November 2007

A bayesian multilevel modeling approach for data query in wireless sensor networks

H. Wang

H. Fang

K. A. Espy

D. Peng

H. Sharif

A Bayesian Multilevel Modeling Approach for Data Query in Wireless Sensor Networks

Honggang Wang[1], Hua Fang[2], Kimberly Andrew Espy[2], Dongming Peng[1], and Hamid Sharif[1]

[1] Department of Computer and Electronics Engneering,University Of Nebraska Lincoln, Omaha, USA,68124
{hwang, dpeng, hsharif}unlnotes.unl.edu
[2] Office of Research, University of Nebraska Lincoln, Lincoln, USA,68588
{jfang2,kespy2}.unl.edu

Abstract. In power-limited Wireless Sensor Network (WSN), it is important to reduce the communication load in order to achieve energy savings. This paper applies a novel statistic method to estimate the parameters based on the real-time data measured by local sensors. Instead of transmitting large real-time data, we proposed to transmit the small amount of dynamic parameters by exploiting both temporal and spatial correlation within and between sensor clusters. The temporal correlation is built on the level-1 Bayesian model at each sensor to predict local readings. Each local sensor transmits their local parameters learned from historical measurement data to their cluster heads which account for the spatial correlation and summarize the regional parameters based on level-2 Bayesian model. Finally, the cluster heads transmit the regional parameters to the sink node. By utilizing this statistical method, the sink node can predict the sensor measurements within a specified period without directly communicating with local sensors. We show that this approach can dramatically reduce the amount of communication load in data query applications and achieve significant energy savings.

Keywords: Bayesian Multilevel Modeling, Wireless Sensor Network.

1 Introduction

In most WSN applications, the typical scenario is to collect and transmit the measured data from each sensor to the centralized sink where the data will be processed and analyzed. However, sensor nodes might be far away from the sink and have to send tremendous real-time data by multiple hops to the sink, which consume significant energy resources. Therefore, to save energy is to reasonably reduce the communication load from the local sensors to the sink.

Statistical modeling techniques have been applied to sensor network query systems [1-3]. However, these studies did not support data queries with specified error bound or clustering structure. Also, they undergo a heavyweight learning phase. Autoregressive multilevel Bayesian models have been widely used outside the

Y. Shi et al. (Eds.): ICCS 2007, Part III, LNCS 4489, pp. 859–866, 2007.

wireless sensor network domain as a way to approximate and summarize time series in many application domains such as finance, communication, weather prediction [14-15]. In this paper, we applied the multilevel Bayesian statistical model to predict sensor values based on multilevel clustering architecture instead of transmitting the real time data directly to sink by each sensor. These techniques take advantages of the recent historical readings to predict the most likely future values. It can drastically reduce the amount of communication from sensors to the sink, detect the abnormal data, and accommodate missing sensor data.

Clustering techniques have also been used in WSN. Many clustering techniques such as K-mean, C-mean, or hierarchical clustering [4-8] have been proposed to improve network performance and save energy in WSN. We propose a query-based two-level clustering structure with consideration of both temporal and spatial correlation, which matches the generic WSN topology. In the following sections, we first present two-level network architecture and discuss the data query in section II. A detailed multilevel Bayesian modeling approach to WSN data query is discussed in section III. We demonstrate the advantages of our approach by the simulation in section IV. Conclusions are reached in the last section.

2 Two Level Network Architecture and Data Query

Hierarchical (clustering) techniques can aid in reducing useful energy consumption [4]. In our proposed hierarchical network structure, the sensor with the highest number of neighbors was selected as the temporary cluster center. Other sensors within a defined radius are then removed and the algorithm looks for a new sensor with the highest number of neighbors. This continues until most sensors are clustered. In our algorithm, the sensor in the cluster with the highest remaining energy is selected as the cluster head. Once the selected cluster heads run out of battery, the new cluster heads will be selected. By this approach, the network is formed into a two-level network architecture. Each sensor joins a local cluster group, forming the level-1(i.e., the sensor level) structure; all the cluster heads form the second tier multi-hop network structure at the cluster level. In this two-level clustering-based network structure, the typical data query application scenario is described as follows: When users submit a query to the sink, each sensor at level-1 senses the local phenomena, sending the sample data to the cluster head. At level-2, the cluster heads summarize these local data, sending them to the sink by one hop or multiple hops. However, in our approach, local sensors and cluster heads only transmit Bayesian model parameters inferred from the historical data instead of transmitting the real-time readings to the sink. All user queries can be answered at the sink within the specified time interval.

