2):271−282.
[14] Wei GY, Ling Y, Guo BF, Xiao B. Prediction-Based data aggregation in wireless sensor networks: Combining grey model and Kalman Filter. Computer Communications, 2011,34(6):793−802.
[15] Zhang CQ, Li ML, Wu MY. A model-aided data gathering approach for wireless sensor networks. Journal of Information and Engineering, 2007,23:1011−1022.
[16] Kim S, Fonseca R, Culler D. Reliable transfer in wireless sensor networks. In: Proc. of the 1st IEEE Int'l Conf. on Sensor and Ad Hoc Communications and Networks. 2004. 449−459.
[17] Li P, Yang W, Xie JY, Zhu HS, Zhang YG, Li XF. A study on mechanisms of secure data aggregation based on weighted fitting analysis. Ruan Jian Xue Bao/Journal of Software, 2013,24:108−116 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/13012.htm
[18] Li YL, Shao W, You L, Wang BZ. An improved PSO algorithm and its application to UWB antenna design. IEEE Antennas and Wireless Propagation Letters, 2013,12:1236−1239.
[19] Bodik P, Hong W, Guestrin C, *et al*. Intel lab data. 2004-06-02. http://db.csail.mit.edu/labdata/labdata.html
[20] Peng AP, Guo XS, Cai W, Xu XM. A data aggregation algorithm for clustering wireless sensor networks based on estimative scheme. Chinese Journal of Sensors and Actuators, 2011,24(1):128−133 (in Chinese with English abstract).

附中文参考文献:

[1] 李建中,高宏.无线传感器网络的研究进展.计算机研究与发展,2008,45(1):1−15.
[8] 向敏,石为人,蒋畅江,罗志勇.基于簇头预测的无线传感器网络节能算法.计算机工程,2008,34(18):27−29.
[17] 李平,阳武,谢晋阳,朱红松,张永光,李晓锋.基于加权拟合分析的 WSN 安全数据融合机制研究.软件学报,2013,24:108−116. http://www.jos.org.cn/1000-9825/13012.htm
[20] 彭爱平,郭晓松,蔡伟,徐晓淼.基于估计机制的分簇传感器网络数据融合算法.传感技术学报,2011,24(1):128−133.

李平(1972－),男,湖南新化人,博士,教授,CCF 会员,主要研究领域为物联网,信息安全,数据挖掘.
E-mail: 1lping9188@163.com

阳武(1990－),男,工程师,主要研究领域为物联网,信息安全,数据挖掘.
E-mail: 1013568049@qq.com

吴佳英(1977－),女,副教授,主要研究领域为无线网络,人工智能.
E-mail: jiayin528@163.com

胡海罗(1989－),男,工程师,主要研究领域为物联网,信息安全,数据挖掘.
E-mail: 332737434@qq.com