S & M 1520

# Application of Energy-efficient Data Gathering to Wireless Sensor Network by Exploiting Spatial Correlation

Ying Li,[1] Xinwang Zheng,[1] Jing Liu,[2] Nina Hu,[2] and Guangsong Yang[2*]

[1]Chengyi College, Jimei University, Jimei Road 199, Xiamen, Fujian 361021, China
[2]Information Engineering School, Jimei University, Yinjiang Road 185, Xiamen, Fujian 361021, China

(Received October 27, 2017; accepted December 18, 2017)

***Keywords***: energy-efficient, spatial correlation, data gathering, wireless sensor network

The paper is focused on the study of energy-saving data gathering strategies based on spatial correlation in wireless sensor networks (WSNs). First, the factors influencing spatial correlation on distortion are discussed, and we prove that representative nodes can be selected to reduce data transmission within a certain range of distortion. Second, the performance of greedy corrected clustering (GCC) and *k*-means algorithms are analyzed. An energy-efficient gathering scheme was proposed by applying GCC and *k*-means algorithm to a typical low-energy adaptive clustering hierarchy (LEACH) protocol; the results of simulation show that the scheme can save energy, reduce distortion, and prolong network lifetime.

## 1. Introduction

A wireless sensor network (WSN) is usually an event-driven system where several nodes try to transmit data when any physical phenomenon of interest is detected. The main objective of the WSN is to reliably estimate event features from the collective information provided by sensor nodes. It is a challenging problem to collect data continuously from a WSN with limited energy and bandwidth.

In a densely deployed WSN, sensing data are likely spatially correlated because one sensor's information can be inferred from its neighboring sensors. Therefore we can remove or reduce the redundancy in the data and reduce communication overhead and energy consumption in a network.

There have been several studies of energy efficient protocols in WSNs. Some approaches seek to optimize communication protocols that spread congestion and energy consumption evenly throughout the network.(1) Many techniques, on the other hand, design a protocol by considering a spatial correlation. Pradhan *et al.* investigated aspects of information theory of correlation in a WSN.(2) Intanagonwiwat proposed a method which exploits spatial correlation inherent in sensor network data combined with a traditional routing protocol.(3) Yang *et al.* applied compress sensing (CS) theory to gather and reconstruct the sparse signals in energy-constrained large-scale WSNs.(4) However, most did not allow for efficient data gathering

[*]Corresponding author: e-mail: gsyang@jmu.edu.cn

http://dx.doi.org/10.18494/SAM.2018.1819

ISSN 0914-4935 © MYU K.K.
http://myukk.org/

which exploit correlations in the WSNs.(5)

In this paper, we propose an energy-efficient data gathering method which divides the network into clusters of spatially correlated sensors by a greedy corrected clustering (GCC) or *k*-means algorithm and suppresses data transmission within a certain distortion level and with minimum energy-expenditure.

The remainder of the paper is organized as follows: The network model and assumptions are introduced, and the factors influencing spatial correlation are discussed in Sect. 2. The correlated clustering methods based on GCC and *k*-means are introduced and a possible way to improve the protocol is proposed in Sect. 3. The results of simulation and relevant analysis are given in Sect. 4. Finally, conclusions are summarized in Sect. 5.

## 2. Correlation Model for WSN and Impact Factors

### 2.1 Model of spatial correlation

The correlation model for information collection by $N$ sensors in an event area is shown in Fig. 1.(6) The sink estimates the event source $S$, according to the observations of the sensor nodes, $n_i$, assuming that the samples are temporally independent. Each observed sample, $X_i$, of sensor $n_i$ is represented as

$$X_i = S_i + N_i, \quad i = 1, \cdots, N, \tag{1}$$

where the subscript $i$ denotes the spatial location of node $n_i$. The event $S_i$ and observation noise $N_i$ are modeled as Gaussian random variables of zero mean and variance $\sigma_S{}^2$ and $\sigma_N{}^2$, respectively.

