http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (12): 2286-2295, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.2286.2295

Efficient Data Gathering based on Linear Regression in Wireless Multivariate Monitoring Sensor Networks

[1]Guofeng Yang, [1]Chunsheng Li and [2]Haotian Yang
[1]School of Computer and Information Technology,
[2]School of Electrical Engineering and Information, Northeast Petroleum University, Daqing, China

Abstract: Gathering sensed information in an energy efficient manner is an important design challenge in the application of wireless sensor networks. The readings of sensors generally exhibit both spatial and temporal redundancies due to redundant node deployment and spatial and temporal correlations between the sensed data. Therefore, in this paper, the distributed regression theory is used to remove the correlation in wireless multivariate monitoring sensor networks. Sensor nodes need not transmit data to one another or the sink and only communicate the regression model parameters. The proposed algorithm reduces amount of data and energy consumption during the data transmission process, thus prolongs the lifetime of the whole networks. In order to validate the algorithm, simulation is carried out to evaluate the energy consumption and prediction accuracy. The result of simulation shows that the proposed algorithm is very suitable for the compression of multivariate monitoring data.

Key words: Data gathering, multivariate, distributed regression, wireless sensor networks

INTRODUCTION

In Wireless Sensor Networks (WSNs), one of the main challenging research topics is to save severely constrained energy resources and effectively extend the lifetime of the network (Anastasi *et al*., 2009). The power consumption in a sensor node can be divided into three parts: sensing consumption, data processing consumption and transmission consumption. Most of power consumption in WSN is used for data transmission (Kimura and Latifi, 2005). Thus, minimizing the size of data will reduce transmission consumption (Ciancio and Ortega, 2004, 2005; Shan *et al*., 2011; Guo *et al*., 2010; Wei *et al*., 2010). Due to the redundant sensor node deployment for fault tolerance of communication connectivity, WSNs exhibit naturally high redundancy in spatio-temporal sampling. The sampled redundant attributes allow a significant reduction of communication overhead by data compression (Tulone and Madden, 2006; Mehmet *et al*., 2004). The reduction of power consumption directly bring into lifetime extension by using the data compression for the network nodes (Kolo *et al*., 2012; Srisooksai *et al*., 2012; Marco *et al*., 2012). The studies of Chou *et al*. (2003), Zixiang *et al*. (2004) give examples of applying the Slepian-Wolf theorem to compress collecting data in WSNs where correlated data streams are physically separated or each sensor node has limited computation capability. The compression schemes allow sensor nodes to compress their sensed data without collaboration and negotiation but need to know prior knowledge of the precise correlation in the data. However, many civil applications of WSNs, such prior knowledge is usually vague. Therefore, the compressive sensing techniques are presented by exploiting compressibility without relying on any specific prior knowledge or assumption on signals (Donoho, 2006; Zheng *et al*., 2012). These schemes provide decentralized compression in WSNs, but they cannot support multi-resolution compression. Wavelet-based compression support multi-resolution storage in a WSN by organizing the network into multiple levels. The Ganesan *et al*. (2003) adopt the three-dimensional discrete wavelet transform (3D-DWT) to generate spatio-temporal summarization of sensing data in each level. The different resolutions of sensing readings are obtained from different levels via drill-down queries. Although DIMENSIONS meets the data compression requirement, it is too complicated for sensor nodes because wavelet-based compression would incur high computation and storage complexity. Also, such complicated wavelet operations are performed at each level of the DIMENSIONS hierarchy. Ciancio and Ortega (2004) analyze the energy consumption of data compression and data reconstruction accuracy in using distributed and non-distributed wavelet transform mode. From the energy point of view, the literature (Ciancio and

Corresponding Author: Guofeng Yang, School of Computer and Information Technology, Northeast Petroleum University, Daqing, China

Ortega, 2004) studies local coefficient quantization distortion of data reconstruction and local coefficient quantization rules. Ciancil *et al.* (2006) based on a DWT propose an energy efficient data representation and routing scheme. Wagner *et al.* (2005, 2006) propose WSN distributed irregular wavelet transform schemes. The program take into account the sensor nodes deployed in space distribution is irregular, uneven and therefore can not be directly applied traditional wavelet transform.

Marcelloni and Vecchio (2009) introduced Huffman coding into wireless sensor nodes. Their simple lossless entropy compression algorithm which was based on static Huffman coding exploits the temporal correlation that exist in sensor data to compute a compressed version using a small dictionary, the size of the ADC resolution. The algorithm was particularly suitable for computational and memory resource constrained sensor nodes. The algorithm is static. Hence, the algorithm cannot adapt to changes in the source data statistics. Tharini and Ranjan (2009) proposed algorithm was a modified version of the classical adaptive Huffman coding. The algorithm does not require prior knowledge of the statistics of the source data and compression is per for med adaptively based on the temporal correlation that exists in the source data. The draw back of this algorithm is that it is computationally intensive. Maurya *et al.* (2011) proposed a compression algorithm that uses median predictor to decorrelate the sensed data. The proposed algorithm is simple and can be implemented in a few lines of code and uses the LEC compression table. The algorithm has similar compression complexity as LEC but lower compression efficiency. Since the LEC algorithm outperforms it, the algorithm will not be used for comparison with our algorithm. Liang and Peng (2010), proposed a scheme called two-modal transmission for predictive coding. In the first modal transmission which is called compressed mode, the compressed bits of error terms falling inside the interval [-R, R].

