

基于聚类模型预测的无线传感网自适应采样技术研究

张美燕^① 蔡文郁^{*②} 周丽萍^①

^①(浙江水利水电学院电气工程学院 杭州 310018)

^②(杭州电子科技大学电子信息学院 杭州 310018)

摘要: 该文利用无线传感网(WSNs)的数据空间相关性,提出一种基于数据梯度的聚类机制,聚类内簇头节点维护簇成员节点的数据时间域自回归(AR)预测模型,在聚类内范围实施基于预测模型的采样频率自适应算法。通过自适应优化调整采样频率,在保证数据采样精度的前提下减少了冗余数据传输,提高无线传感网的能效水平。该文提出的时间域采样频率调整算法综合考虑了感知数据的时空联合相关性特点,仿真结果验证了该文算法的性能优势。

关键词: 无线传感网; 自适应采样; 模型匹配; 模型预测

中图分类号: TP393

文献标识码: A

文章编号: 1009-5896(2015)01-0200-06

DOI: 10.11999/JEIT140175

Clustered Predictive Model Based Adaptive Sampling Techniques in Wireless Sensor Networks

Zhang Mei-yan^① Cai Wen-yu^② Zhou Li-ping^①

^①(*Electrical Engineering Department, Zhejiang University of Water Conservancy and Electric Power, Hangzhou 310018, China*)

^②(*School of Electronics & Information, Hangzhou Dianzi University, Hangzhou 310018, China*)

Abstract: According to the data spatial correlation of Wireless Sensor Networks (WSNs), this study proposes a clustering mechanism based on the data gradient. In the proposed clustering mechanism, the cluster head nodes maintain Auto Regressive (AR) prediction model of the sensory data within each cluster in the time domain. Moreover, the cluster head nodes adjust the temporal sampling frequency based on the implementation of above predicted adaptive algorithm model. By adjusting the temporal sampling frequency, the redundant data transmission is reduced as well as ensuring desired sampling accuracy, so as energy efficiency is improved. The temporal sampling frequency adjustment algorithm takes into account spatial and temporal combined correlation characteristics of sensory data. As a result, the simulation results demonstrate the performance benefits of the proposed algorithm.

Key words: Wireless Sensor Networks (WSNs); Adaptive sampling; Model matching; Model forecasting

1 引言

近年来,随着微机电一体化、短距离无线通信、互连网络等技术的迅速发展,一种全新的信息获取和处理技术——无线传感网(WSNs)应运而生。无线传感网是由大量无处不在的、具有无线通信与计算能力的微小传感器节点构成的自组织分布式多跳通信网络系统,是能根据环境自主完成指定任务的“智能”协同系统^[1,2]。无线传感网实现了大量分布式时空数据的高保真采样,但在一般应用中传感器节点

的采样频率通常是固定的,而自适应采样技术的采样频率可根据被测对象的变化而变化,当观测对象变化趋势慢时降低采样频率,当观测对象变化趋势较快时提高采样频率。本文研究了一种基于聚类预测模型的无线传感网自适应采样技术,所有的传感器节点根据数据梯度关系划分成多个聚类,算法选举了各个聚类中的簇头节点。簇头节点根据感知数据的历史模型和当前采样数据进行判决,根据是否满足采样精度需求,进而利用“乘增半减”的采样频率更新策略调整所有簇成员节点的采样频率。如果聚类内传感采样数据的动态性比较低,则降低采样频率以减少冗余数据的产生与传输,如果传感采样数据的动态性比较高,则提高数据采样频率从而

