Hindawi
Wireless Communications and Mobile Computing
Volume 2019, Article ID 4182563, 10 pages
https://doi.org/10.1155/2019/4182563

Research Article

Novel Energy-Efficient Data Gathering Scheme Exploiting Spatial-Temporal Correlation for Wireless Sensor Networks

Ying Zhou, Lihua Yang, Longxiang Yang, and Meng Ni

Jiangsu Key Laboratory of Wireless Communication, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, China

Correspondence should be addressed to Lihua Yang; yanglh@njupt.edu.cn

Received 14 March 2019; Accepted 8 May 2019; Published 19 May 2019

Guest Editor: Bojan Dimitrijevic

A novel energy-efficient data gathering scheme that exploits spatial-temporal correlation is proposed for clustered wireless sensor networks in this paper. In the proposed method, dual prediction is used in the intracluster transmission to reduce the temporal redundancy, and hybrid compressed sensing is employed in the intercluster transmission to reduce the spatial redundancy. Moreover, an error threshold selection scheme is presented for the prediction model by optimizing the relationship between the energy consumption and the recovery accuracy, which makes the proposed method well suitable for different application environments. In addition, the transmission energy consumption is derived to verify the efficiency of the proposed method. Simulation results show that the proposed method has higher energy efficiency compared with the existing schemes, and the sink can recover measurements with reasonable accuracy by using the proposed method.

1. Introduction

Wireless sensor networks (WSNs), typically consisting of a vast number of densely deployed and collaborative battery-powered sensors, have been widely used in various application fields, such as the environment, industry, and the military [1]. However, the energy constraints are one of the main performance-limiting challenges for the WSNs. In the WSNs, most energy is consumed in three stages: sensing, data processing, and data delivery, and the energy consumed by the data delivery dominates the energy budget. Therefore, the data gathering approach with high energy efficiency is the key to prolong network lifetime.

There are two main types of data gathering methods for WSNs: the mobile sink based data gathering methods [2] and the stationary sink based data gathering methods [3]. In the mobile sink scenario, the long delay is an inevitable problem [4], so both the movement speed and the traveling path for the sink are the difficulties in designing. However, the stationary sink can avoid the above problems. In the stationary sink scenario, since the observed data should be transmitted to the sink by multihop forwarding transmission, the high transmission energy consumption by the sensors is a problem that must be considered, which depends on the routing model and the data reduction technique. Therefore, the high energy efficient data collection with efficient routing is the key in the stationary sink scenario.

Because of the overlap among the observation regions of the different sensors and the gradual variation of the data observed at a sensor over time, the measurements from a WSN are either spatially or temporally correlated [5], which leads to the existing of much redundant information among the observed data. Therefore, data gathering approaches that exploited the inherent correlation have been widely studied to improve the energy efficiency of the WSNs [6]. In [7–11], data reduction methods that utilized temporal correlation or spatial correlation are presented to reduce the transmission energy consumption. Moreover, many data compression algorithms are employed in these data gathering methods to obtain data reduction, where the network coding (NC) and compressed sensing (CS) techniques are applied in the data gathering methods in [7, 8], which utilize the spatial correlation to reduce the data transmissions. In [9–11], the principal component analysis (PCA) methods or different prediction models are applied to reduce the temporal redundancy. However, the redundant information is still excessive

after using the above methods [7–11], which will cause low energy efficiency.

To further improve the energy efficiency, data gathering schemes exploring both temporal and spatial correlations [12–15] are presented. In [12], a collective prediction scheme exploiting spatial-temporal correlation (CoPeST) is given for the energy efficient WSNs. In the CoPeST method, the temporal redundancy and spatial redundancy of data are, respectively, reduced by the prediction approach and similarity based subcluster method. However, the energy cost by the frequent updating of cluster and subcluster topology is large in this case. A framework with dedicated combination of data prediction and compression is discussed for clustered WSNs in [13] (which is called as DPPCA method for convenience). In the framework, the Least Mean Square (LMS) dual prediction algorithm is used to reduce the temporal redundancy, and a centralized PCA technique is utilized to eliminate the spatial redundancy of the sensed data. However, these schemes in [12, 13] are not practical in different monitoring environments due to the employed fixed error threshold in the prediction algorithms. By the spatial and temporal compressions, the multiresolution compression and query (MRCQ) framework is given in [14], which organizes sensor nodes hierarchically and establishes multiresolution summaries of sensing data inside the network. In [15], a neighbor-aided Kronecker compressed sensing scheme is provided for the WSNs. However, the delay of the methods in [13–15] cannot be ignored, because the operation of reducing the temporal redundancy can only be executed after collecting enough data from continuous time intervals. Therefore, these schemes in [13–15] are not suited for the WSNs with the requirement of high real-time transmission.

To solve the above problems and further improve the energy efficiency, we propose an energy efficient data gathering scheme exploiting spatial-temporal correlation for the WSNs. With distinctions to the above approaches, dual prediction and hybrid compressed sensing techniques are jointed to eliminate the redundancy to improve the energy efficiency in the proposed method. Specifically, dual prediction is utilized during the intracluster transmission, and a new error threshold selection method is designed for the prediction stage, which is obtained by optimizing the relationship between the energy consumption and the recovery accuracy. Moreover, hybrid compressed sensing is employed during intercluster transmission, and the cluster heads (CHs) aggregate all the obtained values only when the number of values is no less than the required number of projections for CS reconstruction. Therefore, the proposed method has high energy efficiency and the reasonable quality.