The algorithm of distributed regression has been addressed by many researchers till date. The problem of performing global regression is considered in a vertically partitioned data distribution scenariod (Hershberger and Kargupta, 2001). The authors propose a wavelet transform of the data such that, after the transformation, effect of the cross terms can be dealt with easily. The local regression models are then transmitted to the central station and combined to form the global regression model. The drawback of the algorithm is the need to the synchronization techniques that are unlikely to scale in large, asynchronous systems. Guestrin *et al.* (2004) presented a linear regression framework in a network of sensors using in-network processing of messages (Guestrin *et al.*, 2004). Instead of transmitting the original data, the proposed technique transmits regression coefficients only, thereby reducing the communication energy consumption drastically. However, the major drawback is that their algorithm is not suitable for dynamic data. An algorithm based on multivariate correlation is proposed by Zhu *et al.* (2009). The algorithm can effectively reduce spatial-temporal and multivariate correlations, but all the raw data of cluster members in a cluster must be transmitted directly to the Cluster-Head (CH) and be compressed in the CH. The readings with different attributes but from different nodes are not differentiated and abstracted into a column of the processed data matrix. Before sending data, the CH must perform data preprocessing algorithm to analyze and find out the attribute pairs between which the correlation is large. Song *et al.* (2012) presented a distributed linear regression-based data gathering framework in clustered WSNs. The raw readings can be approximately represented under less than a prespecified threshold while the communication energy consumption can be significantly reduced by the framework. CH nodes perform linear regression operations and use historical sensory data to complete estimation of the actual monitoring measurements. Rather than transmitting original measurements to the sink station, CH nodes transmit constraints on the regression parameters.

In this study, we take into account the data collected by different sensing units in the same sensor node that is, multivariate data or multi-attribute data. In our method, all non-CH nodes collect the original data and select one variable as the base function of the other variables. Each node calculates the regression coefficients of the base variable by using the time as independent variable and the regression coefficients of the non-base variable by using the base variable as independent variable. Non-CH nodes transmit their coefficients to the CH. While the CH receives the coefficients from the active cluster members and sends the model coefficients to the remote sink node.

REGRESSION MODELS IN SENSOR NETWORKS

A wireless sensor network is composed by a set of N energy-constrained sensor nodes that are randomly deployed in two-dimensional field. Nodes can simultaneously a set of environmental attributes, such as temperature, humidity, light intensity, sound intensity, acceleration, video, etc., in which certain correlation universally exists.

Simple linear regression model: The current solutions of data reduction by means of linear regression are performed by using simple linear regression based on the least squares 7. In that case, each sensor node calculates regression coefficients by using the epoch/time as independent variable. Then, the sensor node sends its coefficients to the sink, instead of sending the readings.

Over time, a sensor node measures a function at some time t (e.g., temperature). Then, it collects a set of data points $(t_1, x_1), (t_2, x_2),..., (t_m, x_m)$. Assume a set of basis functions $B = (b_1, b_2,..., b_k)$ are given, the measurements can be approximated by these basis functions. That is, to find basis function coefficients $w = (w_1, w_2,..., w_k)^T$ such that the measurements are approximated as:

$$\sum_{i=1}^{k} w_i b(t_i)$$

where:

$$\tilde{x}(t) = \sum_{i=1}^{k} w_i b(t_i)$$

and k is the number of coefficients. When the number of coefficients is equal to the number of sampling (k = m), each x_j can be calculated exactly. However, such high-degree $\tilde{x}(t)$ fit the noise into the monitoring data and generally generates poor results when used to predict unseen data points (t,x). When the number of coefficients is far smaller than the number of sampling (k = m), the coefficient vector w becomes a compressed representation of the sampling data.

Let $X = (x(t_1), x(t_2),..., x(t_m))^T$ denotes the actual measurements vector with one row for each measurement. The basis matrix of the basis functions at the corresponding sampling time points was defined as matrix B:

$$B = \begin{pmatrix} b_1(t_1) & b_2(t_1) & \dots & b_k(t_1) \\ b_1(t_2) & b_2(t_2) & \dots & b_k(t_2) \\ \vdots & \vdots & \cdots & \vdots \\ b_1(t_m) & b_2(t_m) & \dots & b_k(t_m) \end{pmatrix} \tag{1}$$

where, B is an m×k matrix and x is an m×1 vector. Let $\tilde{X} = (\tilde{x}(t_1), \tilde{x}(t_2), ..., \tilde{x}(t_m))^T$ denote the m×1 vector with one row for each approximation values at t_i sampling time points, then:

$$\tilde{X} = \begin{pmatrix} \tilde{x}(t_1) \\ \tilde{x}(t_2) \\ \vdots \\ \tilde{x}(t_m) \end{pmatrix} = B\omega = \begin{pmatrix} b_1(t_1) & b_2(t_1) & \dots & b_k(t_1) \\ b_1(t_2) & b_2(t_2) & \dots & b_k(t_2) \\ \vdots & \vdots & \cdots & \vdots \\ b_1(t_m) & b_2(t_m) & \dots & b_k(t_m) \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_k \end{pmatrix} \tag{2}$$

To guarantee the error bound of each approximation data and the corresponding sampling data, the approximation errors δ on Root Mean Squared error (RMS) are defined:

$$\begin{aligned} \delta &= \sqrt{\frac{1}{m} \sum_{j=1}^{m} (x(t_j) - \tilde{x}(t_j))^2} \\ &= \sqrt{\frac{1}{m} \sum_{j=1}^{m} (x(t_j) - \sum_{i=1}^{k} \omega_i b_i(t_j))^2} \end{aligned} \tag{3}$$

