Ad Hoc Networks 94 (2019) 101934

Contents lists available at ScienceDirect

Ad Hoc Networks

journal homepage: www.elsevier.com/locate/adhoc

On semantic clustering and adaptive robust regression based energy-aware communication with true outliers detection in WSN

Srijit Chowdhury[a], Ambarish Roy[a], Abderrahim Benslimane[b,*], Chandan Giri[a]

[a] Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, India
[b] CERI/LIA University of Avignon, Avignon, France

ARTICLE INFO

Article history:
Received 29 March 2019
Revised 20 June 2019
Accepted 24 June 2019
Available online 28 June 2019

Keywords:
Energy-aware communication
Semantic clustering
Prediction
Robust regression
Outliers
Wireless sensor networks,

ABSTRACT

To conserve energy and enhance the lifetime of the wireless sensor network (WSN), reducing the amount of data communication by exploiting temporal and spatial correlation of sensed data is well suitable technique. So, instead of sending every data to the destination, it can be worthy of introducing a prediction method to reduce redundant data transmission by exploiting the temporal correlation of sensed data. We show that the prediction accuracy of source data depends not only on the method applied but also on the correctness of the sample data provided by the source nodes. Erroneous sample data (outliers) leads to the wrong prediction. In this paper, we propose an energy efficient SEMantic CLustering (SEMCL) model to mitigate high energy consumption problem in a clustered WSN. Our model produces energy efficient clusters by strong intra-cluster data similarity to exploit spatial correlation of data. We adopt the Robust and Efficient Weighted Least Square method (REWLS) to provide accurate data prediction with negligible errors. Because REWLS method lacks to differentiate true and false outliers and thus to improve further the Quality of Service (QoS) on data accuracy, we propose a separate algorithm, named, True Outlier Detection (TOD). Moreover, to improve the QoS in communications, a reliable backbone network based on the link quality of the data forwarding path has been implemented. Our proposed model has been simulated using real data and compared with the existing techniques to show its efficacy and superiority in terms of QoS on data accuracy, energy consumption, and network lifetime.

1. Introduction and motivation

One of the essential applications of the wireless sensor network (WSN) is to measure various environmental factors affecting the weather conditions. Specific weather conditions can result in partial or extensive failures of a WSN [1,2]. For example, while some environmental factors like temperature and humidity can affect the received signal strength and the link quality of a WSN adversely, a terrible weather condition like natural calamity may ultimately disrupt the working of WSN in a large geographical area. These incidents can hamper the daily routine of society that highly relies on WSNs for its various crucial activities. As the changes in weather conditions are sure to happen and the probability of incoming disaster weather is extremely time-varying, it is essential to measure the environmental factors in a regular short time interval (in seconds or minutes). It is also essential to send these measurements to the network operator or base station (BS) so that proper actions can be taken against the upcoming catastrophic weather before it strikes (for example, disaster alarm can be raised). Sending all the measurements to the BS is undesirable in an energy-constrained WSN as it introduces enormous energy consumption.

In WSN, the data communication between nodes consumes substantial energy compared to the other activities of the sensor nodes like processing or sensing data [3]. To curtail the cost of energy in data communication and to increase the operational network lifetime, reduction of redundant data in communication is required. Some studies [4] show that the weather factors (for example, temperature, light, humidity) have strong temporal correlations. Hence, a prediction method can be used at source nodes to reduce redundant data transmission by exploiting the temporal correlation of sensed data [5]. On the other hand, the nodes in WSN are densely deployed, leading to a high degree of spatial correlation among the data sensed by the neighbouring nodes [6]. The high degree of spatial correlation increases redundant data over the network, which in turn results in the consumption of a significant amount of energy in data communication. Hence, data-aware clustering method based on spatial correlation of data can

* Corresponding author.
E-mail address: abderrahim.benslimane@univ-avignon.fr (A. Benslimane).

https://doi.org/10.1016/j.adhoc.2019.101934
1570-8705/

be adapted to reduce redundant data transmission over the network [7].

From the discussion, it can be argued that the joint consideration of data similarity based clustering and appropriate prediction method can be useful to reduce the number of data communications which in turn can be beneficial for substantial energy savings in WSNs.

Many researchers use well-known Least Mean Square (LMS) filter based prediction algorithm (Refer to Section 2.1) as a suitable prediction method. LMS estimator minimizes mean square error (MSE) and thus produces a very low variance, i.e., the high prediction accuracy for outlier free set of observations. However, it is a well-known fact that in the presence of a single or a fraction of outliers, the estimates by LMS may vary a large from the actual observed values. Therefore, LMS is not considered as a robust estimator. In WSN, observed values of environmental variables may have outliers (inaccurate data), and thus, a robust estimator is required for high prediction accuracy in the presence of outliers. The robustness of an estimator can be measured in terms of breakdown point which measures the maximum fraction of outliers that may have in a given sample without spoiling the estimate completely [8]. A class of robust regression estimators has been proposed which aim to attain the maximum breakdown point and high asymptotic efficiency of estimates simultaneously. In this work, we adopt Robust and Efficient Weighted Least Square (REWLS) estimator [9] due to its outperformance over other existing robust estimators (See Section 4).