The rest of this paper is organized as follows. The background and system model are discussed in Section 2. In Section 3, we introduce our proposed protocol in detail. Section 4 discusses the simulation results and conclusions are given in Section 5.

2. Background and System Model

2.1. Hybrid Compressed Sensing. Compressed sensing [16], as an advanced sampling theory, provides a new data compression solution, and it indicates that only a small fraction of data projections is needed to reconstruct all of the raw data, which contains many zero entries. Assume a data vector $\mathbf{x}$ has a K-sparse representation under a $N \times N$ transform basis, i.e., $\boldsymbol{\Psi}$:

$$\mathbf{x} = \boldsymbol{\Psi}\boldsymbol{\theta} \tag{1}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$. $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_N]^T$, and it has K ($K \ll N$) nonzero entries. Under this premise, the projection transformation is applied $\mathbf{x}$; that is,

$$\mathbf{z} = \boldsymbol{\Phi}\mathbf{x} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_N]\,\mathbf{x} \tag{2}$$

where $\boldsymbol{\Phi}$ is a $M \times N$ ($M \ll N$) measurement matrix, and it satisfies the restricted isometry principle (RIP) [17]. $\boldsymbol{\phi}_i$ is the corresponding coefficient vector of x_i, and $\boldsymbol{\phi}_i = [\phi_{1i}, \phi_{2i}, \ldots, \phi_{Mi}]^T$. Then the M-dimension vector $\mathbf{z}$ can be used to recover the raw data $\mathbf{x}$ by solving a l_1-norm minimization.

According to the theory, only a few measurements are enough to reconstruct all original data for sink by CS in WSNs. Figure 1(a) shows the process of data gathering with pure CS, where each node i codes its data x_i with its corresponding coefficient vector $\boldsymbol{\phi}_i$ in measurement matrix $\boldsymbol{\Phi}$ and then sends out M-dimension vector $\sum_{k=1}^{a} x_k\phi_k$ which is aggregated by its encoded vector $x_k\phi_k$ and the $(a-1)$ achieved data vectors. In this case, every node needs to send out M data. However, the number of original data that needs to be transmitted may be smaller than M in the front nodes of links, and it is not the best choice to directly applying the CS coding on every node.

To avoid this problem, a modified hybrid compressed sensing method is discussed in [18]. In the hybrid CS method, each node i is initially set to transmit its raw data directly at first. Then, once an intermediate node j receives more than $(M - 1)$ raw readings or any encoded readings, the intermediate node switches to the CS aggregation mode. In this case, the unnecessary transmissions are avoided in the data gathering by the hybrid CS. Figure 1(b) shows the process of data gathering with hybrid CS, where each node sends out original readings directly if the number of readings is smaller than M. Otherwise, the node codes the data with corresponding coefficient vectors and sends out M-dimension coded vector later. Obviously, the hybrid CS has a higher energy efficiency.

2.2. System Model. As seen in Figure 2, a cluster-based wireless sensor network with N sensor nodes is considered in this paper. The whole network is divided into p clusters. Both cluster members (CMs) and cluster heads (CHs) continuously generate a set of data that need to be collected by the sink. Let $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_N(t)]^T$ denote the observed data from the network at the time instant t and the data vector can also be written as

$$\mathbf{x}(t) = \left[\mathbf{x}_1^T(t), \mathbf{x}_2^T(t), \ldots, \mathbf{x}_p^T(t)\right]^T \tag{3}$$

where $\mathbf{x}_k(t)$ is the raw data vector of the k-th cluster. The member nodes in a cluster selectively send their data to the

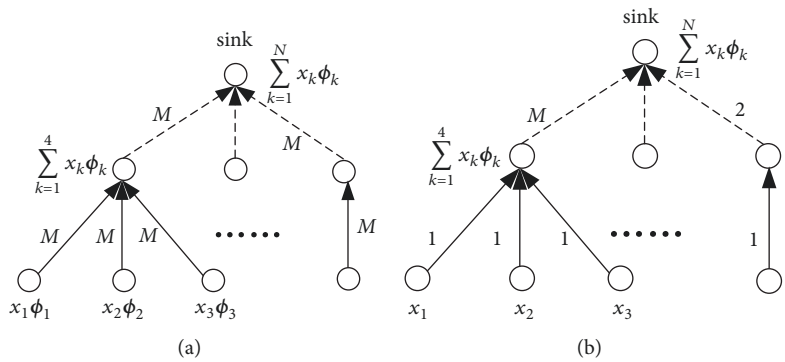

Figure 1: Example of M=3 for difference between pure CS and hybrid CS. (a) Data gathering with pure CS. (b) Data gathering with hybrid CS.

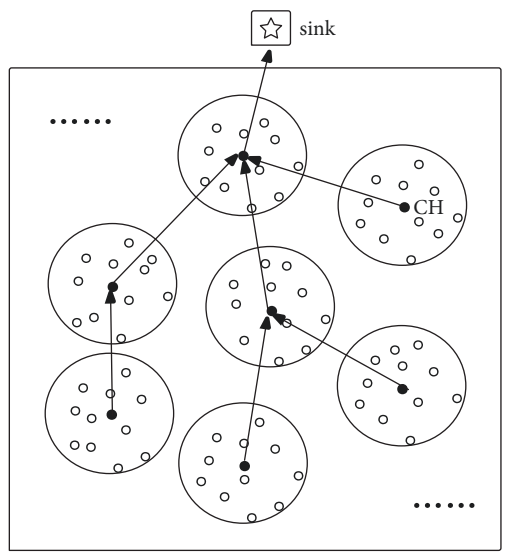

Figure 2: A cluster-based wireless sensor network.