Our two level WSN model consists of a dynamic set of sensors denoted by S, and one sink node. This set of sensors form different clusters $\{S_1, S_2 \ldots\ldots S_n\}$ and all clusters have dynamic cluster heads $\{C_{s1}, C_{s2} \ldots\ldots C_{sn}\}$ by the algorithm we discussed above. Each sensor senses and performs readings on M physical phenomena metrics

$\{M_1, M_2 .....M_n\}$ over time. We assume that each sensor performs a reading on each M_i every T time units. Queries are executed at the sink. The typical query forms are designed as follows:

$$SELECT \quad Sensors \ WHERE \ R(M1, M2....Mn) \ ERROR \ X \ CONFIDENCE \ d\% \ Where \ REGION = \text{Re}\, gion1$$

Where $R(M1, M2....Mn)$ predicted the values of $M1, M2....Mn$ based on the multilevel modeling. X represents an error bound required by the user in the query. The d% is confidence ratio that denotes at least of d% the readings should be within X of their true value, and REGION gives geographical location restrictions of sensor groups.

3 Bayesian Multilevel Modeling in WSN

In this paper, the Bayesian multilevel modeling approach is applied for this two-level generic WSN architecture. The time series measurement model is at level-1 and the Bayesian parameters are transmitted to its cluster head. All cluster heads collect these parameters, inferring the level-2 Bayesian parameters at the cluster level and transmitting them to the sink. When users submit a data query, the sink predictor can answer it within the specified time period.

The level-1 model is expressed as

$$Y_{tij}^{L1} = \beta_{0ij} + \beta_{1ij}T + \beta_{2ij}T^2 + e_{tij}, \quad e_{tij} \sim N(0, \Sigma) \tag{1}$$

where Y_{tij}^{L1} denotes the level-1 (L1) measurement outcomes (e.g., temperature or humidity) at time t for senor i in cluster j; $\beta_{0\mathrm{i\,j}}$ is the initial status of sensor i of cluster j; $\beta_{1\mathrm{ij}}$ and $\beta_{2\mathrm{ij}}$ denote the change rates and acceleration rates associated with time T and quadratic term T^2, respectively. The level-1 errors, e_{tij}, are normally distributed with mean of 0 and covariance matrix Σ under first-autoregressive assumption (AR(1)) which consists of variance, σ^2, and covariance of

$$Cov(e_{tij}, e_{t'ij}) = \sigma^2 \rho^{|t-t'|} \tag{2}$$

where $|t - t'|$ is the lag between two time points; ρ is the auto-correlation and σ^2 is the level-1 variance at each time point. In Bayesian notation, the observer data, Y are distributed according to $f(Y \mid B, \Sigma)$, where f is the normal density, B denotes the β parameters. The outcomes Y_{tij}^{L1} are assumed independently normally distributed with mean of

$$E(Y_{tij}^{L1} \mid B, \Sigma) = \beta_{0ij} + \beta_{1ij}T + \beta_{2ij}T^2 \tag{3}$$

and the covariance matrix Σ. The level-2 model is expressed as

$$\mathrm{B}^{\mathrm{L2}} = \begin{pmatrix} \beta_{0ij} \\ \beta_{1ij} \\ \beta_{2ij} \end{pmatrix} = \begin{pmatrix} \gamma_{00j} & \gamma_{01j} & \gamma_{02j} \cdots & \gamma_{0qj} \\ \gamma_{10j} & \gamma_{11j} & \gamma_{12j} \cdots & \gamma_{1qj} \\ \gamma_{20j} & \gamma_{21j} & \gamma_{22j} \cdots & \gamma_{23j} \end{pmatrix} \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ \cdots \\ X_q \end{pmatrix} + \begin{pmatrix} u_{0ij} \\ u_{1ij} \\ u_{2ij} \end{pmatrix} \tag{4}$$

In Bayesian notation, this specifies the prior $p(B^{L2} \mid \Lambda, G)$ where B^{L2} are the level-2 (L2) outcomes, containing the same β parameters (3×1) as shown in level-1 model, representing the initial status, linear change rate and acceleration (or deceleration) rate of individual sensor *i* of cluster j; Λ is a (3×q) matrix of γ parameters, representing the average initial status (e.g., the initial temperature or humidity) (γ_{00j}), linear change rates (γ_{10j}) and the acceleration rates (γ_{20j}) of cluster j, as well as other γ parameters associated with level-2 *q*×1 predictors (X) , collected by cluster head j; *u* denotes level-2 random effects (or random errors), multivariately and normally distributed with a mean vector of 0 and G covariance matrix.