The correlation model is a power exponential model, as expressed by(6)

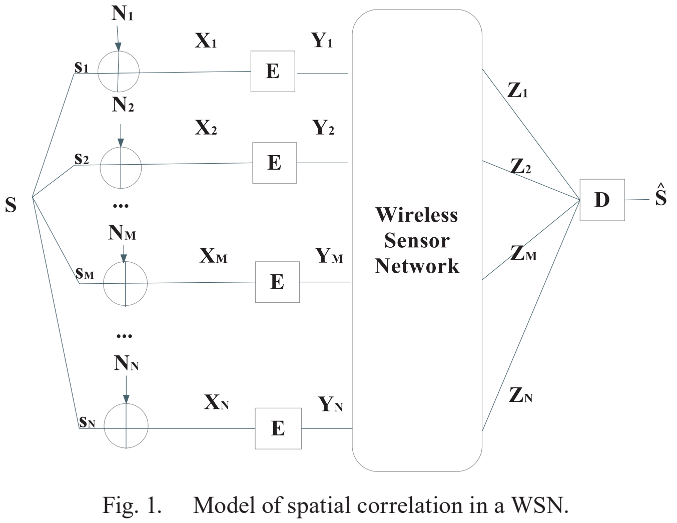

Fig. 1. Model of spatial correlation in a WSN.

$$\rho(i,j) = \frac{E\left[S_i S_j\right]}{\sigma_s^2} = e^{(d(i,j)/\theta)^\alpha}, \tag{2}$$

where $\rho(i,j)$ and $d(i,j)$ are the correlation coefficient and distance between nodes and $n_i$ and $n_j$, respectively. The parameters $\alpha = 1$ and $\theta$ are the controlling parameters for the correlation range between sensors.

Each node encodes its $X_i$ as $Y_i = f_i(X_i)$ and sends it to the sink through the WSN.

$$Y_i = \sqrt{\frac{P_E}{\sigma_S^2 + \sigma_N^2} X_i}, \; i = 1, ..., N, \tag{3}$$

where $P_E$ is power constraint. The encoders and the decoders are labeled $E$ and $D$ in Fig. 1, respectively. The sink decodes each $Y_i$ using the minimum mean squared error (MMSE) estimator, so $\hat{S}$ is expressed as(7)

$$\hat{S}(M) = \frac{1}{M}\sum_{i=1}^{M} Z_i = \frac{1}{M}\sum_{i=1}^{M} \frac{\sigma_S^2}{\sigma_S^2 + \sigma_N^2}(S_i + N_i). \tag{4}$$

The distortion achieved by using $M$ packets to estimate the event $S$ is given as

$$\begin{aligned} D(M) &= E[(S - \hat{S}(M))^2] \\ &= \sigma_S^2 - \frac{\sigma_S^4}{M(\sigma_S^2 + \sigma_N^2)}(2\sum_{i=1}^{M}\rho_{(s,i)} - 1) + \frac{\sigma_s^6}{M^2(\sigma_S^2 + \sigma_N^2)^2}\sum_{i=1}^{M}\sum_{j\neq 1}^{M}\rho_{(i,j)}. \end{aligned} \tag{5}$$

### 2.2 Factors impacting spatial correlation

In WSNs, sensor nodes are usually distributed in a zone and the related information is sent to the sink for centralized processing. When a certain condition (such as temperature, humidity, etc.) exists, the nodes under that condition are aware of this information. There is a strong correlation between the nodes that are close to each other. Therefore, to satisfy the sensing precision, some nodes are chosen, as representative nodes (RN), to send their data rather than having data sent by all nodes in the network. In this way, correlation clusters are formed by taking the representative node as the center; the correlation radius is $R_c$.

By spatially correlating sensing data, the energy consumption of data transmission and collision between sensor nodes will be reduced greatly. Selecting the minimum representative nodes among several nodes is crucial and can be represented as the following.

$$M^* = \arg\min_M\{D(M) < D_{max}\}, \tag{6}$$

where $D_{max}$ is the maximum allowable distortion.

In an area of 500 × 500 $m^2$, 50 nodes were distributed randomly, and some nodes were chosen as the representative nodes. Using the model in Eq. (2), $\theta$ was taken for 10–1000. For each value of $\theta$ the sink calculated the distortion between the collected information from representative nodes and the actual information from all nodes according to Eq. (5). The results are shown in Fig. 2.