CH with prediction model, and the data vector obtained in all CHs is

$$\begin{aligned}\mathbf{y}(t) &= \left[y_1(t), y_2(t), \ldots, y_N(t)\right]^T \\ &= \left[\mathbf{y}_1^T(t), \mathbf{y}_2^T(t), \ldots, \mathbf{y}_p^T(t)\right]^T\end{aligned} \tag{4}$$

where $\mathbf{y}_k(t)$ is the obtained data vector by the k-th CH. Then the CHs forward all the data to the sink with hybrid compressed sensing via multihop communication. Assume that each CH knows the CS projection vectors of all nodes within the cluster. The M-dimension measurement vector obtained in the sink is $\mathbf{z}(t) = \mathbf{\Phi}\mathbf{y}(t)$ where $\mathbf{\Phi}$ is a constant $M \times N$ measurement matrix which meets the RIP condition. The sink reconstructs the data vector from $\mathbf{z}(t)$ with the CS recovery algorithm, such as basis pursuit (BP) and orthogonal matching pursuit (OMP).

3. Proposed Method

3.1. Overall Description of the Proposed Method. In the proposed method, dual prediction [17] and hybrid compressed sensing are jointed to reduce the energy consumption of data gathering in the clustered WSNs. Moreover, an error threshold selection principle is designed to make the proposed method well suitable for a myriad of environments.

Figure 3 shows the diagram of the proposed method for the cluster-based WSN. In the scheme, dual prediction technique is firstly applied in the intracluster transmission to eliminate the temporal redundancy of measurements. With the prediction method, the cluster members selectively transmit their measurements to the CHs, and the CHs use the forecasted values instead of the actual ones which are not received. Next, the hybrid CS is used in intercluster transmission to reduce the spatial redundancy. With the hybrid CS, the CHs aggregate the obtained data or not and then send the coded/uncoded vectors to the sink via multihop communication. Finally, the sink reconstructs the original data with a CS recovery algorithm. The details of the proposed scheme are given in the following.

Stage 1 (intracluster transmission). At every sampling time t, each node i obtains an observed value $x_i(t)$ from the surrounding environment, and then it forecasts $\widehat{x_i}(t)$ using

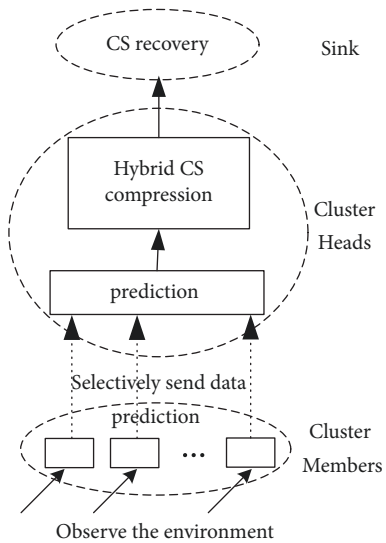

Figure 3: The diagram of the proposed method for the cluster-based WSN.

the preconstructed prediction model. There are two cases that should be considered for the node i as follows:

Case 1. If the error between the forecasted and observed values is over the threshold ε_i, which is prespecified based on the error threshold selection principle (it will be discussed in the Section 3.2), i.e.,

$$\left|\widehat{x_i}(t) - x_i(t)\right| > \varepsilon_i \tag{5}$$

the node i transfers the actual value $x_i(t)$ to its CH.

Case 2. If the forecasted value is acceptable, i.e.,

$$\left|\widehat{x_i}(t) - x_i(t)\right| \leq \varepsilon_i \tag{6}$$

the cluster head calculates $\widehat{x_i}(t)$ with the same prediction model to approximate the actual data. In this case, unnecessary transmissions are avoided.

Stage 2 (intercluster transmission). Let $\mathbf{y}_j(t) = [y_j^1(t), y_j^2(t), \ldots, y_j^{N_j}(t)]^T$ denote the data vector of the cluster head j obtained from the cluster, where N_j is the number of nodes in the cluster j, and $y_j^k(t)$ is an actual value obtained directly or a prediction value calculated by the CH for the k-th node in cluster j. Assume $\mathbf{J} = \{j_1, j_2, \ldots, j_h\}$ is the set of cluster heads which transfer their data to the cluster head j. There are two cases that need to be considered for the cluster head j as follows:

Case 1. If $\sum_{k=1}^{h} N_{j_k} + N_j < M$, the cluster head j sends out the data vector $\mathbf{y}_j^*(t)$ without aggregation, where

$$\mathbf{y}_j^*(t) = \left[\mathbf{y}_{j_1}^T(t), \mathbf{y}_{j_2}^T(t), \ldots, \mathbf{y}_{j_h}^T(t), \mathbf{y}_j^T(t)\right]^T \tag{7}$$

and the number of elements in $\mathbf{y}_j^*(t)$ is

$$q_j = \sum_{k=1}^{h} N_{j_k} + N_j \tag{8}$$

Case 2. If $\sum_{k=1}^{h} N_{j_k} + N_j \geq M$,

$$\mathbf{y}_j^*(t) = \sum_{k=1}^{h} \mathbf{\Phi}^{j_k} \mathbf{y}_{j_k}(t) + \mathbf{\Phi}^j \mathbf{y}_j(t) \tag{9}$$

and

$$q_j = M \tag{10}$$

Stage 3. The sink obtains the data vector

$$\begin{aligned} \mathbf{z}(t) &= \mathbf{\Phi}\mathbf{y}(t) \\ &= \sum_{k=1}^{p} \mathbf{\Phi}^k \mathbf{y}_k(t) \end{aligned} \tag{11}$$

and reconstructs the original data with the CS recovery algorithm, such as basis pursuit (BP) and orthogonal matching pursuit (OMP).