2014-01-26 收到, 2014-05-26 改回

国家自然科学基金(61102067)和浙江省自然科学基金(Y15F030066)资助课题

*通信作者: 蔡文郁 dreampp2000@163.com

保证数据采样质量，由此实现网络能率高效与数据采集质量两者之间的均衡性。

目前相关研究工作大概都基于如下假设：(1)在无线传感网的实际应用场景中，传感器节点采集到的采样数据通常是连续的，从总体上表现出一定的连续性和稳定性；(2)受无线多跳通信不可靠、传感器节点失效以及节点能量的限制，感知数据的获取、处理与传输必然存在一定范围的误差值。因此，无线传感网所获取的数据集在时间域和空间域上都是无限的，实际上只能通过抽样获取所需的数据内容，而且所获得的数据并非实际的精确值，因此收集到的数据越多并不意味着结果越精确。基于上述这种思想，有一些研究者开启了无线传感网中数据采样优化技术的研究。无线传感网数据处理方法一般采用先进行网内数据处理然后再汇聚传输的方式以减少所需传输的数据量。目前网内数据处理可分为数据近似代替(data approximation)和数据汇聚处理(data aggregation)两种方法^[3]：数据近似代替通过对感知数据进行分布式建模，大量减少感知数据的传输量，从而延长网络生存周期。数据近似代替方法可基于不同模型：概率模型^[4]，时间序列分析模型^[5]，数据挖掘模型^[6]和数据压缩模型^[7]。数据汇聚处理使用 AVG, TOP, COUNT, SUM 等聚集函数来降低感知数据的传输量，但存在的最大问题是感知数据中大量的原始信息丢失，只能提供较为粗糙的统计结果量。

与本文想法较为相近的有以下研究：文献[8]提出传感器节点的采样时间间隔由 Sink 节点根据带宽来分配，通过 Kalman 滤波预测下一时刻采样值，如果预测值和真实值的误差大于预设阈值，则根据当前误差调整采样时间间隔，但该方法没充分利用感知数据的相关性；文献[9]提出了一种利用线性回归模型来动态调整采样频率的策略，在单个传感器节点上设置预测模型，根据实际采样数据期望值之间的差异大小来调整采样频率，但是该方法增加了节点之间的通信负荷；文献[10]提出了一种根据数据空间相关性选择聚集域中采样节点的方法，数据相关性较强的节点中，只有遴选出的某些特殊节点需要采集感知数据，因此降低了数据间的冗余，但是该方法并未考虑时间域采样频率的调整；文献[11]提出一种在 Nyquist 采样频率基础上的时间域采样频率的跟随调整，但是节点之间的额外通信所需耗费很多能量。文献[12]提出了一种基于预测模型的无线传感网内数据纠错方法，在时间域建立误差模型，以提高感知数据的可靠性。针对以上算法的不足之处，本文引入时间域数据预测模型技术，从而在分

布式聚类内实现自适应采样，提高网络能效水平。文献[13]创新地将自适应采样的算法应用于无线体感网(body sensor networks)中，考虑了无线体感网的特殊限制和需求。

2 预测模型匹配自适应采样算法

2.1 算法原理

本文所研究的基于感知数据预测模型的自适应采样技术中，首先将传感器节点根据数据梯度关系划分成多个聚类，各个聚类中的簇头节点根据感知数据的历史模型和当前采样数据的变化情况决定聚类内所有成员节点的采样频率(即下一采样时刻)，只有当真实值(以高频率采样的实测值为准)与预测值的误差大于预设阈值时，簇头节点才提高节点采样频率，实现网络能率高效及数据精度质量之间的均衡。当感知数据的动态性较高时，提高采样频率从而保证采样质量，当感知数据的动态性较低时，根据数据历史模型，仅发送那些必需发送的数据，降低采样频率以减少冗余数据的产生与传输。

本文算法的主要原理如下：每个聚类内的簇头节点根据聚类内成员节点的历史数据建立一个预测模型，用预测模型估计的估计值来代替采样数据，并用较低频率的采样数据来更新预测模型，如图 1 所示。

当采样数据与估计数据差较大时，逐步增大采样频率；否则，逐步降低采样频率。每个聚类内簇头节点的在线数据预测模型框架如图2所示。通过历史数据的模型特征提取建立预测树，当数据与预测模型不一致时进行采样频率的更新调整。