Though REWLS is a robust and efficient estimator, it lacks to differentiate between true and false outliers within a temporal dataset. It may result in inefficient prediction accuracy. For example, in the case of environmental parameters like light or wind speed, there may be a sudden change in values due to the rapid changes in the weather condition. For that reason, new observations of the environmental parameters may contain very much deviated values as compared to their previous readings. In this case, REWLS may detect these new values as outliers when compared to the previous set of observations. So, the true outlier detection is essential for high prediction accuracy, and we propose a separate algorithm for this purpose.

1.1. Contributions

In this work, the joint use of distributed semantic clustering and robust regression based data prediction technique has been proposed which yields a high percentage of data reduction (up to 99.5%) in communication and enhances the operational network lifetime. The term 'semantic clustering' refers to the clustering based on data analogy. The main contributions of the proposed work can be summarised as follows:

1. An energy-efficient and distributed SEMantic CLustering (SEMCL) method has been proposed, which forms clusters with spatially correlated nodes. SEMCL adopts the REWLS method for better data accuracy of sensed values as compared to the LMS based approaches [6,10,11].
2. REWLS based data prediction technique is applied to obtain high data reduction in communication and providing a high quality of service (QoS) on data accuracy. As discussed above, the REWLS method lacks to differentiate true and false outliers and thus to improve further the QoS on data accuracy, a separate algorithm, named, True Outlier Detection (*TOD*) has been proposed.

1.2. Paper organization

The remaining parts of the paper are organized as follows: Section 2 describes the related works. Background theories related to the existing methods and the proposed method are discussed in Section 3. Section 4 demonstrates the proposed SEMCL method. The simulation results and performance analysis are described in Section 5. Finally, Section 6 concludes the work.

2. Related works

Some studies [1,2] explore the effects of weather factors (e.g., temperature, humidity, etc.) in WSN and show that disastrous weather may cause severe damages to a WSN. Thus, reading the values of weather factors and forwarding these values to the BS on a regular and short time intervals is imperative, but the implementation of this task is not very straightforward in an energy-constrained network like WSN. To resolve the issue of high energy consumption in data communication and to increase the network lifetime of a WSN, a variety of clustering techniques [12] have been proposed which use various paradigms including computational intelligence(CI) [13], prediction [14] and data similarity [6]. Due to the computational complexity, CI based meta-heuristic methods are comparatively less efficient than prediction and data similarity-based approaches. Some of the existing and relevant prediction and similarity-based redundant data reduction techniques are discussed below as our proposed work considers the combination of both the prediction and the data analogy based approaches to reducing redundant data in communication.

2.1. Prediction based approaches

In [15], three prediction based data aggregation protocols have been proposed where a combination of Grey model and Kalman filter has been used. In this scheme, a double-queue mechanism is used to synchronize the source node and the sink. Prediction combined with Kalman Filter (PKF) [16] combines a predictor with a Kalman filter. Compared to previous works based on Kalman Filter (KF), PKF requires less computational effort while improving the reconstruction quality. Some other prediction based data aggregation protocols [17,18] use autoregression and its variants. These works exploit time series data for better approximation and aggregation.

Prediction-based Data-aware Clustering (PDC) [14] exploits the temporal correlation of data and provides high prediction accuracy and low computation and communication costs. Spatial correlation of data has also been utilized in [14] to form clusters based on the similarity of data. This work does not consider the occurrence of the clustering process in every round. Instead, clusters remain unchanged until the sensor values exceed an error threshold. The main drawback of [14] is that it did not consider the residual energy of nodes in the cluster head (CH) selection. Thus, the energy load may not be evenly distributed over the network.

In some research works, the prediction is used to reduce redundant data in communication, and the source nodes do not need to transmit every sensed data to the destination node. These prediction approaches prevent the transmission of temporally correlated data and thus help in reducing the number of transmissions in the network, which in turn minimizes the energy consumption in communication. Some recent works [10,19] use dual prediction algorithm based on normalized least mean square (nLMS) method. In dual prediction method, the prediction is made at both the source node and the destination node. The prediction algorithm Optimal Step Size Least Mean Square (OSSLMS) has been proposed in [10], which minimizes Mean Square Deviation (MSD) to get optimal step size. Optimal step size is required for the efficient convergence of the Least Mean Square (LMS) filtering process. In another LMS based approach [11], Hierarchical least mean square (HLMS) technique has been used. In this work, two levels of hierarchical LMS