3.2. Error Threshold Selection Principle. As illustrated in Section 3.1, a node i will send its observed value to the CH if $|\widehat{x_i}(t) - x_i(t)| > \varepsilon_i$ where ε_i is a user given error threshold. It means that the value ε_i determines the number of transmissions in every time instant, and it influences the accuracy of data recovery in the sink. Therefore, it is important to select an appropriate error threshold ε_i for the data gathering scheme with prediction.

According to the Central Limit Theorem, we assume unbiased predictions and errors normally distributed. In this paper, we select a simple autoregressive (AR) model to

predict in the proposed method, which is only an example for understanding easily. Moreover, the proposed error selection principle is universal for different prediction methods when prediction error distribution models of these methods can be estimated.

A l-order AR predictor can be denoted as

$$\widehat{x_i}(t) = \sum_{k=1}^{l} \xi_k x_i (t-k) \tag{12}$$

where $\xi_1, \xi_2, \ldots, \xi_l$ are the parameters of the prediction model, which can be constructed by the Yule-Walker equations or the least square method. Assume the observed value can be written as

$$x_i(t) = \sum_{k=1}^{l} \xi_k x_i (t-k) + e_i(t) \tag{13}$$

where $e_i(t)$ is the prediction error with zero mean and variance σ_i^2, i.e., $e_i(t) \sim N(0, \sigma_i^2)$. To a fixed error threshold ε_i, the probability of a node to send its observed data out is

$$\alpha = 2 - 2\Phi\left(\frac{\varepsilon_i}{\sigma_i}\right) \tag{14}$$

where $\Phi(\cdot)$ is the CDF of Gaussian white noise. Clearly, to make sure that the scheme with dual prediction is more energy efficient, the additional computation power must be much less than the reduced energy of transmission; that is,

$$(2-\alpha) E_p < (1-\alpha) E_s \tag{15}$$

where E_p and E_s are the energy costs of a single prediction and sending a value to the CH, respectively. In general, E_s is larger than E_p, and suppose $E_s = kE_p$. Thus, the error threshold needs to satisfy

$$\Phi\left(\frac{\varepsilon_i}{\sigma_i}\right) > \frac{k}{2k-2} \tag{16}$$

In practice, the value σ_i is unknown. Authors in [10] use $\sigma_{x_i}\sqrt{1-\sum_{j=1}^{p}\phi_j \rho_{x_i}(j)}$ instead of the unknown value σ_i, where σ_{x_i} and $\rho_{x_i}(j)$ are the standard deviation and correlation coefficient of a stationary time series $\mathbf{x}_i$, respectively.

Combining (14) and (16), it is obvious that a bigger ε_i leads to lower energy consumption but decreases the recovery accuracy. Therefore, the choice of ε_i must be weighed between the energy consumption and the recovery accuracy.

The energy consumption for one node sending T data to its CH in the proposed method is

$$\begin{aligned} E_T &= (2-\alpha) TE_p + \alpha TE_s \\ &= 2\Phi\left(\frac{\varepsilon_i}{\sigma_i}\right) TE_p + \left[2 - 2\Phi\left(\frac{\varepsilon_i}{\sigma_i}\right)\right] TE_s \end{aligned} \tag{17}$$

The mean square error (MSE) of the process is

$$\text{MSE} = \frac{1}{T}\sum_{j=1}^{T} \eta(j) \times e_i^2(j) \tag{18}$$

where $e_i(j) \le \varepsilon_i$ and

$$\eta(j) = \begin{cases} 0, & \textit{send out } x_i(j) \\ 1, & \textit{not send out } x_i(j) \end{cases} \tag{19}$$

Thus, one can obtain the following:

$$\text{MSE} \le (1-\alpha)\varepsilon_i^2 \tag{20}$$

Formulate the optimization as

$$\begin{aligned} \min \quad & 2\Phi\left(\frac{\varepsilon_i}{\sigma_i}\right) E_p + \left[2 - 2\Phi\left(\frac{\varepsilon_i}{\sigma_i}\right)\right] E_s \\ s.t. \quad & \Phi\left(\frac{\varepsilon_i}{\sigma_i}\right) > \frac{k}{2k-2} \\ & \left[2\Phi\left(\frac{\varepsilon_i}{\sigma_i}\right) - 1\right]\varepsilon_i^2 \le \lambda \varepsilon_i^2 \end{aligned} \tag{21}$$

where λ is a user given parameter which indicates the acceptable maximum recovery accuracy. In practice, the value of λ is between 0 and 1, and it reflects the ratio of expected prediction error to the maximum error. A larger λ means less energy cost (i.e., high energy efficiency) and lower recovery accuracy. The users can determine an appropriate λ depending on the tradeoff between energy efficiency and recovery accuracy for different applications; that is because different applications have different requirements for energy efficiency and recovery accuracy. For example, some applications require high energy efficiency and low recovery accuracy, while others require low energy efficiency and high recovery accuracy. Therefore, the value of λ needs to be determined according to the specific application requirements, and we only give the performance of proposed method under a given value in the simulation.