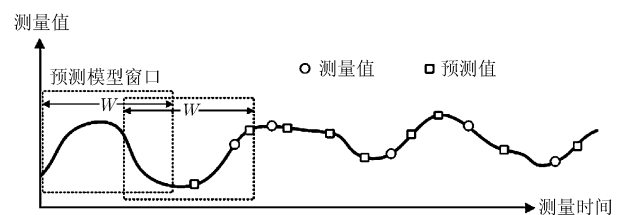


图 1 算法原理

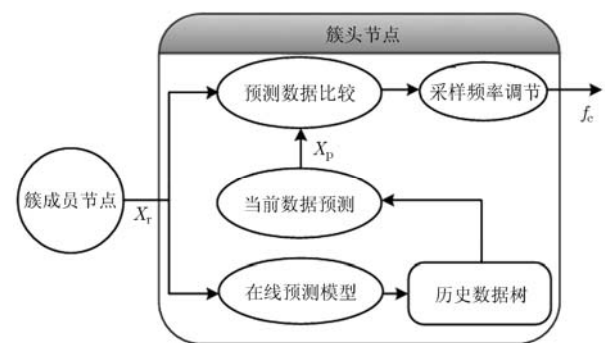


图 2 每个聚类内集中式预测模型

但是预测模型与特定的应用场景有着很强的相关性,而 Weiner 和 Kalman 滤波器等预测计算方法对传感器节点而言过于复杂。本文中各个聚类中簇头节点所采用的预测模型采用计算量较小的线性自回归(Auto Regressive, AR)模型,并采用递归最小二乘(LS)估计方法进行参数估计。如式(1)和式(2)所示:

$$X_p[n] = \sum_{i=1}^W a_i[n] \cdot X_r[n-i] \quad (1)$$

$$e[n] = X_r[n] - X_p[n] \quad (2)$$

其中 $X_p[n]$ 是预测数据值, $X_r[n]$ 是实测数据值, $e[n]$ 是误差值, $a_i[n]$ 是 AR 模型各阶系数。

2.2 数据相关性聚类机制

数据相关性聚类机制的主要思路是将数据相关性较强(体现为物理空间位置和数据均值较为接近)的传感器节点归在一个簇,每个聚类内的传感器节点进行集中式的数据模型预测,簇头节点负责维护每个聚类内的数据预测模型,减低需传输的冗余数据,提高能量使用效率。每个传感器节点成为簇头节点的概率计算如下:节点以自身剩余能量和其一跳邻居数(节点度数)的综合权值作为最大概率值来实施簇头节点的选举,则最大概率值公式为

$$\omega_i = p \frac{E(i) \times |N_i|}{\sum_{j \in N_i} E(j)} \quad (3)$$

式(3)中 p 代表簇头节点所占的比率, $p \in (0,1]$, 实际上簇头节点所占总节点数的上限比率一般设为 20%, 因此 $p \in (0,0.2]$; $E(i)$ 是节点 N_i 的剩余能量, $|N_i|$ 是节点 N_i 的度数(包含自身)。

一旦传感器节点被选举为簇头节点,立即在其一跳邻居节点内广播簇头标志信息,诱导非簇头节点加入自身簇。非簇头节点选择最优簇头节点的思想如下:选择簇头节点与簇成员节点之间数据梯度值最小的一跳邻居簇头作为簇头节点。如图 3 所示。

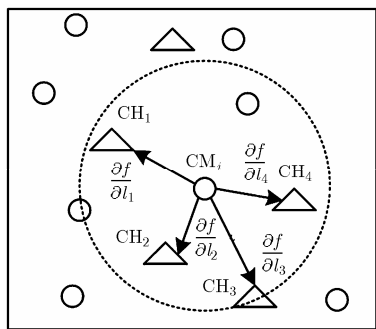


图3 基于数据梯度的簇头节点遴选示意图

假设非簇头节点 CM_i 有 4 个簇头节点 (CH_1, CH_2, CH_3, CH_4) 可供选择,各个方向的数据梯度值如式(4)所示。这部分的研究可以参考作者已发表的会议论文^[14]。

$$\left. \begin{aligned} \frac{\partial f}{\partial l_1} \Big|_{CM_i} &= \frac{|\bar{E}_{CM_i} - \bar{E}_{CH_1}|}{d(CM_i, CH_1)} \\ \frac{\partial f}{\partial l_2} \Big|_{CM_i} &= \frac{|\bar{E}_{CM_i} - \bar{E}_{CH_2}|}{d(CM_i, CH_2)} \\ \frac{\partial f}{\partial l_3} \Big|_{CM_i} &= \frac{|\bar{E}_{CM_i} - \bar{E}_{CH_3}|}{d(CM_i, CH_3)} \\ \frac{\partial f}{\partial l_4} \Big|_{CM_i} &= \frac{|\bar{E}_{CM_i} - \bar{E}_{CH_4}|}{d(CM_i, CH_4)} \end{aligned} \right\} \quad (4)$$