To minimize approximation errors, the optimization problem is stated as:

$$\begin{aligned} w^* &= \arg\min_{w} \delta \\ &= \sqrt{\frac{1}{m} \sum_{j=1}^{m} (x(t_j) - \tilde{x}(t_j))^2} \end{aligned} \tag{4}$$

Setting the gradient of this quadratic objective to zero gives the optimal coefficients in matrix form:

$$w^* = (B^T B)^{-1} B^T X \tag{5}$$

Let $A = B^T B$ and $c = B^T X$. The equations are following as:

$$A = B^T B = \begin{pmatrix} <b_1 \cdot b_1> & <b_1 \cdot b_2> & \dots & <b_1 \cdot b_k> \\ <b_2 \cdot b_1> & <b_2 \cdot b_2> & \dots & <b_2 \cdot b_k> \\ \vdots & \vdots & \cdots & \vdots \\ <b_k \cdot b_1> & <b_k \cdot b_2> & \dots & <b_k \cdot b_k> \end{pmatrix}$$

$$c = B^T X = \begin{pmatrix} <b_1 \cdot x> \\ <b_2 \cdot x> \\ \vdots \\ <b_k \cdot x> \end{pmatrix}$$

We can transform the Eq. 5 to $w^* = (A)^{-1} c$, namely:

$$c = Aw^* \tag{6}$$

where, A denotes the dot-product matrix, where each element is the dot product between two basis functions. c is the projected measurement vector, where each element denotes simply the projection of the measurement vector into the space of a particular basis function. When the measurement vector and the basis functions are given, the optimal regression weights can be computed with simple matrix operations.

Over time, it is necessary to update the linear regression model for reconstruction of sampling data. We fit the coefficients of our basis functions with respect to the sampling data collected in the last T minutes. Suppose that the matrix A and c have been computed for the sampling data at times $t_1,..., t_{m-1}$ and a new measurement at time t_m are obtained as the following:

$$A = \begin{pmatrix} <b_1(t_m) \cdot b_1(t_m)> & <b_1(t_m) \cdot b_2(t_m)> & \dots & <b_1(t_m) \cdot b_k(t_m)> \\ <b_2(t_m) \cdot b_1(t_m)> & <b_2(t_m) \cdot b_2(t_m)> & \dots & <b_2(t_m) \cdot b_k(t_m)> \\ \vdots & \vdots & \cdots & \vdots \\ <b_k(t_m) \cdot b_1(t_m)> & <b_k(t_m) \cdot b_2(t_m)> & \dots & <b_k(t_m) \cdot b_k(t_m)> \end{pmatrix}$$

$$c = \begin{pmatrix} <b_1(t_m) \cdot x(t_m)> \\ <b_2(t_m) \cdot x(t_m)> \\ \vdots \\ <b_k(t_m) \cdot x(t_m)> \end{pmatrix}$$

So, the matrix A of the basis functions and the projected measurement vector c are updated by the increment operation expression 7:

$$A \leftarrow A+A(t_m)\ c \leftarrow c+c(t_m) \tag{7}$$

Similar to the operation expression (7), if measurement t_1 falls outside the time sliding window, the linear regression model is updated according to the Eq. 8:

$$A \leftarrow A-A(t_1)\ c \leftarrow c-c(t_m) \tag{8}$$

Thus, when new measurements are received at any time, the dot-product matrix A of the basis functions and the projected measurement vector c can be updated by implementing the increment operations as well as the basis function coefficients of linear regression model can be computed by solving the linear system c = Aw*.

Multivariable linear regression model: In wireless multivariate monitoring Sensor Networks, a sensor node is able to perform monitoring of more than one variable. Moreover, the multivariate correlation is usually strong. The correlation happens due to the fact that each sensor node gathers correlated data from one or more attributes at a given time. It is observed in the nature of physical phenomena (Mehmet *et al*., 2004). The simple linear regression model is able to work over correlation, but it is not able to work over the multivariate correlation (more than one variable). In our solution, we use multivariate linear regression model to work over the multivariate correlation. The purpose of our paper is to apply the multivariate correlation method to improve prediction accuracy on WSN data reduction.

Suppose that a sensor node has P sensing units, the collected attribute is x_j, j = 1, 2,…, P. The overall operation of the regression-based compression scheme is as follows:

$$y_i = 1+\beta x_{i,1}+\beta x_{i,2}+...+1+\beta_p x_{i,p-1} \tag{9}$$

where, y_i denotes an attribute called response value and $x_{i,1}$, $x_{i,2}$,..., $x_{i,p-1}$ are the remaining p-1 attributes at a given time i. We can pack all response values for all actual measurements into an m-dimensional vector:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

We can pack all predictors into a m×(p-1)+1 matrix:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(1-p)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(1-p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{m(1-p)} \end{pmatrix}$$

We can pack the regression coefficients into a p-dimensional vector:

$$\beta = \begin{pmatrix} 1 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

Using linear algebra notation, the model 9 can be compactly written:

$$Y = X\beta$$

In order to estimate β, we take a least squares approach to minimize Eq. 10:

$$\sum_{i=1}^{m}\left[y_i-(1+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\cdots+\beta_p x_{i,p-1})\right]^2 \tag{10}$$

And the regression coefficients can be determined by Eq. 11 through the least square evaluation:

$$\beta = (X^TX)^{-1}XY \tag{11}$$

Then, the missing sense data can be predicted according to Eq. 11.

Distributed algorithm using regression: The network model was assumed that a set of energy-constrained sensor nodes were randomly deployed in M×M two-dimensional field. The following assumptions are made for the sensor network. All sensor nodes are not mobile and unaware of their location. The immobile sink node is only and considered to be a powerful node endowed with enhanced communication and computation capabilities and no energy constraints. Sensor nodes can adjust the transmitting power according to the distance, namely, radio transmitting power of nodes is controllable. Sensor nodes are fitted with the same radio communication model. The radio channel is symmetric so that the energy required to transmit m-bit message from node i to node j is identical to the energy required to transmit m-bit message from j to i.