To solve the above optimization problem in Equation (21), $\Phi(\varepsilon_i/\sigma_i)$ is treated as a variable. In this case, the above optimization is translated into a simple linear programming problem, which can be solved by the existing algorithms [19]. Once $\Phi(\varepsilon_i/\sigma_i)$ is calculated, ε_i can be obtained according to the corresponding σ_i and Gaussian distribution table.

Specifically, to ensure the energy efficiency of the system, a small transmit probability is required. In the meanwhile, the accuracy of intracluster transmission should be guarantee to a certain extent. In this case, a prediction model with higher estimated accuracy (but higher computation complexity in general) should be considered.

3.3. Procedures of the Proposed Method. Algorithm 1 illustrates the proposed threshold selection-based data collection algorithm. As discussed in Section 3.1, the procedure of the proposed method is as the following 4 paragraphs:

Firstly, each cluster member calculates its error threshold according to the error threshold selection principle which is discussed in Section 3.2.

Then, intracluster transmission begins. The cluster member predicts the current value and compares the predicted value with the actual observed one. If the error between them

Each cluster member obtains its error threshold ε_i using (21)
Stage 1: Intra-cluster Transmission
for each cluster member i
  forecasts $\widehat{x_i}(t)$ for the current sampling time t using the preconstructed prediction model;
    calculates the error between $\widehat{x_i}(t)$ and the observed value $x_i(t)$;
  if $|\widehat{x}_i(t) - x_i(t)| > \varepsilon_i$
    the node i transfers the actual value to its cluster heads
  else
    the corresponding cluster head calculates $\widehat{x_i}(t)$ with the same prediction model;
  end
end
Stage 2. Inter-cluster Transmission
After stage 1, each cluster head j obtains the values of its members
  $\mathbf{y}_j^*(t) = [\mathbf{y}_{j_1}^T(t), \mathbf{y}_{j_2}^T(t), \ldots, \mathbf{y}_{j_h}^T(t), \mathbf{y}_j^T(t)]^T$.
for each cluster head j in the multi-hop link
  calculates the number of values which have been obtained by the node j from all front cluster heads $\mathbf{J} = \{j_1, j_2, \ldots, j_h\}$ in the link and its own members;
  if $\sum_{k=1}^{h} N_{j_k} + N_j \geq M$
    $\mathbf{y}_j^*(t) = \sum_{k=1}^{h} \mathbf{\Phi}^{j_k}\mathbf{y}_{j_k}(t) + \mathbf{\Phi}^j\mathbf{y}_j(t)$
  else
    $\mathbf{y}_j^*(t) = [\mathbf{y}_{j_1}^T(t), \mathbf{y}_{j_2}^T(t), \ldots, \mathbf{y}_{j_h}^T(t), \mathbf{y}_j^T(t)]^T$
  end
  sends out $\mathbf{y}_j^*(t)$;
end
Stage 3. Reconstruction
The sink obtains the data vector
$$\mathbf{z}(t) = \mathbf{\Phi}\mathbf{y}(t) = \sum_{k=1}^{p} \mathbf{\Phi}^k \mathbf{y}_k(t)$$
and reconstructs the original data with the CS recovery algorithm.

Algorithm 1: Threshold selection-based data collection algorithm.

is larger than the error threshold, the cluster member sends the observed reading to its corresponding cluster head. Else, the cluster head predicts the current value with the same prediction model.

Next, once the cluster heads collect all data of their members, the intercluster transmission begins. If the number of data that needs to be transmitted for one cluster head is less than M, the cluster head sent these data out directly. Else, the cluster head aggregates these data to M dimension with compressed sensing which is discussed in Section 2.1 and sends the coded data out after.

Finally, the sink reconstructs the original data according to the obtained data vector and some CS recovery algorithms.

Ignore the computation complexity of the reconstruction process. Assume the complexity of the applied prediction model is O_{pred}, and the computation complexity of stage 1 is $(N - p)[(2 - \alpha)O_{pred} + 2]$. The computation complexity of stage 2 is about $O(MN)$ which depends on the complexity of the matrix multiplication $\mathbf{\Phi}\mathbf{y}(t)$. Thus, the total computation complexity of Algorithm 1 is about $(N-p)[(2-\alpha)O_{pred}+2] + O(MN)$.

3.4. Transmission Energy Cost Analysis. In this paper, the reduction of the communication cost means the energy saving while guaranteeing the user-defined data accuracy. As shown in Figure 3, the scheme concludes two communication stages: intracluster transmission with prediction and intercluster transmission with compressed sensing.

Assume that there are N nodes in network (p of them are cluster heads), and each node has T data for gathering. In intracluster transmission, each node costs E_p energy to forecast the current value, and then the node sends its observed data to its cluster with probability α; otherwise, the cluster head needs to predict the data with probability $(1 - \alpha)$. The expected energy cost of each cluster member for this process is $\alpha E_s + (1 - \alpha)E_p$. Therefore, the energy cost of intracluster transmission is

$$E_1 = T\left[(2 - \alpha)(N - p)E_p + \alpha(N - p)E_s\right] \quad (22)$$

In intercluster transmission, the energy cost of each cluster head k for each time is $q_k E_t$, where q_k is the number of data that CH k needs to send out (which is defined in Section 3.1), and E_t is the energy cost for a cluster head sending out data. Therefore, the transmission energy cost of intercluster transmission is

$$E_2 = T\sum_{k=1}^{p}(q_k E_t) \quad (23)$$

where p is the number of clusters in network.