式(4)中 $\frac{\partial f}{\partial l_j} \Big|_{CM_i}$ 代表了从 CM_i 到 CH_j 的数据梯度

值, $d(CM_i, CH_j)$ 代表了从 CM_i 到 CH_j 的欧式物理距离, \bar{E}_{CM_i} 和 \bar{E}_{CH_j} 代表了数据窗口内的数据序列的均值。簇成员节点选择数据梯度值最小的簇头节点作为簇头,如式(5)所示:

$$\begin{aligned} CH &= \left\{ CH_j \mid \frac{\partial f}{\partial l_j} \Big|_{CM_i} \right. \\ &= \left. \min \left\{ \frac{\partial f}{\partial l_1} \Big|_{CM_i}, \frac{\partial f}{\partial l_2} \Big|_{CM_i}, \frac{\partial f}{\partial l_3} \Big|_{CM_i}, \frac{\partial f}{\partial l_4} \Big|_{CM_i} \right\} \right\} \quad (5) \end{aligned}$$

综上所述,本文提出的聚类机制充分考虑了感知数据相关性以及传感器节点空间位置,形成的簇分布能够体现感知数据的空间关联性。

2.3 数据预测模型及方法

根据平稳数据预测精度和速度的需求,本文采用二阶 AR 预测模型,预测值 \hat{X}_{i+1} 如式(6)如下:

$$\hat{X}_{i+1} = \alpha(X_i - \bar{X}) + \beta(X_{i-1} - \bar{X}) + e_{i+1} \quad (6)$$

其中 X_i 为真实值, α 和 β 为 AR 模型的参数, e_{i+1} 为预测误差, \bar{X} 是 W 个历史观测数据的均值。对于给定的 W 个历史数据,参数 α 和 β 可由样本自相关函数计算得出。

$$\left. \begin{aligned} \alpha &= [\rho_1(1 - \rho_2)] / (1 - \rho_1^2) \\ \beta &= (\rho_2 - \rho_1^2) / (1 - \rho_1^2) \end{aligned} \right\} \quad (7)$$

而样本自相关函数的计算可由式(8)得到:

$$\rho_k = \frac{\sum_{i=1}^{N-k} (X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2} \quad (8)$$

可见,预测误差可以由预测值和真实值之差得到:

$$e_{i+1} = \hat{X}_{i+1} - X_{i+1} \quad (9)$$

本文中的数据预测模型没有采用定期更新的机

制，因为定期更新会引入不必要的更新次数，只有当误差大于一定阈值时才启动重新计算二阶 AR 系数的过程。AR 模型参数估计采用了常用的 Yule-Walker 方程法^[15]。

2.4 采样频率更新机制

采样频率的自适应调整采用加增半减的方法来进行调整：当采样误差小于预设阈值时，减半聚类中各个传感器的采样频率，否则增加采样频率，增加的步进值为上限阈值和下限阈值的 M 分之一。最后限制调整过的采样频率在上限阈值和下限阈值范围之内。为了衡量误差的精度，预测误差采用了平均相对误差 (Mean Relative Error, MRE) 指标来表征预测数据偏离实际数据的程度。

$$\text{MRE} = \frac{1}{W} \sum_{i=1}^W \left| \frac{\hat{x}_i - x_i}{x_i} \right| \quad (10)$$

其中 W 为数据窗口长度，本文根据所采用数据集的采样频率属性，各个参数值设置为

$$\text{Th}_{\min} = 1 \text{次}/10 \text{min} = 0.1 \times \frac{1}{60} \text{Hz}$$

$$\text{Th}_{\max} = 1 \text{次}/1 \text{min} = 1 \times \frac{1}{60} \text{Hz}$$

$$M = 9, \theta_{\text{MRE}} = 5\%$$

$$f_i = 1 \text{次}/1 \text{min} = 1 \times \frac{1}{60} \text{Hz}$$

采样频率更新策略如表 1 所示。

表 1 采样频率更新策略

#基于预测模型的聚类内采样频率调整
Begin
#CH广播初始化采样频率 f_i ;
While(每一个采样间隔 T)
{
#CH接收节点数据，进行预测模型比对;
if(平均相对误差 $\text{MRE} \leq$ 阈值 θ_{MRE})
$f_c = \frac{1}{2} \times f_c$
else
$f_c = f_c + \beta \quad \beta = \frac{\text{Th}_{\max} - \text{Th}_{\min}}{M}$
#采样频率阈值限制;
if($f_c \geq \text{Th}_{\max}$) $f_c = \text{Th}_{\max}$;
if($f_c \leq \text{Th}_{\min}$) $f_c = \text{Th}_{\min}$;
#CH广播调整后采样频率 f_c ;
}
End