The Bayesian method requires to know the joint distribution of the data Y and unknown parameters, θ, which denotes both fixed coefficients γ and covariance matrix ψ (including G and Σ) in our study. The joint distribution can be written as:

$$P(Y, \theta) = P(\theta) P(Y \mid \theta) \tag{5}$$

where P (θ) is called the prior and P (Y| θ) is called the likelihood. As we observed the data Y, Bayes' Theorem was used to get the posterior distribution as follows:

$$P(\theta \mid Y) = \frac{P(\theta) P(Y \mid \theta)}{\int P(\theta) P(Y \mid \theta) d\theta} \tag{6}$$

specifically,

$$P(\gamma, \psi \mid Y) = \frac{f(Y \mid \gamma, \psi) P(\gamma \mid \psi) P(\psi)}{\int\int f(Y \mid \gamma, \psi) P(\gamma \mid \psi) P(\psi) d\gamma d\psi} \tag{7}$$

As the parameters γ are of primary interest, we have

$$P(\gamma \mid Y) = \int p(\gamma, \psi \mid Y) d\psi \tag{8}$$

In general, analytically performing the above integration has been a source of difficulty in application of Bayesian inference and often Markov Chain Monte Carlo (MCMC) simulation is one way to evaluate the integrals. In this study, we used one of MCMC procedures, Metropolis-Hastings sampling procedure, to implement this approximation [16-18].

4 Simulation and Analysis

We used SAS software [10] to simulate and test our approach. Our simulation was based on 50 random deployed sensors. With our clustering algorithm, all sensors form Cluster A and B. Cluster A has 20 sensors deployed while Cluster B has 30 sensors. The temperature data were collected at different clusters across different areas with a significant temperature difference. In our simulation, we used the first order radio model presented in [4]. In the specified radio model, the radio dissipates E_{elec} = 50 nJ/bit to run the transmitter or receiver circuitry and E_{amp} = 100 pJ/bit/m2 for the transmit amplifier. To transmit a k-bit message a distance d meters, E_{Tx} was used by sensors. To receive a message, the sensors spent E_R.

$$E_{tx}(k,d) = E_{elec} \cdot k + E_{amp} \cdot k \cdot d^2 \tag{9}$$

$$E_{rx}(k,d) = E_{elec} \cdot k \tag{10}$$

After the clusters were formed and cluster heads were selected, the sink calculated the routing hops among cluster heads. In addition, an index matrix was created for time, area and sensor IDs. The two measured areas represented by the two sensor class heads were coded as 0 and 1, respectively. Individual sensors (IDs) were considered nested within each cluster represented by corresponding cluster heads, for instance, sensor IDs ranged from 1 to 20 for Class Head 1, and 21 to 50 for Class Head 2. Time started from 0 and extended to the assumed 14.5 hours with 0.5 hour interval. Based on Model (4), a univariate response vector of y_{ti} was created. For example, each sensor might have had 30 half-hour time points and one cluster had 20 sensors while the other had 30 sensors. The data generator [11-12] was validated with parameter estimates from Potthoff and Roy's data[13]. Table 1 presents partial local parameters generated by each sensor at level-1, to be transmitted to the cluster heads.

Table 1. Selected Model Parameters at Sensor Level

Parameters	Sensor ID	Estimates	Parameters	Sensor ID	Estimates
Intercept	5	69.59662	Intercept	23	79.86074
Slope	5	0.307631	Slope	23	0.590479
Acceleration/ Deceleration	5	-0.00325	Acceleration/ Deceleration	23	-0.00355
Intercept	6	69.50935	Intercept	24	80.6984
Slope	6	0.403908	Slope	24	0.348969
Acceleration/ Deceleration	6	-0.00203	Acceleration/ Deceleration	24	-0.00375
		...			

Similarly, Table 2 shows the level-2 Bayesian model parameters based on the local colleted data, to be transmitted to the sink. The parameters β_0, β_s, and β_a represent the initial temperature, linear change rate and deceleration rate at the two areas, respectively. Based on these parameters, the sink predicts the next half hour temperature value.