The existence of the temporal as well as spatial correlations brings the potential to significantly develop and implement the efficient communication protocols well-suited for the WSN paradigm. Here, the distributed algorithm is presented to exploit the spatio-temporal correlation characteristics of the clustered sensor network based on regression model that can approximate the raw data while significantly reducing the communication energy consumption.

In order to take advantage of the existence of nodes of different abilities inside a WSN, data gather processing makes use of the classical LEACH protocol (Heinzelman *et al.*, 2002; Heinzelman *et al.*, 2000). The nodes organize themselves into local clusters, with one node acting as the CH. All non-CH nodes collect the original data, calculate the coefficients by performing regression for original measurements and transmit their coefficients to the CH. While the CH receives the coefficients from the active cluster members and sends the model coefficients to the remote sink node. The processing principle of the distributed solution is derived as following.

Suppose that a sensor node has N sensing units, the collected attribute is A_j, j = 1, 2, L, N. Initially, each node selects an attribute as the base attribute according to the correlation coefficient matrix of multivariate sampling data. The correlation coefficient of the attribute between X and Y is denoted as Eq. 12:

$$r_{X_1X_2} = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{D(X_1)}\sqrt{D(X_2)}} = \frac{\sum_{i=0}^{N-1}[(x_{1_i} - E(X_1))(x_{2_i} - E(X_2))]}{\sqrt{\sum_{i=0}^{N-1}(x_{1_i} - E(X_2))^2}\sqrt{\sum_{i=0}^{N-1}(x_{2_i} - E(X_2))^2}} \tag{12}$$

Where:

$$E(X_1) = \tfrac{1}{N}\sum_{i=0}^{N-1} x_{1_i}$$

and:

$$\mathrm{Cov}(X_1, X_2) = E((X_1\text{-}E(X_1)(X_2\text{-}E(X_2)))$$

When the absolute value of $r_{X_1X_2}$ is 1, the relationship between X1 and X2 is complete correlation. If sensor node uses attribute X_1 as independent variable, all the data points of attribute X_2 lie on the regression line. The smaller the absolute value $r_{X_1X_2}$ of is, the lower the correlation is and the more scattered the data points are. If each node can collect H attributes, the relationship among all attributes is defined as the correlation coefficient matrix R with the size H×H, in which the element of the jth column in the ith row denotes correlation coefficient between the i-th and the jth attribute. The best attribute Xj as independent variable is selected by Eq. 13:

$$\mathrm{opt\_fit}_j = \max_j\left(\sum_i |r_{ij}|\right),\ \ i = 1, 2, \cdots, H,\ j = 1, 2, \cdots, H \tag{13}$$

```
Dis_regress(i)
{//initialize parameters
Max_time_window←size of the sampling time sliding window
MSG_interval←timer interval to send messages.
ε←precision
CLUSTER_id←the cluster id number of sensor nodes.
For (each node I) do //select the base variable
    $\left\{\mathrm{opt\_fit}_j = \max_j\left(\sum_j |r_{ij}|\right);\right\}$
For (each node i) do {A←0; c←0; w←0;}
For (each node i) do //build the regression model
    {if (T<Max_time_window)}
        {A = A+A(T); w = w+w(T);}
     else
        (A = A+A(T); w = w+w(T); A = A-A(T1); w = w-w(T1);)
  For (each MSG_interval)
       ($w^* = (A)^{-1}$ c;
       $\beta = (X^TX)^{-1}$ XY;
       Send_message (CLUSTER_id, w, β);
       }
    }
}
```

Fig. 1: Regression algorithm

If the estimate error of the attribute $\mathrm{opt\_fit}_j$ using the time as independent variable is higher than the threshold, a node will re-selected sub-optimal attribute Xj as independent variable.

When the base attribute has been chosen, the node performs the regression algorithm, shown in Fig. 1. Each node N_i maintains a matrix A(i) and a vector c(i) that summarize, respectively the effect of this node's measurements in the dot-product matrix and the projected easurement vector for its base attribute. When the node collects a new value, its local matrix and vector are updated using the incremental rule and an event is scheduled to delete this value when it falls outside the time window. The node computes the non-base attributes on the base attribute and transmits the calculated regression coefficients to the CH. The raw data of nodes is no longer required to be transmitted.

An alternative to transmitting all of the measurements is to build a regression model of this data in the network and transmit only the model coefficients. These lead to lesser packet transmissions and reduce redundancy, thereby helping in prolonging the network lifetime.

Table 1: Ten successive measurements for different environmental attributes

Frequency	Power	Current	Voltage	Panel temperature	Pipe temperature	Tank temperature	Tank level
50.00	6.29	11.32	378.00	11.16	21.20	28.91	1.21
50.00	6.29	11.40	377.00	11.18	21.22	28.91	1.21
49.78	5.98	11.05	377.00	11.18	21.21	28.92	1.21
49.15	5.79	10.88	378.00	11.18	21.22	28.91	1.21
50.00	6.14	11.10	379.00	11.18	21.21	28.91	1.21
49.52	5.95	11.02	381.00	11.20	21.22	28.91	1.21
49.71	6.11	11.29	383.00	11.19	21.21	28.91	1.21
50.00	6.09	11.20	378.00	11.19	21.23	28.92	1.21
50.00	6.07	11.05	372.00	11.20	21.22	28.90	1.21
49.04	5.60	10.55	365.00	11.20	21.22	28.91	1.21