2.5 频率更新机制分析

采样频率的自适应调整由各个聚类中的簇头节点管理，簇头节点在接收到采样数据后将下一个采样频率附带在确认数据包中发送给各个簇成员节

点。由于簇头节点接收到感知数据后也会发送确认数据包，因此采样频率的更新并不花费额外的通信负荷。各个聚类内数据采集的频率不一样，但是簇头节点可以通过模型预测的方法以统一的频率将数据上传到汇聚节点。

3 仿真

本文采用 Berkeley 大学的 Intel 实验室^[16]所做的真实采样数据，这份数据来自于 54 个传感器，有温度、湿度、亮度以及电压等读数，大约每隔 30 s 从 54 个传感器读取一次数据，共有一个月的数据，本文仿真只采用了其中温度值数据。本文定义了不同滑动窗口长度的数据序列相关性系数的量化公式：

$$C_{uv} = \frac{\left| [\mathbf{D}_u - \bar{E}(\mathbf{D}_u)] \times [\mathbf{D}_v - \bar{E}(\mathbf{D}_v)]^T \right|}{W \times \sqrt{\text{Var}(\mathbf{D}_u)} \times \sqrt{\text{Var}(\mathbf{D}_v)}} \quad (11)$$

式(11)中 C_{uv} 是窗口长度为 W 的数据序列 \mathbf{D}_u 和 \mathbf{D}_v 的相关性系数， $\bar{E}(\mathbf{D}_u)$ 和 $\text{Var}(\mathbf{D}_u)$ 分别表示数据序列 \mathbf{D}_u 的均值和均方差。 $C_{uv} \in [-1, 1]$ ，其中 $C_{uv} = -1$ 代表数据序列 \mathbf{D}_u 和 \mathbf{D}_v 负相关， $C_{uv} = 0$ 代表数据序列 \mathbf{D}_u 和 \mathbf{D}_v 不相关， $C_{uv} = 1$ 代表数据序列 \mathbf{D}_u 和 \mathbf{D}_v 正相关。以一天数据量为窗口长度 ($W=1100$)，节点 1 与节点 2 的相关性系数为 0.9393，节点 1 与节点 16 的相关性系数为 0.7199，节点 1 与节点 2 的相关性明显高于节点 1 与节点 16 的相关性系数。

当簇头比例分别为 $p = 0.05$ 和 $p = 0.15$ 时，数据相关性聚类结果如图 4 所示，不同聚类采用不同形状 of 节点集表示，其中标志节点号的为每个簇相应的簇头节点。

如图 5 所示，可以看到节点 1 在 2004 年 3 月 1 日~3 月 6 日的温度数据变化趋势在时间维度上传感器节点的数据呈现很强的重复性，因为在时间维度上可以利用各个聚类内传感器节点采样频率的自适应调整进行网络能耗优化。当采样频率分别为 $\text{Th}_{\min} = 1 \text{次}/10 \text{min} = 0.1 \times \frac{1}{60} \text{Hz}$ 时，节点 2 在 2004 年 3 月 1 日的实测数据、预测数据和估计误差分别如图 6 所示。

假设节点能耗主要考虑由采样过程所产生，因此网络整体能耗体现在传感器数据的采样频率上，数据采样频率越高，网络能耗越大。在平均相对误差 $\theta_{\text{MRE}} = 5\%$ 和簇头比例 $p = 0.2$ 的条件下，利用数据相关性聚类机制获得的聚类划分结果为 9 个聚类，每个时刻所有聚类的采样频率平均值如图 7 所示。由图 7 可以发现，相比固定采样频率方式(采样频率为 $\text{Th}_{\min} = 1 \text{次}/1 \text{min} = 1 \times \frac{1}{60} \text{Hz}$)，本文提出的基于聚类模型预测的自适应采样技术可以降低网络整体能量消耗，多次仿真所获取的平均采样频率可

