Data Collection Optimization Method for Wireless Sensor Networks Based on Linear Regression

IOP Conf. Series: Materials Science and Engineering **569** (2019) 032064 doi:10.1088/1757-899X/569/3/032064

Data Collection Optimization Method for Wireless Sensor Networks Based on Linear Regression

Meng Zhang*, Xiaomei Zhang and Yinghui Huang

School of Information Science and Technology, Nantong University, Nantong, Jiangsu, 226019, China

*18036160804@163.com

Abstract. In the event monitoring applications of wireless sensor networks, the sensing data collected by sensor nodes near the same monitoring area has a great spatial-temporal correlation. In order to reduce the amount of data transmission in the network and the energy consumption of communication among nodes, an energy-efficient distributed data collection optimization strategy based on linear regression for wireless sensor networks is proposed. linear regression model of local sensing data is constructed to represent and predict the actual sensing data monitoring values of sensor nodes. Within the allowable range of errors, the node does not need to transmit the actual monitoring sensing data to the sink node, but only transmits the parameter information of the regression model basis function. Without losing the basic structural characteristics of data, the communication overhead caused by frequent data transmission between sensor nodes is effectively reduced, and the linear regression model of sensing data adopts the incremental update method with low computational complexity. he simulation results show that the data collection optimization strategy based on linear regression can effectively predict and estimate perceptual data with less network energy consumption, and achieve the goal of reducing network energy consumption.

1. Introduction

Data collection is the basic function of wireless sensor networks and the basis of most monitoring applications. The main research goal of data collection technology in wireless sensor networks is to reduce the energy consumption of the network, prolong the life cycle of the network, and avoid the huge overhead of redeploying the wireless sensor network monitoring system in the process of data collection.

At the Moboicom 2002 meeting, Deborah Estrin pointed out in an invitation report that the energy required for a sensor node to transmit 1 bit of information 100 m away is equivalent to the energy consumed to execute 3,000 computational instructions. Sahingoz compared the power consumption of Mica2dot nodes in communication and computation modes, and found that transmitting 1 bit data is equivalent to running 2090 clock cycles of node microcontrollers. It proves that the energy consumption of nodes is much less than that of communication. That is to say, the main factor affecting the total energy consumption of wireless sensor networks is the communication energy consumption in networks [1].On the basis of node clustering, Slepian-Wolf bound coding and other distributed source coding techniques are used to compress the sensing data information and optimize the information rate allocation in the cluster, so as to minimize the communication energy consumption [2] [3].Zhang J, Tang J, Chen F synthesize the advantages of prediction model and clustering technology, and propose a hierarchical data collection framework for wireless sensor

networks based on integrated adaptive prediction model. In this framework, cluster head implements data collection and effective prediction analysis of cluster nodes. According to the analysis results of network state and performance, the energy consumption of communication and prediction calculation is balanced, and the prediction model is adaptively selected to achieve energy-efficient data aggregation processing in wireless sensor networks[4]. If considering from the aspect of node scheduling, it is also an effective way to save energy consumption by allocating and optimizing the active slots of nodes, using stochastic Petri net model and designing reasonable scheduling strategy to find a more suitable sleep wake-up mechanism for nodes[5][6][7].

Regression analysis is a statistical analysis method to determine the quantitative relationship between two or more variables. In the application of event monitoring sensor networks, sensing data has temporal and spatial correlation to a certain extent, which reduces the network energy consumption caused by a large amount of data transmission. In this paper, a linear regression model is established based on historical data measured by regression analysis method. By solving the parameters of the corresponding base function of the model, the nodes only transmit the relevant parameter vectors, which reduces the amount of redundant data transmission. Moreover, the model can receive new actual measurements by simple incremental updating method, and the parameters of the basis function can also be solved by the linear regression model constructed at any time.