From Fig. 2, we can see that as $\theta$ and the number of representative nodes increase, the observed event distortion decreases because of the highly redundant data sent by the sensor nodes that are close to each other.

Moreover, for a fixed number of representative nodes, the minimum distortion can be achieved by choosing the nodes which are located as close to the source as possible and as far apart from each other as possible.

Therefore, we can exploit the spatial correlations between sensing data by choosing appropriate representative nodes among all the nodes to reduce the data forwarded to the sink. This method cannot only save energy without degrading the achieved distortion at the sink but can also reduce the conflict within the wireless medium.

## 3. Data Gathering Method Based on Spatial Correlation

### 3.1 Greedy corrected clustering method

As we discussed in Sect. 2.2, correlation of data can be considered in the design process of data gathering. The GCC algorithm is a clusters method useful for correlation.(8) When the GCC algorithm is used, sets of correlated clusters are formed and all the nodes within a cluster are considered highly correlated. The information is observed by multiple sensor nodes in the event area creating redundant reports. Only a few nodes need to report the sensory data, and the remaining nodes can remain in a silent state to save energy.

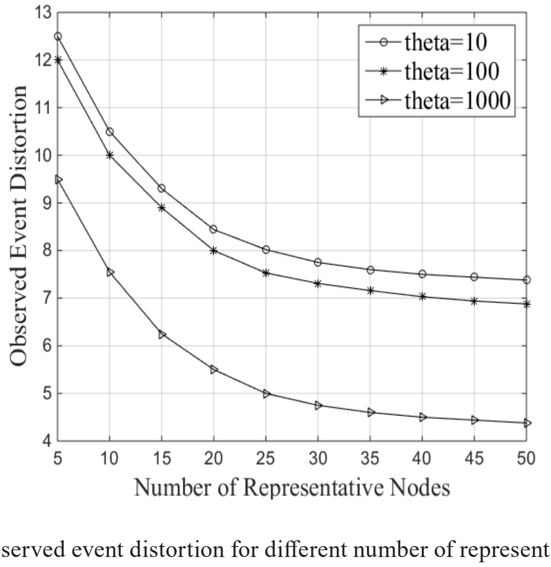

Fig. 2. Observed event distortion for different number of representative nodes.

Suppose we are given $n$ points, and want to find $k$ clusters. A cluster is a subset of the $n$ points, called $C_j$. The GCC algorithm is show in Table 1.

First, randomly select a node from all nodes and a node $j$ from the no-cluster nodes. Then calculate the distance $d(i,j)$ between the two nodes. Put the nodes which meet the condition $d(i,j) < \xi$ into the set $C_k(i,j)$; $k$ is the number of the cluster; $\xi$ is correlation threshold. If the node does not satisfy the conditions, choose two of the most relevant nodes from the cluster members and perform the above steps until the nodes form cluster $k$.

### 3.2 *K*-means method

*K*-means clustering is another algorithm used to classify or group objects based on features into $k$ numbers of groups, some criteria such as Bayesian information criterion (BIC) or minimum description length (MDL) can be used to estimate $k$ automatically.

Given a set $X$ of $n$ points in a $d$-dimensional space and an integer $k$, the task is choosing a set of $k$ points $\{c_1, K, c_k\}$ in the $d$-dimensional space to form clusters $\{C_1, C_2, \ldots, C_k\}$ such that Eq. (7) is minimized.

$$Cost(C) = \sum_{i=1}^{K} \sum_{x \in C_i} L_2^2 (x - c_i) \tag{7}$$

The method is very simple to implement. The grouping is done by minimizing the sum of the squares of the distances between a data point and the corresponding cluster centroid. Therefore, the purpose of $k$-means clustering is to classify data relatively evenly. The process of the $k$-means algorithm is shown in Table 2.