Table 3 gives partial predicted temperatures at the sink with error bound and confidential interval, which responds to the queries submitted by the user at the sink.

Table 2. Model Parameters at Cluster Level

	Cluster Head 1			Cluster Head 2		
	β	SE	95% CI	β	SE	95% CI
β_0	69.980	0.128	(69.729, 70.231)	80.187	0.109	(79.973, 80.401)
β_s	0.307	0.025	(0.258, 0.356)	0.448	0.024	(0.401, 0.495)
β_a	-0.003	0.001	(-0.00496, -0.00104)	0.003	0.001	(0.001, 0.005)

Table 3. Selected Predicted Values with Error Bounds at Sink

Region	Time (hour)	Cluster	Predicted Temperature	SE	95% Confidence Interval	
					Lower Bound	Higher Bound
			...			
0	8	1	74.15	0.0773	74.00	74.30
0	8.5	1	74.30	0.1164	74.08	74.53
0	9	1	74.56	0.1432	74.28	74.84
0	9.5	1	74.65	0.1492	74.36	74.95
0	10	1	74.73	0.1578	74.42	75.04
			...			

Figure 1 (a) indicates the predicted temperature values of 20 sensors at each .5 hour in Cluster A and the solid red line represents the estimated temperature by Cluster Head A over 14.5 hours. Figure 1(b) presents the predicted temperature of each sensor and the green line is the temperature trajectory estimated at the corresponding cluster head in Cluster B within the same time interval. To show the significant temperature difference in the two areas, we compare the estimated temperature of the two areas in Figure 1 (c).

Figure 1(d) presents the residuals of the predicted values of each sensor. We found that all the predicted values were controlled within the $\pm$ 1.5 standard deviation. This simulation shows that our approach can satisfy the user controllable error bound requirements. We also compared the energy consumption with the general approach based on 50 random deployed sensors based on equation (9) and (10) within 14.5hours time interval. We compared the general data aggregation approach with our multilevel Bayesian approach in the same WSN topology and found that our approach has slightly higher energy consumption than General Data aggregation approach in the initial 1.5 hour time window. That is because the Bayesian model needs to

transmit more parameters than real temperature data at the beginning, however, with longer time period (1.5-14.5 hours), our approach has achieved significantly less energy consumption than the linear-increasing energy consumption of the General Data Aggregation approach when no parameters update is needed.