2. Distributed linear regression model for sensor networks

According to the specific application environment of the network and the performance indicators of storage space and processing capacity of sensor nodes, the nearest sensing datas of sensor nodes in a certain time interval are selected. Assuming that where and represent sampling time points, can be affected by measurement errors. Using perceptual datas to construct functions. The approximation error is very small (within the confidence interval of the data collected by the monitoring system). The form of function depends on the specific problem. Here can be expressed in the form of equation (1):

$$p(t) = \sum_{j}^{n} \lambda_j A_j(t) \tag{1}$$

The number of neutral terms is n and the specific basis function A_j depend on the actual problem. In general, the selected basis function can be $A_j(t) = t^{j-1}$. Then the equation (1) can be expressed as the *t*-th polynomial of $n-1$, that is:

$$p(t) = \lambda_1 + \lambda_2 t + \lambda_3 t^2 + ... + \lambda_n t^{n-1} \tag{2}$$

Choosing $n = m$ can calculate the corresponding value exactly, but it is easy to interfere with the data when calculating higher-order functions. When predicting the corresponding p value of unforeseen t, its accuracy will be affected. It is better to choose a value far less than m, that is $n << m$.By choosing the values of coefficients, the estimated values of the functions corresponding to the measured values are obtained. In wireless sensor network applications, a third-order polynomial function model is constructed based on the assumption that 50 temperature measurements recently collected by nodes are selected: $p(t) = \lambda_1 + \lambda_2 t + \lambda_3 t^2 + \lambda_4 t^3$; the estimated measurement value $p_i(i = 1,2,3,...,50)$ is enough. And the nodes do not need to transmit 50 actual measurements. After constructing the function model, only four parameter values, namely $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, need to be transmitted in the network as the compressed representation of measurement values, so the amount of information transmission in the network is reduced. If the coefficients are obtained by linear regression model, the polynomial representation model needs to be transformed into matrix representation. In this way, the node does not need to solve the higher order polynomial solution, but only needs to maintain the correlation matrix. Let the coefficient n dimension vector be

$\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)^T$ .The *m*-dimension vector of the actual measured value is $p = (p_1, p_2, ..., p_m)^T$ .The basis function matrix of the corresponding time sampling point t_i is:

$$U = \begin{bmatrix} A_1(t_1) & A_2(t_1) & ... & A_n(t_1) \\ A_1(t_2) & A_1(t_2) & ... & A_n(t_2) \\ ... & ... & ... & ... \\ A_1(t_m) & A_2(t_m) & ... & A_n(t_m) \end{bmatrix}$$

For matrix element $m_{ij} = A_j(t_i)$ $m_{ij} = A_j(t_i)$, the predictive function m-Dimension vector $p = (p(t_1), p(t_2), ..., p(t_m))^T$ of equation (1) at t_i sampling time point is expressed as equation (3):

$$P\begin{bmatrix} P(t_1) \\ P(t_2) \\ ... \\ P(t_m) \end{bmatrix} = U\lambda \begin{bmatrix} A_1(t_1) & A_2(t_1) & ... & A_n(t_1) \\ A_1(t_2) & A_1(t_2) & ... & A_n(t_2) \\ ... & ... & ... & ... \\ A_1(t_m) & A_2(t_m) & ... & A_n(t_m) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ ... \\ \lambda_m \end{bmatrix} \quad (3)$$

Then the approximation error vector Δ can be expressed as an equation, i.e.

$$\Delta = U\lambda - p \quad (4)$$

In order to minimize the approximation error Δ of the estimated value, the optimal objective is to minimize the norm of the approximation error vector:

$$Min(\|\Delta\| = (\sum_{i=1}^{m} \Delta_i^2)^{1/2}) \quad (5)$$

Combining equation (4) and (5) to optimize the objective, we can get:

$$Min(\|\Delta\|^2 = \|U\lambda - p\|^2 = \sum_{i=1}^{m} (\sum_{i=1}^{m} u_{ij}\lambda_j - p_i)^2) \quad (6)$$

By calculating the differential of $\|\Delta\|^2$ for each $\lambda_k (k = 1, 2, ..., n)$ and making the result 0, the minimum value of $\|\Delta\|$ can be obtained:

$$\frac{d\|\Delta\|^2}{d\lambda_k} = \sum_{i=1}^{m} 2(\sum_{i=1}^{m} u_{ij}\lambda_j - p_i)u_{ik} = 0, k = [1, n] \quad (7)$$