Table 2: Coefficient of the correlation analysis

	Frequency	Power	Current	Voltage	Panel temperature	Pipe temperature	Tank temperature	Tank level
Frequency	1.000000	0.911674	0.820903	0.3865990	-0.355050	-0.218040	0.031860	-4.1E-15
Power	0.911674	1.000000	0.957566	0.5610840	-0.545530	-0.367800	-0.028420	2.63E-15
Current	0.820903	0.957566	1.000000	0.6952360	-0.509100	-0.287040	0.090692	7.61E-16
Voltage	0.386599	0.561084	0.695236	1.0000000	-0.345860	-0.256770	0.241327	2.38E-15
Panel temperature	-0.355050	-0.545530	-0.509100	-0.3458600	1.000000	0.666667	-0.247590	2.96E-14
Pipe temperature	-0.218040	-0.367800	-0.287040	-0.2567700	0.666667	1.000000	0.092848	1.78E-13
Tank temperature	0.031860	-0.028420	0.090692	0.2413270	-0.247590	0.092848	1	1.98E-13
Tank level	-4.1E-150	2.63E-15	7.61E-16	2.38E-150	2.96E-14	1.78E-13	1.98E-13	1
Time	-0.315590	-0.552600	-0.557840	-0.4743500	0.870388	0.565752	-0.226280	0

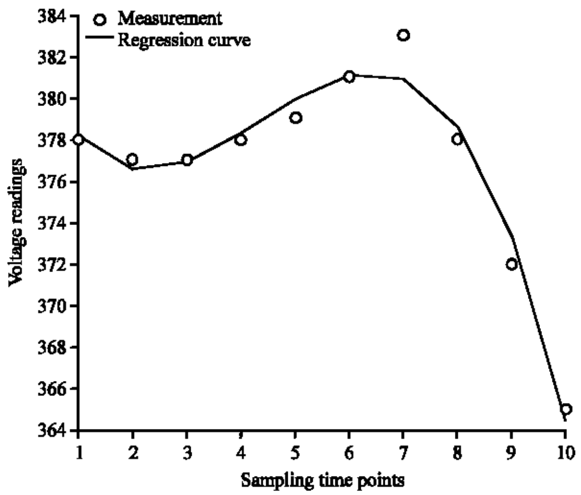

Fig. 2: Voltage regression curve of ten sampling time points

For example, instead of extracting the original measurement from node Ni every 10 sec, suppose that we have ten raw readings for every attribute during the sampling time, shown in Table 1. The correlation coefficient r results in Table 2 show that there is a greater correlation between the voltage variable and other variables gathered by the sensor nodes than with the time variable. Thus, we select the voltage variable as the base attribute. In order to perform simple computing, we wish to fit the last 10 sampling points with a degree-three polynomial: f(t) = $w_0+w\,t_1+w\,t_2^2+w\,t_3^{3}$ and only need to extract 4 parameters from the voltage readings: w_0, w_1, w_2 and w_3. More generally, given a set of basis functions of the voltage readings (e.g., 1, t, t^2 and t^3), we would like to continuously fit their parameters and thereby reduce the dimensionality of the voltage readings. The model coefficient vector was computed by Eq. 6 that is, -0.1340, 1.7494, -5.9044, 382.4667. Therefore, the degree-three polynomial is Eq. 14. The real line denotes the regression prediction curve of ten voltage values in Fig. 2:

$$Y(t) = -0.1340+1.749t+-5.90144t^2+382.4667t^3 \quad (14)$$

The other variables use voltage as independent variable to extract 2 parameters from temperature readings: β_1 and β_2 which was computed by Eq. 11. The red lines denote the regression estimate curves of ten prediction values other than voltage readings in Fig. 3.

EXPERIMENTS AND EVALUATION

Here, to analyze the validity of the regression strategy, we implemented it in a small WSN which contains frequency, power, current, voltage, panel temperature, pipe temperature, tank temperature and tank level readings gathered by multisensors in a solar water pressure monitoring system at intervals of 10 sec. These readings were held during the day, between 1 February and 5 April 2013. Thus, the data gathered for our simulation comes from a reality scenario. We compare the distributed linear regression-based strategy against the standard clustered LEACH and regression algorithm