Table 1
GCC algorithm.

| | |
|---|---|
| Input: | The coordinates of all nodes. |
| Output: | Cluster $C_k(i,j)$ |
| Step 1. | Start: $k = 1$. |
| Step 2. | Choose a node $n_i$ from the no-cluster nodes. Calculate the distance $d(i,j)$ between nodes $n_i$ and $n_j$. |
| Step 3. | If $d(i,j) < \xi$ go to Step 5. |
| Step 4. | Put $n_i$ and $n_j$ into $k$ clusters $C_k = \{n_i, n_j\}$. |
| Step 5. | If $n_i$ (no-cluster node) exists, go to Step 2. |
| Step 6. | Stop. |

Table 2
*k*-means algorithm.

| | |
|---|---|
| Input: | The coordinates of all nodes. |
| Output: | Clusters $C_k(i,j)$ |
| Step 1. | Randomly pick $k$ cluster centers $\{c_1, K, c_k\}$. |
| Step 2. | For each $i$, set the cluster $c_i$ to be the set of points in $X$ that are closer to $c_i$ than they are to $c_j$ for all $i \neq j$. |
| Step 3. | For each $i$, let $c_i$ be the center of cluster $c_i$ (the mean of the vectors in $c_i$). |
| Step 4. | Repeat until convergence. |

### 3.3 Proposed data gathering scheme

The data gathering model is a hierarchical model shown in Fig. 3. In the first level, the entire sensor field is divided into several correlated sub-regions and a subset of nodes is selected as representative of the regions using the GCC or *k*-means algorithm. In the second level, these RNs later execute a dynamic low-energy adaptive clustering hierarchy (LEACH) algorithm to gather data during each round. In each round, only the representative nodes collect data using the dynamic clustering protocol.

The operation of our scheme is divided into rounds. Each of these rounds consists of 2 phases: a set-up phase and a steady-state phase. During the set-up phase, cluster-heads are determined and the clusters are organized. During the steady-state phase, data transference to the base station occurs. Our scheme works with rounds in the same way as a typical LEACH protocol.

1) At the beginning of the network setup phase, the sink advertises a broadcast packet including the information of correlated radius $R_c$ to all nodes, correlated clusters are formed using the GCC or *k*-means algorithm, and RNs are selected from all nodes.
2) In each correlated cluster, the RN receives the raw sensing data from other ordinary nodes (ON) and calculates the accuracy of the information between the RN and each ON to determine whether it meets the distortion constraint. If the distortion of an ON is larger than the threshold, the node is labeled as non-correlated node (NCN) and forms a new cluster independently.
3) The sink collects all information on the average distortion of every cluster and number of NCNs from RNs and determent whether $R_c$ is appropriate. If $R_c$ is not, an updated value of $R_c$ is broadcast again.

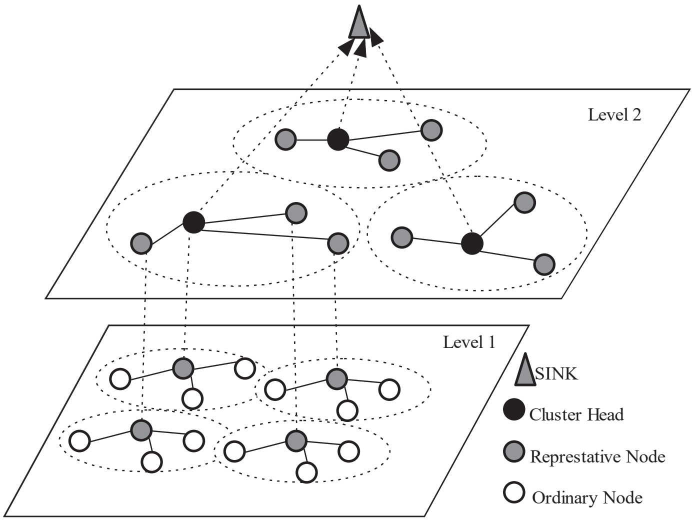

Fig. 3. Hierarchical model of data gathering based on spatial correlation.