According to equation (4), the following matrix equation equivalent to equation (7) can be deduced, namely:

$$(U\lambda - y)^T U = 0 \quad (8)$$

$$U^T(U\lambda - y) = 0 \quad (9)$$

$$U^T U\lambda = U^T y \quad (10)$$

Because the defined base function is $A_j(t) = t^{j-1}$, the base function matrix U is column full rank matrix. For any column full rank matrix U , we can get that $U^T U$ is positive definite, so $U^T U$ exists. According to equation (10), we can get the D solution of coefficient vector as follows:

$$\lambda = (U^T U)^{-1} U^T y \tag{11}$$

Set

$$O = U^T U = \begin{bmatrix} < A_1 \cdot A_1 > & < A_1 \cdot A_2 > & \dots & < A_1 \cdot A_n > \\ < A_2 \cdot A_1 > & < A_2 \cdot A_2 > & \dots & < A_2 \cdot A_n > \\ \dots & \dots & \dots & \dots \\ < A_n \cdot A_1 > & < A_n \cdot A_2 > & \dots & < A_n \cdot A_n > \end{bmatrix} \tag{12}$$

$$k = U^T p = \begin{bmatrix} < A_1 y > \\ < A_2 y > \\ \dots \\ < A_n y > \end{bmatrix} \tag{13}$$

According to equation (12), (13), equation (11) can be written as: $\lambda = O^{-1} k$, that is:

$$O\lambda = k \tag{14}$$

Among them, O is the quantity product matrix of the base function and k is the projection of the base function of the measured value vector. So far, the optimal regression coefficients can be obtained by solving the typical linear system of Equal Formula (14) with known measured values and basis functions.

3. Model parameter optimization

In the application of event monitoring in wireless sensor networks, with the extension of monitoring time, the amount of monitoring data collected by sensor nodes is also increasing. Due to the limitation of energy, storage and processing capacity of sensor nodes, the nodes can only store sampled data for a certain period of time. When the linear regression model is used to calculate the data representation coefficients, the updating operation of the model can be calculated in the following incremental way[8].

Assuming that the number product matrix O and the projection vector k of the basis function in the sampling period from t_1 to t_{m-1} have been calculated, the new measurements in t_m are as follows:

$$O(t_m) = \begin{bmatrix} < A_1(t_m) \cdot A_1(t_m) > & < A_1(t_m) \cdot A_2(t_m) > & \dots & < A_1(t_m) \cdot A_n(t_m) > \\ < A_2(t_m) \cdot A_1(t_m) > & < A_2(t_m) \cdot A_2(t_m) > & \dots & < A_2(t_m) \cdot A_n(t_m) > \\ \dots & \dots & \dots & \dots \\ < A_n(t_m) \cdot A_1(t_m) > & < A_n(t_m) \cdot A_2(t_m) > & \dots & < A_n(t_m) \cdot A_n(t_m) > \end{bmatrix}$$

$$k(t_m) = \begin{bmatrix} < A_1(t_m) y(t_m) > \\ < A_2(t_m) y(t_m) > \\ \dots \\ < A_n(t_m) y(t_m) > \end{bmatrix}$$

Then the number product matrix and projection vector of the basis function in the new sampling period are as follows:

$$O \leftarrow O + O(t_m); k \leftarrow k + k(t_m) \tag{15}$$

The sliding window mechanism is used for matrix O and vector k scale control[9].The system considers the calculation, storage capacity and application requirements of nodes to set the sliding window size.With the increasing scale of matrix O and vector k, when the data of time t exceeds the setting of sliding window, the updated matrix O and vector k can be calculated according to formula (16).

$$O \leftarrow O - O(t_m); k \leftarrow k - k(t_m) \tag{16}$$

To sum up, the node can take the regression coefficient $O\lambda = k$ by calculating the linear system, and the matrix and vector parameters of the linear regression system model can be updated incrementally.

4. Experimental test and performance analysis

In order to test the performance of the proposed distributed data collection optimization algorithm, simulation experiments are carried out for network energy consumption analysis. The test uses NS2 network simulation tools to build wireless sensor network scenarios. The algorithm was added to the improved LEACH protocol, and the typical LEACH [10] and LEACH-C protocols [11] were compared in data acquisition. At the same time, the optimization effect of network energy consumption using the two protocols was analyzed.

In order to test the impact of clustering-based WSN data acquisition algorithm on the network life cycle and overall energy consumption, the network simulation tool NS2 is selected as the simulation platform[12].LEACH protocol is a typical distributed clustering routing protocol. Its hierarchical data forwarding mechanism produces much less network energy consumption than planar routing protocol. In the experiment, the linear regression optimization process is added to the LEACH improved protocol. While the cluster head node receives the sensing information of the nodes in the cluster, it also calculates the linear regression model, and replaces the original sensing data with the parameter information of the transmission regression model. Cluster head implements the estimation, prediction and fault-tolerant processing of sensing data. The protocol still retains the LEACH cluster head selection method[13]. However, change the direct communication between cluster head node and Sink node to multi-hop data transmission between cluster heads. In the experiment, 100 sensor nodes are randomly deployed in the plane area of 100m*100m. The location coordinates of the base station are set to (50,80). The initial energy of each node is 2J. The energy attenuation model is shown in Fig.1.