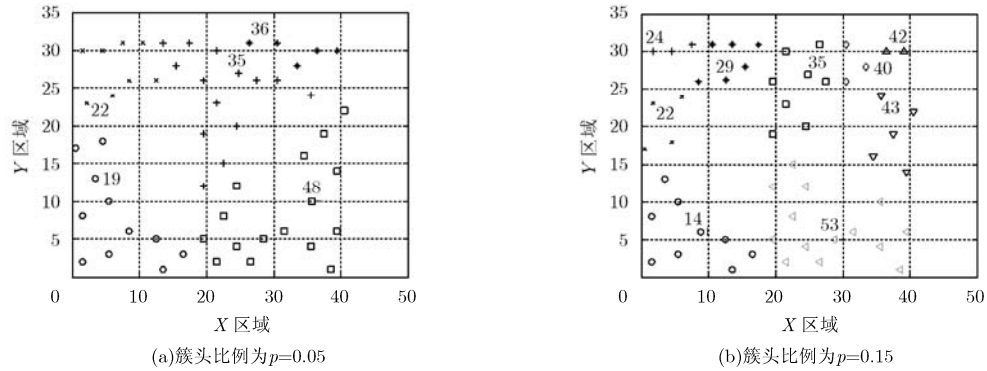


图 4 数据相关性聚类划分结果

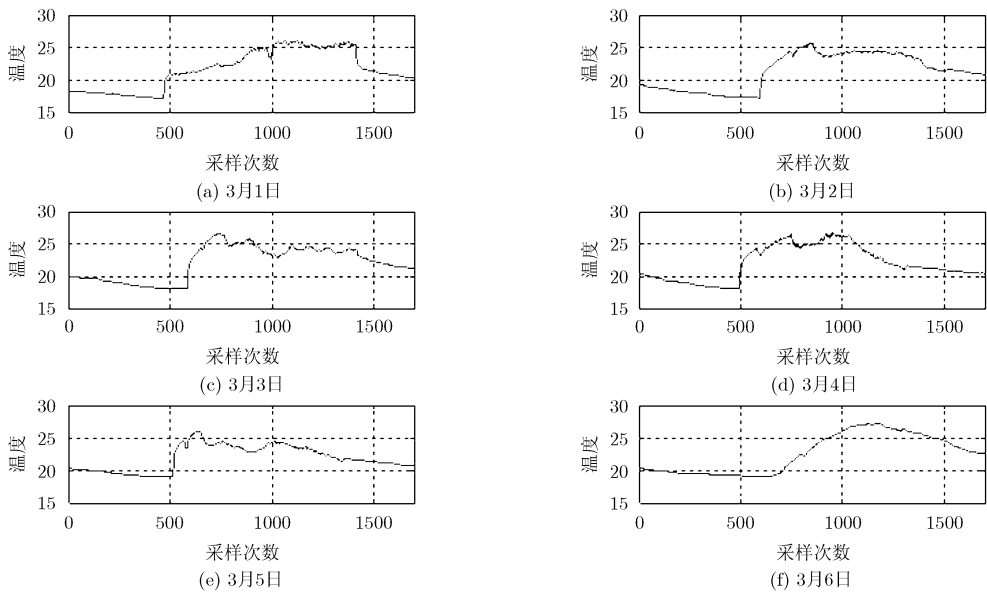


图 5 节点 1 在不同天的温度数据变化趋势

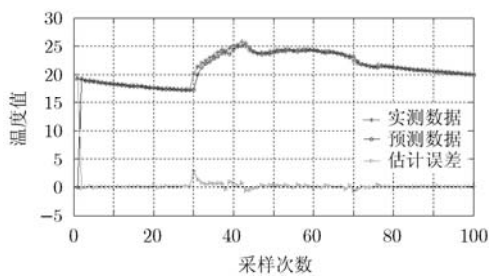


图 6 AR 预测误差分析

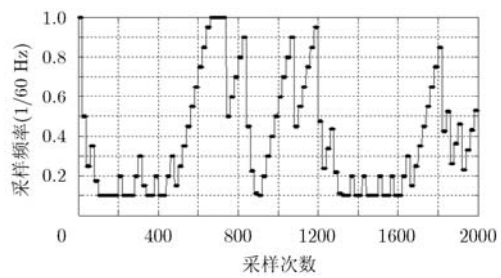


图 7 采样频率自适应调整值

达 $\bar{f} = 0.52 \times \frac{1}{60}$ Hz，也就是说能够获取将近一半的能率提高。

4 结束语

本文利用无线传感网的数据空间相关性，提出了一种基于数据梯度的聚类机制，簇头节点维护簇成员节点的数据时间域 AR 预测模型，在聚类范围内实施基于预测模型的采样频率自适应调整算法。