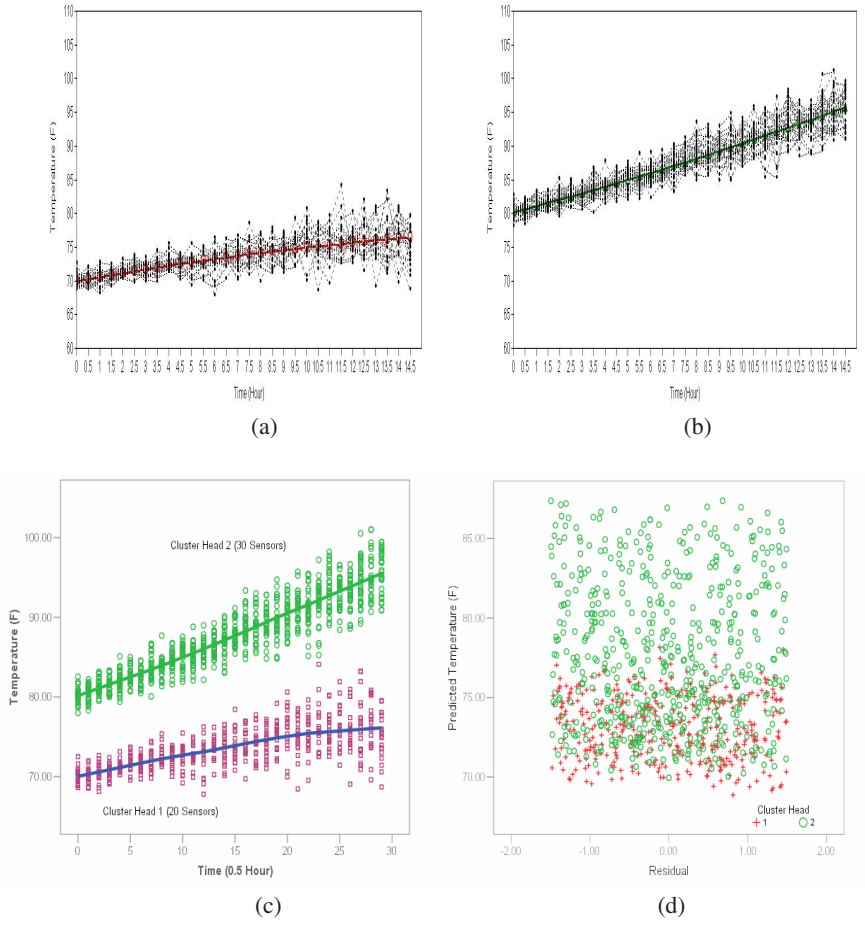

Fig. 1. Predicted values of each sensor against estimated value by each cluster head in two areas over 14.5 hours

5 Conclusions

In this paper, we proposed a multilevel Bayesian modeling approach to the query application in the WSN multilevel architecture, utilizing both temporal and spatial correlation to predict parameters at different levels. Our approach relies mostly on local Bayesian models computed and maintained at each sensor. In order to adapt the local model to variations in the data distribution, each sensor continuously maintains its local model, and notifies the sink only of significant changes. As we showed, our

approach can provide a significant reduction in communication load over the existing general data aggregation approach, and can also effectively predict future values with controllable error bounds. By using this approach, significant energy consumption is saved in typical data query applications.

References

1. Jain, A., E.Y., and Wanf, Y. Adaptive stream management using kalman filters. In SIGMOD, 2004.
2. Kotidis, Y., Snapshot queries: towards data-centric sensor networks. In Proc. Of the 21th Intl. Conf. on Data Engineering, April 2005.
3. Chu, D., Desphande, A., Hellerstein, J., and Hong, W. Approximate data collection in sensor networks using probabilistic models. In ICDE, April 2006.
4. Heinzelman, W. R., Chandrakasan, A., and Balakrishnan, H., An Application-Specific Protocol Architecture for Wireless Microsensor Networks, IEEE Transactions on Wireless Communications, vol. 1, no. 4, pp. 660¿C670, October 2002.
5. Lin, C.R., Gerla, M., Adaptive Clustering for Mobile Wireless Network, IEEE J. Select. Area Commun., vol15, pp. 1265-1275, Sept 1997.
6. Ryu, J.H., Song, S., Cho, D.H., Energy-Conserving Clustering Scheme for Multicasting in Two-tier Mobile Ad-Hoc Networks, Elec-tron. Lett., vol. 37, pp. 1253- 1255, Sept 2001.
7. Hou, T.C., Tsai, T.J., Distributed Clustering for Multimedia Support in Mobile Multihop Ad Hoc Network, IEICE Trans. Commun., vol. E84B, pp. 760-770, Apr 2001.
8. Han, J., Kamber, M., Data Mining: Concepts and Techniques, San Diego: Academy Press, 2001.
9. Productive Patterns Software Web site. http://www.predictivepatterns.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overvie w.htm
10. SAS Institute Inc. (2003). SAS/STAT userí»s guide, version 9.1. Cary, NC: SAS Institute Inc..
11. Fang, H. (2006). %hlmdata and %hlmrmpower: Traditional repeated measures vs. HLM for multilevel longitudinal data analysis - power and type I error rate comparison. Proceedings of the Thirty-First Annual SAS Users Group Conference, SAS Institute Inc., Cary, NC.
12. Fang, H., Brooks, G. P., Rizzo, M. L., & Barcikowski, R. S. (2006). An empirical power analysis of multilevel linear model under three covariance structures in longitudinal data analysis. 2006 Proceedings of the Joint Statistical Meetings, American Statistical Association, Quality Industry and Technology Section [CD-ROM]. Seattle, Washington: American Statistical Association.
13. Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51, 313-326.
14. Raudenbush S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods. London: Sage publications, Inc..
15. McCulloch, C. E., & Searle, S. R. (2001). Generalized, linear, and mixed models. New York: John Wiley & Sons, Inc.
16. Carlin, B. P. & Louis, T. A.(2000). Bayes and Empirical Bayes Methods for Data Analysis. NY: Chapman & Hall/CRC.
17. Little, R.J., & Rubin, D.B. (2002). Statistical analysis with missing data (2nd edition). New York: John Wiley.
18. Ross, S. M. (2003). Introduction to probability models. San Diego, CA: Academic Press.