According to the principle of wireless communication, the transmission power decreases exponentially with the increase of transmission distance. If the distance between the sending node and the receiving node is l, when l is less than the constant threshold l_{Thres}, the transmission power decreases as l^2, that is, the free space attenuation model. When l is greater than l_{Thres}, the transmission power decreases as l^4, i.e. the multi-path attenuation model. The energy consumption $E_T(k,l)$ generated by transmitting k bits data is composed of two parts, $E_{T-elec}(k)$ energy consumption of transmitting circuit and $E_{T-am}(k,l)$ energy consumption of power amplifier, as shown in equation (17). The energy consumption $E_R(k)$ generated by receiving k bits of data is only caused by the energy consumption of the circuit, as shown in equation (18).

$$E_T(k,l) = E_{T-elec}(k) + E_{T-am}(k,l) = \begin{cases} k \times E_{elec} + k \times \xi_s \times l^2 & l < l_{Thres} \\ k \times E_{elec} + k \times \xi_m \times l^2 & l \geq l_{Thres} \end{cases} \tag{17}$$

$$E_R(k) = E_{R_{T-elec}}(k) = k \times E_{elec} \tag{18}$$

E_{elec} is the energy consumption of transmitting (receiving) 1 bit data. ξ_s and ξ_m represent the energy required to transmit 1 bit data power amplifier under free apace attenuation model and multi-path attenuation model, respectively.

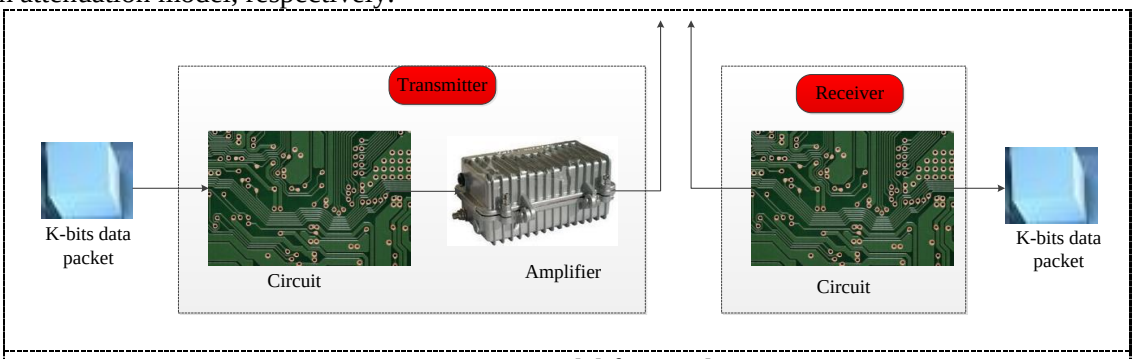

Figure 1. Energy attenuation model for wireless communication.

In addition, the energy consumption of the first regression model for cluster head nodes is: $E_{re} = n_{CH} \times E_{com}$, where n_{CH} denotes the number of cluster head nodes and E_{com} denotes the energy consumption of the first regression model for cluster head nodes. The simulation assumes that $n_{CH} = 5$, $E_{elec} = 50nJ / bit$, $\xi_s = 10pJ / bit / m^2$, $\xi_m = 0.0013pJ / bit / m^2$, $E_{com} = 5nJ / bit$, bandwidth is 1Mbps, message length is 500 bytes, sending and receiving delay is $25\mu s$, simulation time is 500 seconds, the time interval for each round of cluster head election is 20 seconds, the number of linear regression model data sampling is 20, and the update period of regression model is 60 seconds.

Figure 2. shows the total energy consumption of cluster head nodes with LEACH protocol and linear regression strategy at simulation intervals of 20 seconds. From the experimental results, it can be seen that the total energy consumption of each round of cluster head nodes in LEACH protocol is between 4J and 5J, and the total energy consumption of each round of cluster head nodes in linear regression strategy is between 1.5J and 2.5J, which is significantly lower than that of LEACH protocol. Although the energy consumption of cluster head nodes in LEACH protocol decreases after the simulation time reaches 380s, it does not actually reduce the total energy consumption. However, with the increase of simulation time, the energy consumption of nodes in the network has approached the initial energy of nodes. Some selected cluster head nodes consume 2J energy in their work and die in the middle. The total energy consumption of cluster head nodes in normal work is calculated experimentally. In addition, the experimental results in Figure 2. show that the total energy consumption of cluster heads with 60 s interval of simulation time increases, because cluster heads need to recalculate the parameters of regression model in each update cycle of regression model and send the updated parameters to the base station. In this way, more computing and communication energy consumption is generated than other simulation time.