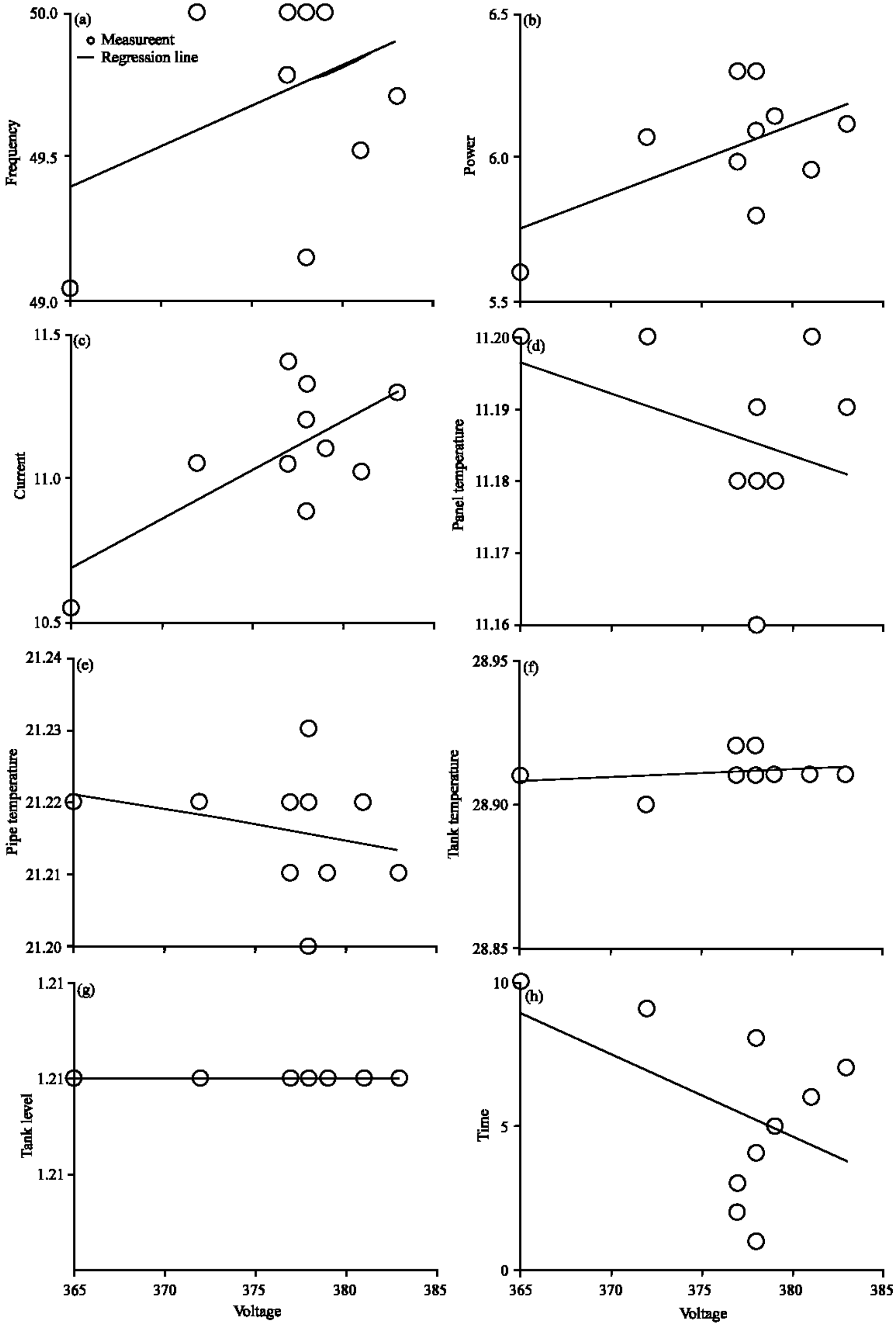

Fig. 3(a-h): Regression curves of different environmental attributes under voltage variable as independent variable presented above. We provide a particular analysis of the proposed algorithm for the energy consumption and the prediction accuracy. The NS-2 software is used to implement and simulate the network system.

Evaluation of the energy consumption: For power consumption used for transmitting and receiving, we adopt a simple radio model by Heinzelman *et al.* (2002). Specifically, each node needs to run the circuitry for the power amplifier. Let e_l^r (J bit^{-1}) be the power consumption over the link l, when it receives one unit of data and e_l^t (J bit^{-1}) be the power consumption when one unit of data is sent over the link l. We have:

$$e_l^r = \varepsilon_{elec}$$
$$e_l^t = \varepsilon_{elec} + \varepsilon_{amp} d_l^{\alpha}$$

where, ε_{elec} is a distance-independent constant that denotes the energy consumption to run the transmitter or receiver radio electronics and ε_{amp} is the coefficient of the distance-dependent term that denotes the transmit amplifier. α is the path loss exponent which is usually between 2 and 4 for free-space and short-to-medium-range radio communication. For the experiments described in this study, the main simulation of the WSN are set as Table 3.

For these experiments, each node begins with only 0.5 J initial energy and 200 bytes control packets to send to the sink node. The CH node was determined at the beginning of each round which lasts for 20 sec. Node will generate energy consumption whenever a sensor in network transmits or receives data or performs regression operation. Figure 4 shows how the total energy consumption of the network at each round varies as the simulation time runs on for the proposed protocol and LEACH protocols. The simulation results demonstrate that the CH nodes of the proposed algorithm required less energy in the simulation time than LEACH protocol. This is because a much smaller amount of packets was transmitted to the sink by CH using the regression estimate model to provide a structured prediction of the original data. The energy consumption increased slightly in each regression period for computing the model coefficients. The number of dead nodes in the network at each round is shown in Fig. 5. It can be seen from the figure that the number of dead nodes in our scheme is less than leach protocol. The reason is that each node just transmits the coefficients whose amounts are far less than the original amounts of data. So, the proposed algorithm saves energy and belongs the network lifetime.

Table 3: System parameters of the simulation scenarios

Parameter	Acronym	Value/ type
Transceiver average power consumption (nJ bit^{-1})	ε_{elec}	50.0
Transmitter power gain coefficient (pJ/bit/m^2)	ε_{amp}	100.0
Path loss coefficient	α	2.0
Initial energy of the node (J)	Eini	0.5
Regression estimate energy cost (nJ bit^{-1})	Ecom	5.0
The bandwidth of the channel (Mb sec^{-1})	Band_width	1.0
Data message size (bytes)	Data_size	500.0
Transmission delay (μsec)	Tran_delay	25.0
The interval of each round (sec)	Round_time	20.0
Simulation time (sec)	Sim_time	600.0
Regression period (sec)	R-period	30.0

Figure 6 shows that the base attribute regression estimate curve deviates from the actual measurements spot. The absolute value error between the measurements