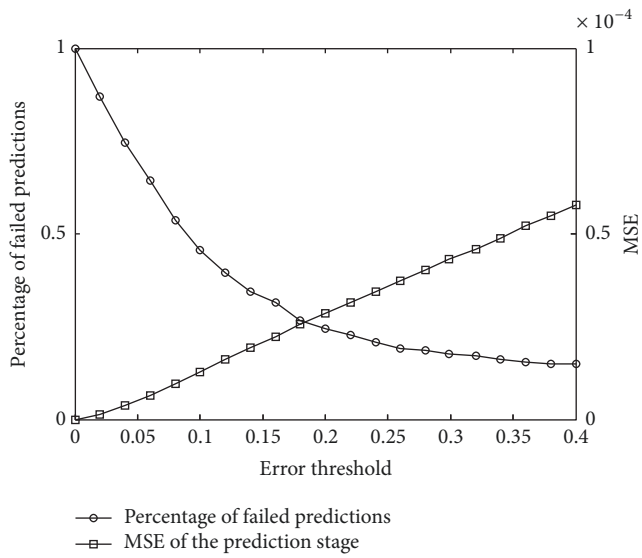

Figure 4: Percentage of transmission nodes and MSE performance under the different error thresholds.

Combining (22) and (23), the total transmission energy cost is

$$E_{total} = T\left[(2-\alpha)(N-p)E_p + \alpha(N-p)E_s + \sum_{k=1}^{p}(q_k E_t)\right] \quad (24)$$

4. Results and Discussion

To evaluate the performance of the proposed method, the cluster-based WSNs with fixed area are considered. For simplicity, the LEACH [20] is used for the clustering approach in the simulation, and the probability for each node to become a cluster head is set to 0.1. In the simulation, the number of nodes N of the synthetic network takes value from 500 to 1500, and each node in the network has T=1000 values that need to be gathered. The parameter λ of the proposed method ranges from 0.01 to 0.95. A simple AR predictor is used in the simulation, and ten history data values can be cached by each node for constructing the parameters of the prediction model. The energy cost model of MICAZ [21] is used for energy estimation of our work. In MICAZ model, the transmission cost of one bit is 600 nJ and the computation energy per clock cycle is 3.5 nJ. Assume that data is with 16 bits. In comparison with the proposed method, the available schemes, i.e., DPPCA [13] and CHCS [18], are also simulated, specifically, the fixed threshold of the DPPCA method which with prediction program is set to 0.1 according to the reference. All simulations are run in the MATLAB software.

Figure 4 shows the percentage of failed predictions and MSE curves versus the different error threshold at one node for the first stage of the proposed method. In Figure 4, MSE is the mean square error of the prediction stage of the proposed method. Clearly, if the prediction fails in a time instant, the node will need to send the raw data to its CH. As the error threshold increases, the percentage of failed predictions decreases and the MSE performance is degraded. When the error threshold is larger than 0.2, the percentage of failed predictions tends to decrease slightly, while the MSE performance is still degraded rapidly, because the prediction error obeys an approximate Gaussian distribution in general.

Figure 5 gives the error threshold curve versus the different for the proposed method λ. In Figure 5, λ is a user given parameter, and a larger λ means a larger MSE performance for the first stage of the proposed method. In the simulation, k=30 is considered for determining the error threshold. From Figure 5, the error threshold increases rapidly when λ is larger than 0.8. According to the MSE performance in Figure 4, the MSE performance will be too poor in this case. Therefore, a reasonable λ should be chosen in practical application.

Figure 6 gives the number of transmissions curves versus the number of clusters in network for the proposed method. From Figure 6, one can see that the total number of transmissions increases as the number of clusters increases in network, because more clusters in network mean more intercluster transmissions, and while prediction model is applied in intracluster transmission, the number of intercluster transmissions is larger. In addition, as the number of clusters increases, the total numbers of transmissions with different λ tend to be the same. It is because the number of intracluster transmissions will be reduced (note that its limit value will tend to zero) as the number of clusters increases, and the value of λ only affects the intracluster transmission, so the gaps between the numbers of intracluster transmissions with different λ become smaller as the number of clusters increases.

Figure 7 shows the total number of transmissions curves versus the number of nodes in the network for different methods. In Figure 7, "normal method" is one without any compression operation during the transmission. From Figure 7, one can see that the transmission energy consumption of the proposed method is smaller than those of the available

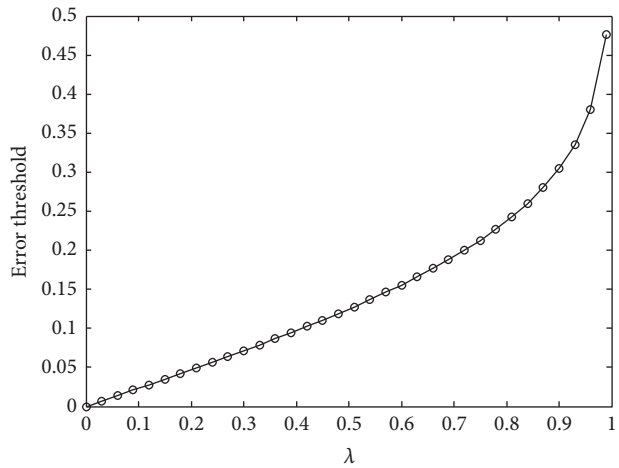

Figure 5: Error threshold against different λ for the proposed method (N=1000).