## 4. Simulation and Evaluation

### 4.1 Performance of GCC and *k*-means

Using the scenario in Fig. 3, 200 nodes are randomly distributed in a region 200 × 200 m$^2$, and the sink is set in the center. The GCC algorithm and the *k*-means algorithm are used to cluster the nodes. Each correlation cluster member in the cluster is connected with its cluster head, and the cluster head serves as the representative node to which is sent the collected data from other members. The correlation model is an exponential model as shown in Eq. (2).

The result of clustering GCC and *k*-means are shown in Figs. 4 and 5, respectively. We can see from Figs. 4 and 5 that nodes are well-distributed when the *k*-means algorithm is used.

The distortions of correlation clustering are show in Fig. 6. With the increase in the number of nodes, the average distortion tends to decrease gradually, both in GCC and *k*-means. This

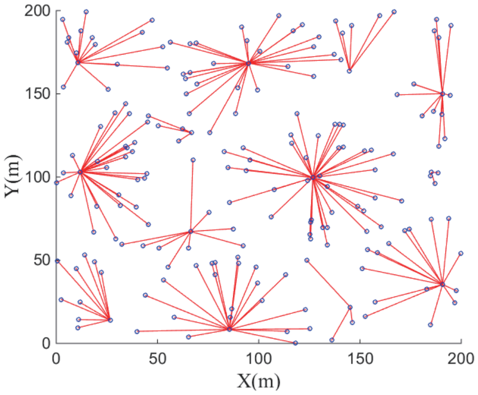

Fig. 4. (Color online) GCC correlation clustering ($R_c$ = 50).

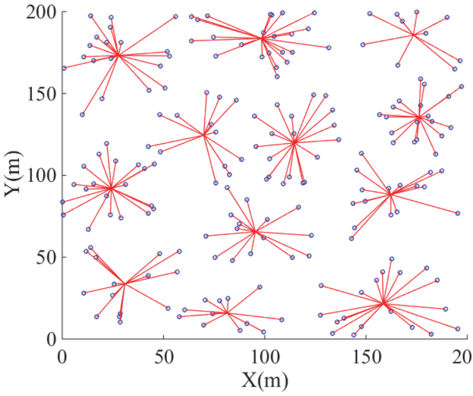

Fig. 5. (Color online) *k*-means correlated clustering ($K$ = 12).

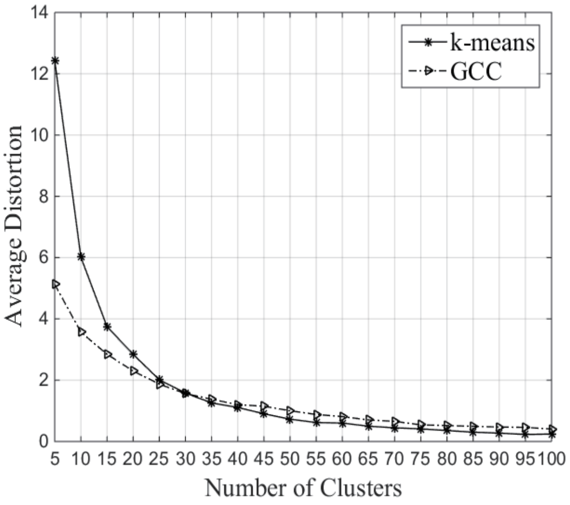

Fig. 6. Distortion of GCC and *k*-means.

means that a higher cluster density indicates a stronger correlation among data points which results in a better estimation.

These results show that when the number of clusters is small, the average distortion of the *k*-mean algorithm is larger, but the distortion becomes smaller when the number of clusters increases. This occurs because the *k*-means algorithm ensures that the average distance of the sensor nodes to their corresponding centroid is the same, so that the final location of the centroid is a given distance from each sensor node, resulting in less distortion.

### 4.2 Energy efficiency of proposed scheme

We evaluated the effectiveness of our scheme with simulations. In a simulation, *N* sensor nodes are randomly distributed in a square region 200 × 200 m$^2$ in size with a sink in the center of the region. The parameters used in the simulation are summarized in Table 3.

The results in Fig. 7 show the relationship between the number of sensor nodes that remained alive and the number of rounds. It can be seen from the figure that the life-span of the WSN using correlation is longer than that of a traditional LEACH.