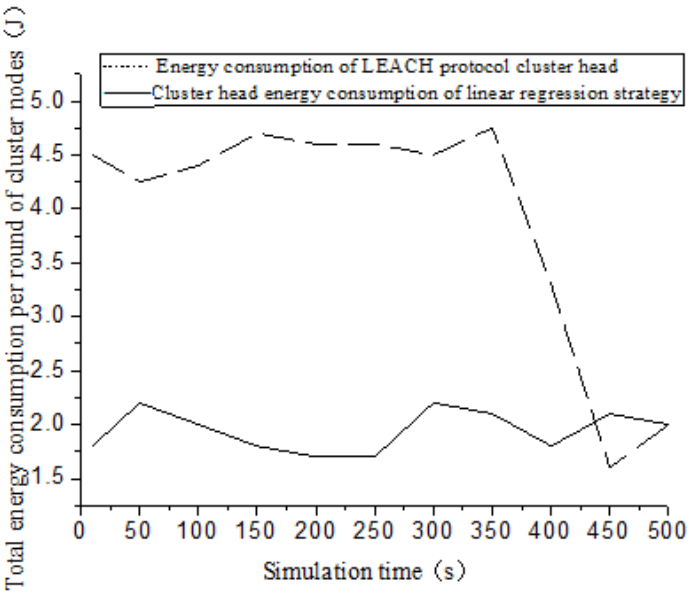

Figure 2. The relation between cluster head energy consumption and time.

In order to simulate the effect of measured changes on the energy consumption of the algorithm in the simulation environment, the updating cycle of the linear regression model is set to 10s, 30s and 60s respectively, and the total energy consumption of each cluster head node is shown in Figure 3. As can be seen from the experimental results, the shorter the regress renewal period is, the higher the mutation frequency is. The more times cluster head nodes calculate and retransmit regression model parameters is, the greater the total energy consumption is. In practical environmental monitoring applications, the frequency of updating and retransmitting the regression model parameters will be very low when the measured values are in linear change in most cases, and the total energy consumption of the network will also be reduced.

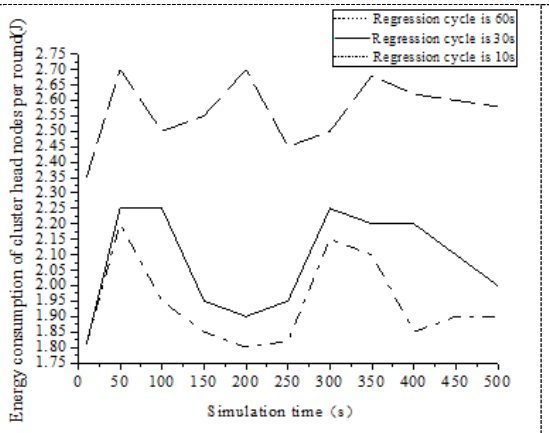

Figure 3. Node energy consumption in different regression periods

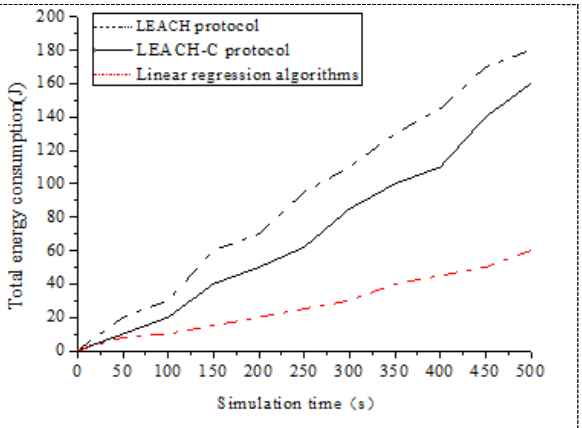

Figure 4. The relationship between total network energy consumption and simulation time

In the simulation time of 500 seconds, when the update period is 60 seconds, the total energy consumption of nodes with linear regression strategy is shown in Figure 4.