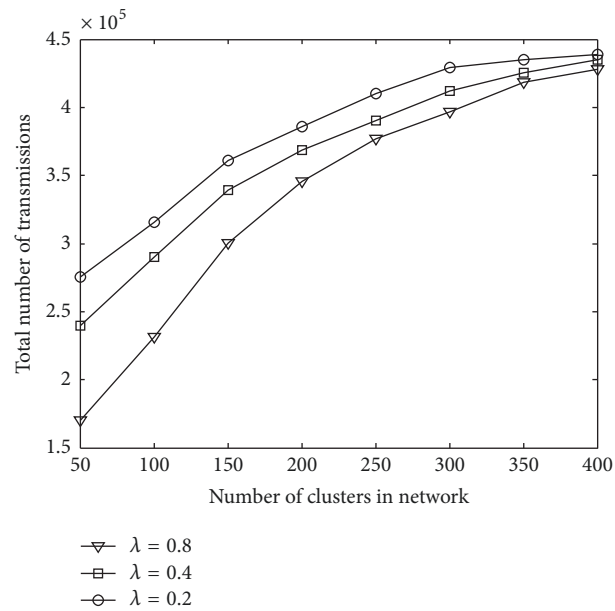

Figure 6: Total number of transmissions against the number of clusters in network for the proposed method (N=1000).

methods. Compared with the normal method, the proposed method reduces the transmission energy consumption by 50%. Moreover, the proposed method can reduce the transmission energy cost about 40% and 20% when compared with the CHCS and DPPCA methods, respectively; that is because the proposed method uses dual prediction with adaptive error threshold and hybrid compressed sensing. In addition, as the number of nodes increases, the proposed method will have better energy efficiency, which also shows that the proposed method is suitable for the large-scale networks.

Figure 8 gives the MSE performance against the number of nodes for different methods. In Figure 8, the CHCS in [18] and the DPPCA in [13] are compared with the proposed method. From Figure 8, one can see that the MSE performance improves as the number of nodes increases. It is because the distribution area for all nodes of the network is assumed to be fixed, and the node density increases with the increase of total number of nodes. In this case, the overlap degree of observed area for nodes is higher, which means that the spatial correlation of observed data increases. Therefore, the MSE performance improves when the compressed sensing with fixed compression ratio is applied in intertransmission stage. Moreover, one can see that the MSE performance of the CHCS method is the best due to its only having the intercluster compression, but it costs the most energy in transmission stage which can be seen from Figure 7. The MSE performance of the DPPCA method is a little better while $\lambda = 0.8$, but the advantage gradually diminished as the network size increases, and the DPPCA method always costs much more energy than the proposed method which can be seen from Figure 7. When λ is smaller than 0.4, the MSE performance of the proposed is better, and meanwhile the proposed method still has a higher energy efficiency. Therefore, taking into consideration the

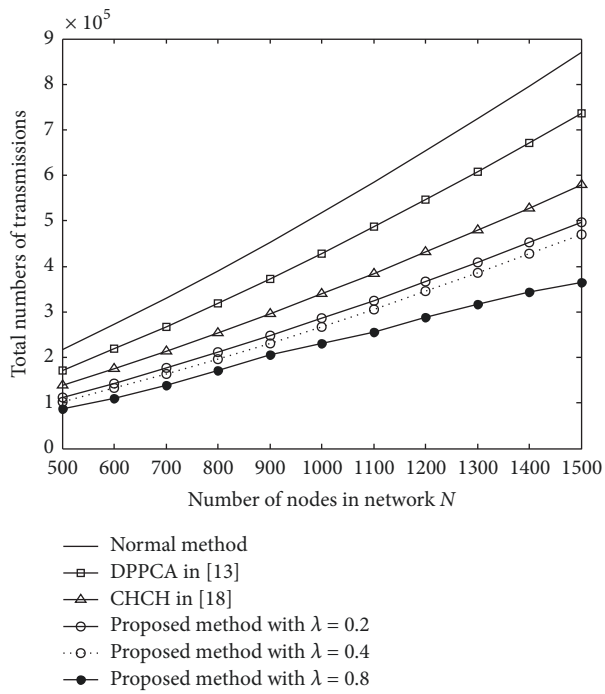

Figure 7: Total number of transmissions against the number of nodes in network for different methods.

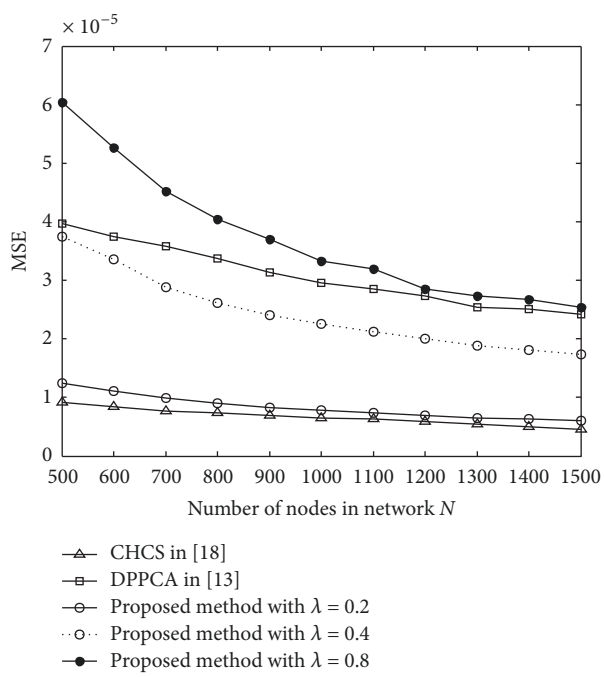

Figure 8: MSE performance of different methods against the number of nodes in network.

energy cost and the MSE performance, the proposed method with adaptive threshold has a distinct advantage in practical applications.