In our scheme, after the node selection phase during each round, only the representative nodes remain active while non-representative nodes go to sleep. Consequently, the number of active nodes is much smaller than that of LEACH. Since only the RNs participate in the data dissemination, the number of data transmissions is greatly decreased. Therefore, the energy consumption is greatly reduced.

It also seen in Fig. 7 that the lifetime using *k*-means is longer than that using GCC, because the *k*-means can find the optimal cluster size to minimize the maximum distance between any point and its nearest centroid.

Table 3
Simulation parameters.

| Parameter | Meaning | Value |
|---|---|---|
| $n$ | Size of data packet | 2000 b |
| $a$ | Size of control packet | 100 bit |
| $E_{DA}$ | Aggregate energy consumption | 5 nJ/b/signal |
| $E_{schedule}$ | Energy consumption of schedule | 5 nJ/b/signal |
| $E_{init}$ | Initial energy | 10 J |
| $P_r$ | Receive power | 1 mJ/b |
| $B_t$ | Threshold of battery | 1 J |

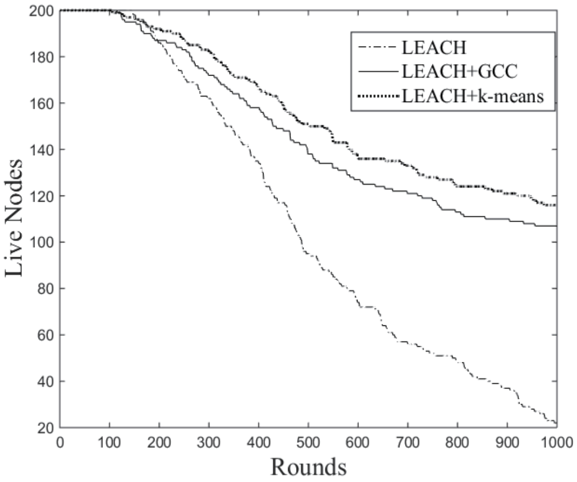

Fig. 7. LEACH combined with GCC and *k*-means.

## 5. Conclusions

Spatial correlation between the data collectors is not only effectively used to ensure that the distortion lies within a certain range, but also to avoid transmitting too much data and consuming too much energy.

Our work shows that the energy consumption of the nodes can be decreased and the lifetime of the system increased with an acceptable level of distortion in data by exploiting spatial correlation, and the transmission of redundant nodes can thereby be controlled.

Future work includes the study of some adjustable scheme to achieve the adaptive correlated radius, finding a dynamic optimal correlated radius, and determining the optimal numbers of correlative clusters. Therefore, some approaches to efficient medium access and reliable event transport by exploiting spatial correlation in WSNs will also be considered.

## Acknowledgments

The paper is supported by the Science Foundation of Fujian Province (No. 2015J01267), the Science Foundation of Fujian Education Department (Nos. JA15650 and JA15257), and the Training Program of Fujian Excellent Talents in University.

## References

1 N. D. Pham, T. D. Le, K. Park, and H. Choo: Int. J. Commun. Syst. **23** (2010) 1311.
2 S. S. Pradhan, J. Kusuma, and K. Ramchandran: IEEE Signal Process. Mag. **19** (2002) 51.
3 C. Intanagonwiwat, R. Govindan, and D. Estrin: 6th Annu. ACM/IEEE Int. Conf. Mobile Computing and Networking (MOBICOM) 56.
4 G. S. Yang, M. B. Xiao, and S. Q. Zhang: J. Networks **8** (2013) 197.
5 A. Al-qamaji and B. Atakan: 25th Signal Processing and Communications Applications Conf. (2017).
6 J. O. Berger, V. D. Oliviera, and B. Sanso: J. Am. Stat. Assoc. **96** (2001) 1361.
7 M. C. Vuran, Ö. B. Akan, and I. F. Akyildiz: Comput. Networks **45** (2004) 245.
8 R. K. Shakya, Y. N. Singh, and N. K. Verma: IET Wireless Sens. Syst. **3** (2013) 266.