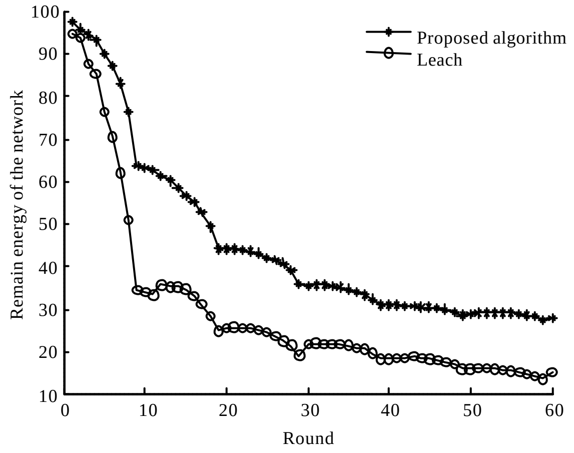

Fig. 4: Total energy consumption of the network at each round

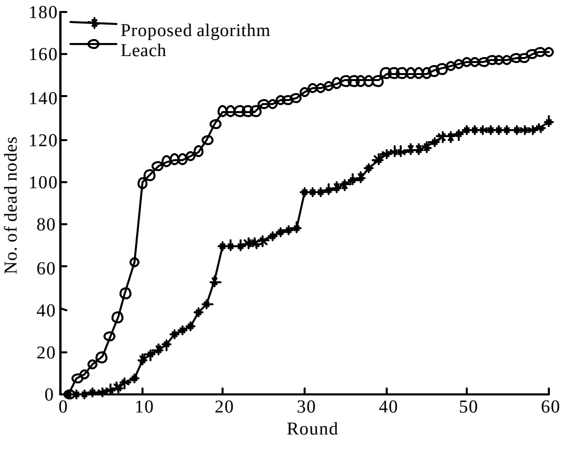

Fig. 5: Number of dead nodes in the network at each round

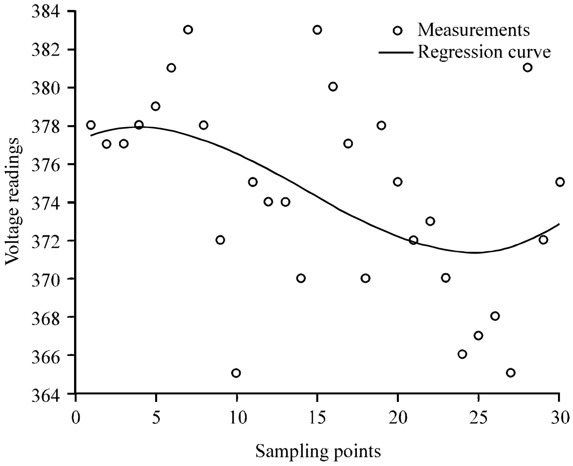

Fig. 6: Base attribute regression curve at the varied sampling time points

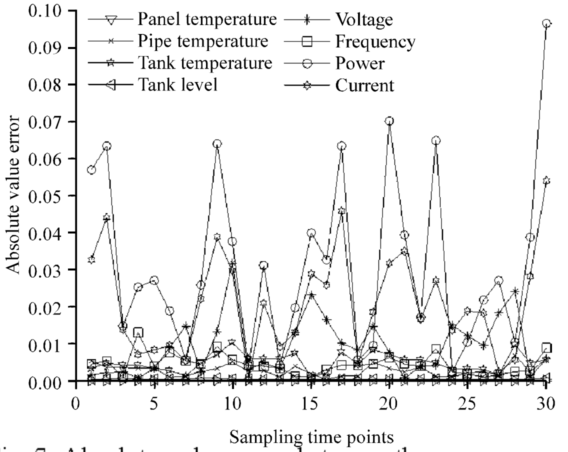

Fig. 7: Absolute value error between the measurements and regression prediction values

and regression prediction values are shown in Fig. 7. The variations in the absolute value error per round over time are small for each attributes. For the power sampling, the absolute value error of the last sampling time point is the biggest up to 9.5% which is below the certain prespecified error threshold.

CONCLUSION

In this study, an effective distributed regression model is used to implement the wireless multivariate monitoring sensor networks. The proposed algorithm uses the correlation between base and non-base variables to compute the regression coefficients of non-base variables. The algorithm runs independently on each node. Rather than transmitting sensor readings at a continuous rate, our scheme allows each node to locally compute the regression coefficients. After finding the optimal base variable and distributed regression computing, the node transmits the regression coefficients to sink by the CH nodes. The sink has the coefficients of the estimate model to predict the approximation of the monitoring data. Experimental results demonstrate that the algorithm is capable of accurately summarizing and estimating values of sensor measurements small amounts of communication and obtain more savings in the energy as compared with LEACH.

REFERENCES

Anastasi, G., M. Conti, M. di Francesco and A. Passarella, 2009. Energy conservation in wireless sensor networks: A survey. Ad Hoc Networks, 7: 537-568.

Chou, J., D. Petrovic and K. Ramchandran, 2003. A distributed and adaptive signal processing approach to reducing energy consumption inn sensor networks. Proceedings of the INFOCOM 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, March 30-April 3, San Francisco, CA, USA., pp: 1054-1062.

Ciancil, A., S. Pattem, A. Ortega and B. Krishnamachari, 2006. Energy-efficient data representation and routing for wireless sensor networks based on a distributed wavelet compression algorithm. Proceedings of the 5th International Conference on Information Processing in Sensor Networks, April 19-21, USA., pp: 309-316.

Ciancio, A. and A. Ortega, 2004. A distributed wavelet compression algorithm for wireless sensor networks using lifting. Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP), Vol. 4. May 17-21, 2004, Montreal, Canada, pp: 633-636.