5. Conclusions

In this paper, a novel energy-efficient data gathering scheme exploiting both spatial and temporal correlations is proposed for clustered WSNs. The method uses dual prediction to reduce the temporal redundancy during the intracluster transmission, and the hybrid CS technique is utilized to reduce the spatial redundancy during the intercluster transmission. In this way, the transmission energy cost is decreased greatly. To enhance the availability of the proposed method for different environments, an error threshold selection principle based on the tradeoff between the MSE and the

energy cost is designed. Compared with the CHCS and DPPCA methods, the proposed scheme can greatly reduce the transmission energy cost. In addition, the proposed method offers adequate data recovery accuracy, and it can perform well in large-scale and dense WSNs.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC1314903, in part by the National Natural Science Foundation of China under Grant 61372124 and Grant 61427801, the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 18KJ13510034), and the 11th Batch of China Postdoctoral Science Fund Special Funding (No. 2018T110530).

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–105, 2002.

[2] Y. Gu, F. Ren, Y. Ji, and J. Li, "The evolution of sink mobility management in wireless sensor networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 507–524, 2016.

[3] M. Dong, K. Ota, and A. Liu, "RMER: reliable and energy-efficient data collection for large-scale wireless sensor networks," *IEEE Internet of Things Journal*, vol. 3, no. 4, pp. 511–519, 2016.

[4] C.-F. Cheng and C.-F. Yu, "Data gathering in wireless sensor networks: a combine-TSP-reduce approach," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2309–2324, 2016.

[5] M. Duarte, S. Sarvotham, D. Baron, M. Wakin, and R. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Proceedings of the Conference Record of the Thirty-Ninth Asilomar Conference onSignals, Systems and Computers, 2005.*, pp. 1537–1541, Pacific Grove, California.

[6] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: a survey," *Ad Hoc Networks*, vol. 7, no. 3, pp. 537–568, 2009.

[7] N. Nguyen, D. L. Jones, and S. Krishnamurthy, "Netcompress: Coupling network coding and compressed sensing for efficient data communication in wireless sensor networks," in *Proceedings of the 2010 IEEE Workshop on Signal Processing Systems, SiPS 2010*, pp. 356–361, USA, October 2010.

[8] Y. Zhou, L. H. Yang, and L. X. Yang, "Improved compressed network coding scheme for energy-efficient data communication in wireless sensor networks," *Ksii Transactions on Internet Information Systems*, vol. 9, no. 7, pp. 2488–2511, 2017.

[9] N. A. Alduais, J. Abdullah, A. Jamil, and H. Heidari, "Performance evaluation of real-time multivariate data reduction models for adaptive-threshold in wireless sensor networks," *IEEE Sensors Letters*, vol. 1, no. 6, pp. 1–4, 2017.

[10] H. Jiang, S. Jin, and C. Wang, "Prediction or not? an energy-efficient framework for clustering-based data collection in wireless sensor networks," *IEEE Transactions on Parallel Distributed Systems*, vol. 22, no. 6, pp. 1064–1071, 2010.

[11] U. Raza, A. Camerra, A. L. Murphy, T. Palpanas, and G. P. Picco, "Practical data prediction for real-world wireless sensor networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 8, pp. 2231–2244, 2015.

[12] M. Arunraja and V. Malathi, "Collective prediction exploiting spatio temporal correlation (CoPeST) for energy efficient wireless sensor networks," *KSII Transactions on Internet and Information Systems*, vol. 9, no. 7, pp. 2488–2511, 2015.

[13] M. Wu, L. Tan, and N. Xiong, "Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications," *Information Sciences*, vol. 329, pp. 800–818, 2016.

[14] Y.-C. Wang, Y.-Y. Hsieh, and Y.-C. Tseng, "Multiresolution spatial and temporal coding in a wireless sensor network for long-term monitoring applications," *Institute of Electrical and Electronics Engineers. Transactions on Computers*, vol. 58, no. 6, pp. 827–838, 2009.

[15] L. Quan, S. Xiao, X. Xue, and C. Lu, "Neighbor-aided spatial-temporal compressive data gathering in wireless sensor networks," *IEEE Communications Letters*, vol. 20, no. 3, pp. 578–581, 2016.

[16] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[17] G. M. Dias, B. Bellalta, and S. Oechsner, "A survey about prediction-based data reduction in wireless sensor networks," *ACM Computing Surveys*, vol. 49, no. 3, 2016.

[18] R. Xie and X. Jia, "Transmission-efficient clustering method for wireless sensor networks using compressive sensing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, pp. 806–815, 2014.

[19] V. Chvatal, *Linear Programming*, vol. 5, no. 1, W. H. Freeman and Company, New York, NY, USA, 1983.

[20] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Annual Hawaii International Conference on System Siences (HICSS '00)*, vol. 2, p. 8020, IEEE, January 2000.

[21] B. Gedik, L. Liu, and P. S. Yu, "ASAP: an adaptive sampling approach to data collection in sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 12, pp. 1766–1783, 2007.