As can be seen from Figure 4, compared with the typical LEACH protocol and LEACH-C protocol, the addition of linear regression strategy reduces the overall energy consumption of the network under the same amount of data to be transmitted. After adding the regression model, when the simulation time reaches 500s, the energy consumption of the network is about 100J less than that of LEACH and

LEACH-C protocols. Because the correlation of sampling data in time is considered in the regression strategy when nodes transmit sampling data, a linear regression model is constructed based on their own historical sampling data. Within the closed range of monitoring error, nodes can upload regression model parameters to represent actual sensing data. Although the calculation energy consumption of nodes is increased, the communication energy consumption between nodes is greatly reduced.

5. Conclusion

In this paper, the basic idea and principle of distributed data collection optimization algorithm for wireless sensor networks based on linear regression are described in detail. The incremental updating method of regression model parameters is introduced and the complexity of the algorithm is analyzed. The distributed data acquisition process of the algorithm is illustrated by an example. In the test and analysis of network energy consumption performance, a node is randomly deployed in the monitoring area, and a cluster tree-based wireless sensor network model is constructed. The changes of total energy consumption of cluster-head nodes with simulation time are tested, and the changes of total energy consumption of cluster-head nodes with different regression cycles are set. The experimental results show that the proposed distributed data collection optimization strategy based on linear regression has good performance in prolonging the network life cycle and reducing the total energy consumption of the network, which reflects the feasibility and energy efficiency of the algorithm.

Acknowledgments

Thank you very much for the help and encouragement of my colleagues in this unit, which helped me to complete my paper.

References

[1] Sahingoz, O K . (2013) Larger scale wireless sensor networks with multi-level dynamic key management scheme. J. Journal of Systems Architecture, 59(9):801-807.
[2] Shu, Q., Hu, Q., Zheng, J. (2013) A Dependable Slepian-Wolf Coding Based ClusteringAlgorithm for Data Aggregation in Wireless Sensor Networks. In: International Conference on Wireless Communications & Signal Processing. Hangzhou. pp. 151-156.
[3] Goela, N., Gastpar, M. (2014) Reduced-Dimension Linear Transform Coding of Correlated Signals in Networks. J. Transactions on Signal Processing., 60(6):3174-3187.
[4] Zhang, J., Tang, J., Chen, F. (2016) Energy-Efficient Data Collection Algorithms Based on Clustering for Mobility-Enabled Wireless Sensor Networks. In: International Conference on Cloud Computing & Security. Chengdu. pp. 72-77.
[5] Iwata, M., Tang, S., Obana, S. (2018) Energy-Efficient Data Collection Method for Sensor Networks by Integrating Asymmetric Communication and Wake-Up Radio. J. Sensors, 18(4):1121-1124.
[6] Su, S., Yu, H. (2015) Minimizing tardiness in data aggregation scheduling with due date consideration for single-hop wireless sensor networks. J. Wireless Networks, 21(4):1259-1273.
[7] Lei, L., Wang, H., Lin, C., et al. (2014) Wireless channel model using stochastic high-level Petri nets for cross-layer performance analysis in orthogonal frequency-division multiplexing system. J. Communications Iet, 8(16):2871-2880.
[8] Kolo, J.G., Ang, L.M., Shanmugam, S.A., et al. (2013) A Simple Data Compression Algorithm for Wireless Sensor Networks. J. Advances in Intelligent Systems and Computing, 188:327-336.
[9] Kumar, R., Calders, T. , Gionis, A., et al. (2015) Maintaining Sliding-Window Neighborhood Profiles in Interaction Networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. New York. pp. 72-78.

[10] Ammar, A.B., Dziri, A., Terre, M., et al. (2016) Multi-hop LEACH based cross-layer design for large scale wireless sensor networks. In: Wireless Communications & Mobile Computing Conference. Xi'an. pp. 129-133.
[11] Tripathi, M., Gaur, M.S., Laxmi, V., et al. (2014) Energy efficient LEACH-C protocol for Wireless Sensor Network. In: International Conference on Computational Intelligence & Information Technology. Paris. pp. 335-339.
[12] Zuozhen, L., Bo, L.I., Qiao, Q.U., et al. (2013) Design of frame aggregation simulation platform based on NS2. J. Computer Engineering and Applications, 161:307-311.
[13] Arumugam, G.S., Ponnuchamy, T. (2015)EE-LEACH: development of energy-efficient LEACH Protocol for data gathering in WSN. J. Eurasip Journal on Wireless Communications & Networking, 2015(1):1-9.