Ciancio, A. and A. Ortega, 2005. A distributed wavelet compression algorithm for wireless multihop sensor networks using lifting. Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP), Vol. 4. March 18-23, 2005, Philadelphia, USA, pp: 825-828.

Donoho, D.L., 2006. Compressed sensing. IEEE Trans. Inform. Theory, 52: 1289-1306.

Ganesan, D., D. Estrin and J. Heidemann, 2003. DIMENSIONS: Why do we need a new data handling architecture for sensor networks? Comput. Communic. Rev., 33: 143-148.

Guestrin, C., P. Bodi, R. Thibau, M. Paski and S. Madden, 2004. Distributed regression: An efficient framework for modeling sensor network data. Proceedings of the International Conference on Information Processing in Sensor, April 26-27, 2004, Berkeley, pp: 1-10.

Guo, L., B. Wang, Z. Liu and W. Wang, 2010. An energy equilibrium routing algorithm based on cluster-head prediction for wireless sensor networks. Inform. Technol. J., 9: 1403-1408.

Heinzelman, W.B., A.P. Chandrakasan and H. Balakrishnan, 2002. An application specific protocol architecture for wireless microsensor networks. IEEE Trans. Wireless Commun., 1: 660-670.

Heinzelman, W.R., A. Chandrakasan and H. Balakrishnan, 2000. Energy-efficient communication protocol for wireless microsensor networks. Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, January 4-7, 2000, IEEE Computer Society, Washington, DC., USA.

Hershberger, D.E. and H. Kargupta, 2001. Distributed multivariate regression using wavelet-based collective data mining. J. Parallel Distrb. Comput., 61: 372-400.

Kimura, N. and S. Latifi, 2005. A survey on data compression in wireless sensor networks. Proceedings of the International Conference on Information Technology, Vol. 2. April 4-6, 2005, Coding. Comput., pp: 8-13.

Kolo, J.G., S.A. Shanmugam, D.W.G. Lim, L.M. Ang and K.P. Seng, 2012. An adaptive lossless data compression scheme for wireless sensor networks. J. Sensors, 12: 1-20.

Liang, Y. and W. Peng, 2010. Minimizing energy consumptions in wireless sensor networks via., two-modal transmission. Comput. Communic. Rev., 40: 12-18.

Marcelloni, F. and M. Vecchio, 2009. An efficient lossless compression algorithm for tiny nodes of monitoring wireless sensor networks. Comput. J., 52: 969-987.

Marco, F.D., G. Shen, R.G. Baraniuk and A. Ortega, 2012. Signal compression in wireless sensor networks. Phil. Trans. R. Soc., 370: 118-135.

Maurya, A.K., D. Singh and A.K. Sarje, 2011. Median predictor based data compression algorithm for Wireless Sensor Network. Int. J. Smart Senser Ad Hoc Networks, 1: 62-65.

Mehmet, V.C., A.B. Ozgur and I.F. Akyildiz, 2004. Spatio-temporal correlation: Theory and applications for wireless sensor networks. Computer Networks, 45: 245-259.

Shan, L., J. Wang, Y. Zhao and Y. Liu, 2011. Synchronous aggregation scheduling with minimal latency in multihop sensornet. Inf. Technol. J., 10: 1626-1631.

Song, X., C.R. Wang, J. Gao and X. Hu, 2012. DLRDG: Distributed linear regression-based hierarchical data gathering framework in wireless sensor network. Neural Comput. Appl., 10.1007/s00521-012-1248-z

Srisooksai, T., K. Keamarungsi, P. Lamsrichan and K. Araki, 2012. Practical data compression in wireless sensor networks: A survey. J. Network. Comput. Applic., 35: 37-59.

Tharini, C. and P.V. Ranjan, 2009. Design of modified adaptive huffman data compression algorithm for wireless sensor network. J. Comput. Sci., 5: 466-470.

Tulone, D. and S. Madden, 2006. PAQ: Time series forecasting for approximate query answering in sensor networks. Proceedings of the 3rd European Conference on Wireless Sensor Networks, February 13-15, 2006, Zurich, Switzerland, pp: 21-37.

Wagner, R., H. Choi, R. Baraniuk and V. Delouille, 2005. Distributed wavelet transform for irregular sensor network grids. Proceedings of the 13th Workshop on Statistical Signal Processing (SSP), July 17-20, 2005, Bordeaux, France, pp: 1196-1201.

Wagner, R.S., R.G. Baraniuk and S. Du, D.B. Johnson and A. Cohen, 2006. An architecture for distributed wavelet analysis and processing in sensor networks. Proceedings of the 5th International Conference on Information Processing in Sensor Networks (IPSN). April 21, 2006, Piscataway, USA, pp: 243-250.

Wei, W., B. Zhou, A. Gao and Y. Mei, 2010. A new approximation to information fields in sensor nets. Inform. Technol. J., 9: 1415-1420.

Zheng, H., S. Xiao, X. Wang and X. Tian, 2012. Energy and latency analysis for in-network computation with compressive sensing in wireless sensor networks. Proceedings of the INFOCOM, March 25-30, 2012, Orlando, FL., pp: 2811-2815.

Zhu, T.J., Y.P. Lin, S.W. Zhou and X.L. Xu, 2009. An adaptive multiple-modalities data compression algorithm using wavelet for wireless sensor net or s. J. Communic., 30: 48-53.

Zixiang, X., A.D. Liveris and S. Cheng, 2004. Distributed source coding for sensor networks. IEEE Signal Process. Maga., 21: 80-94.