通过自适应优化调整采样频率，提高无线传感网的能耗水平：如果感知数据的动态性较高，提高采样频率从而保证采样数据质量，如果感知数据的动态性较低，仅仅发送那些必需发送的数据，降低采样频率以减少冗余数据的产生与传输。本文属于一种考虑空间相关性的时间域采样频率调整算法，由于综合考虑了感知数据的时空联合相关性等特点，因此在提高能耗水平方面具有较大优势。

参 考 文 献

- [1] Li J Z and Gao H. Survey on sensor network research[J]. *Journal of Computer Research and Development*, 2008, 45(1): 1-15.
- [2] 蔡文郁, 张美燕, 蒋一波. 基于时空联合性的无线传感网覆盖采样技术[J]. *传感技术学报*, 2013, 26(2): 260-265.
Cai Wen-yu, Zhang Mei-yan, and Jiang Yi-bo. Coverage sampling technology based on spatial temporal joint union for wireless sensor networks[J]. *Chinese Journal of Sensors and Actuators*, 2013, 26(2): 260-265.
- [3] Deligiannakis A, Kotidis Y, and Roussopoulos N. Dissemination of compressed historical information in sensor networks[J]. *The VLDB Journal*, 2007, 16(4): 439-461.
- [4] Masoum A, Meratnia N, and Havinga P J M. An energy-efficient adaptive sampling scheme for wireless sensor networks[C]. The 8th IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Melbourne, Australia, 2013: 231-236.
- [5] Tulone D and Madden S. PAQ: time series forecasting for approximate query answering in sensor networks[C]. Proceedings of the 3rd European conference for Wireless Sensor Networks (EWSN 2006), Zurich, Swiss, 2006: 21-37.
- [6] Sachidananda V, Khelil A, Noack D, *et al.* Information quality aware co-design of sampling and transport in wireless sensor networks[C]. The 6th Joint IFIP Wireless and Mobile Networking Conference (WMNC 2013), USA, 2013: 1-8.
- [7] Zhou S W, Lin Y P, Zhang J M, *et al.* A wavelet data compression algorithm using ring topology for wireless sensor networks[J]. *Journal of Software*, 2007, 18(3): 669-680.
- [8] Jain A and Chang E Y. Adaptive sampling for sensor networks[C]. Proceedings of the 1st International Workshop on Data Management for Sensor Networks (DMSN'04), Toronto, Canada, 2004: 10-16.
- [9] Li Jin-bao and Li Jian-zhong. Data sampling control, compression and query in sensor networks[J]. *International Journal of Sensor Networks*, 2007, 2(1): 53-61.
- [10] Aplippi C, Anastasi G, Francesco M D, *et al.* An adaptive sampling algorithm for effective energy management in wireless sensor networks with energy-hungry sensors[J]. *IEEE Transactions on Instrumentation and Measurement*, 2010, 59(2): 335-344.
- [11] Gedik B, Liu L, and Yu P S. ASAP: an adaptive sampling approach to data collection in sensor networks[J]. *IEEE Transactions on Parallel Distributed Systems*, 2007, 18(12): 1766-1783.
- [12] Mukhopadhyay S, Schurgers C, Panigrahi D, *et al.* Model-based techniques for data reliability in wireless sensor networks[J]. *IEEE Transactions on Mobile Computing*, 2009, 4(8): 528-542.
- [13] Qi Xin, Keally M, Zhou Gang, *et al.* AdaSense: adapting sampling rates for activity recognition in body sensor networks[C]. IEEE 19th Real-Time and Embedded Technology and Applications Symposium (RTAS), Philadelphia, USA, 2013: 163-172.
- [14] Zhang Mei-yan, Cai Wen-yu, and Zhou Li-ping. A sensing data Ddriven clustering algorithm for adaptive sampling in wireless sensor networks[C]. 2012 International Applied Mechanics, Mechatronics Automation Symposium (IAMMAS2012), Shenyang, China, 2012: 748-752.
- [15] Da Silva S, Dias J Nior M and Lopes Junior V. Damage detection in a benchmark structure using AR-ARX models and statistical pattern recognition[J]. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 2007, 29(2): 174-184.
- [16] Samuel Madden: Intel Berkeley Research Lab[OL]. <http://www.intel-research.net/berkeley/index.asp>. 2014.1.
- 张美燕：女，1983年生，讲师，研究方向为无线传感器网络、新型能源技术、物联网技术。
- 蔡文郁：男，1979年生，副教授，研究方向为无线通信、物联网、无线传感网及嵌入式技术。
- 周丽萍：女，1965年生，讲师，研究方向为无线传感器网络技术。