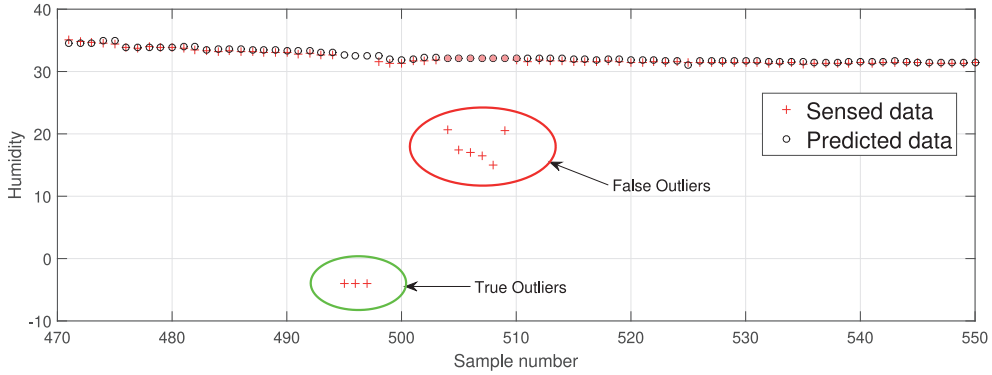


Fig. 1. True outlier detection problem.

filter is applied to improve the convergence speed of the LMS filter when compared with the non-hierarchical LMS techniques.

Regression method can be adapted successfully to predict the values of the parameters having a robust temporal relationship (i.e., values of such parameters change gradually over time) [20]. This statistical method has been applied widely in many real fields for its high prediction accuracy. Authors in [21] use a spatial adaptive estimation of non-parametric regression to detect the fault in autonomous systems where the fault may occur inside or outside of the systems. The work in [22] estimates the upper and lower error bounds of a real-time traffic prediction system. In [23], a regression-based model has been proposed for the prediction of health metrics, which has a significant impact on clinical practice.

2.2. Clustering methods based on data similarity

Data similarity based clustering has been proposed in some research works [6,24]. In [24], a semantic clustering model has been proposed based on the fuzzy system. This work targets the reduction in energy consumption and improvement of data accuracy. The model exploits the spatial correlation of data to find semantic neighborhood relationships. However, this work does not use any predictive scheme and ignores the temporal correlation of data. Distributed Similarity-based Clustering and Compressed Forwarding (DSCCF) [6] constructs isoclusters by data similarity. A dual prediction method is also applied to reduce the intra-cluster data communications. The dual prediction method uses adaptive nLMS algorithm to exploit the temporal correlation of data.

The proposed work is similar to DSCCF with the difference that instead of nLMS based prediction used in DSCCF, adaptive REWLS based prediction method has been proposed for its advantages over LMS based prediction method as discussed in Section 1. Also, a new data similarity based clustering method has been proposed which achieves superior performance as compared with the existing DSCCF clustering method in terms of formation of data similar clusters, clustering overhead and energy dissipation (See Section 5.2).

3. Background theory

3.1. Adaptive LMS method

Some existing prediction algorithms [6,10,11] use the autoregressive model to predict time series data from some previous sample data. Autoregressive parameters are estimated by using adaptive LMS method [6]. The LMS filter accepts a sample data stream $u[t]$ of length F_L at the time instant t and calculates the prediction $v[t]$ as a linear combination of the F_L number of previous samples as

$$v[t] = w^T[t]u[t] \tag{1}$$

where $w[t]$ is the weight vector.

The error $e[t]$ is calculated by comparing the output $v[t]$ with the desired signal $b[t]$ as

$$e[t] = v[t] - b[t] \tag{2}$$

To minimize the mean square error, the filter updates the weight vector at each time instant t as

$$w[t+1] = w[t] + \mu u[t]e[t] \tag{3}$$

where μ is the step size.

After multiple rounds of prediction and weight adaptations using Eqs. (1), (2) and (3), the output signal converges with the desired signal.

3.2. REWLS Method [9]

The robust regression estimator REWLS simultaneously attains full efficiency and maximum breakdown point under errors having a normal distribution. In REWLS, weights are adaptively measured from a base estimator using the empirical distribution of the residuals.

Considering initial estimates of regression and scale respectively $\boldsymbol{\beta}_{0p}$ and σ_p, the standardized residuals are defined as

$$e_k = \frac{y_k - \boldsymbol{x}_k^T \boldsymbol{\beta}_{0p}}{\sigma_p} \tag{4}$$

where σ_p is the standard deviation of p samples.

If $|e_k|$ is large, then $(\boldsymbol{x}_k, y_k)$ is considered as an outlier. The proportion of outliers in the sample is defined as

$$d_p = \max_{k>k_0} \left\{ F^+(|e|_{(k)}) - \frac{(k-1)}{p} \right\}^+ \tag{5}$$

Here $\{.\}^+$ denotes positive part, F^+ is the distribution of $|X|$ where $X \sim F$, $k_0 = max\{k : |e|_{(k)} < \theta\}$, where $|e|_{(1)} \le |e|_{(2)} \le \cdots \le |e|_{(p)}$ are the order statistics of the standardized absolute residuals and θ is some large quantile of F^+. Thus $\lfloor nd_n \rfloor$ observations with largest standardized absolute residuals are eliminated. The adaptive cut-off value becomes

$$t_p = |e|_{(k_p)} \tag{6}$$

where $k_p = p - \lfloor pd_p \rfloor$. It is observed that $k_p > k_0$ and $t_p > \theta$. Using this adaptive cut-off value the adaptive weights are measured as

$$w_k = \begin{cases} 1 & \text{if } |e_k| < t_p \\ 0 & \text{if } |e_k| \ge t_p \end{cases} \tag{7}$$

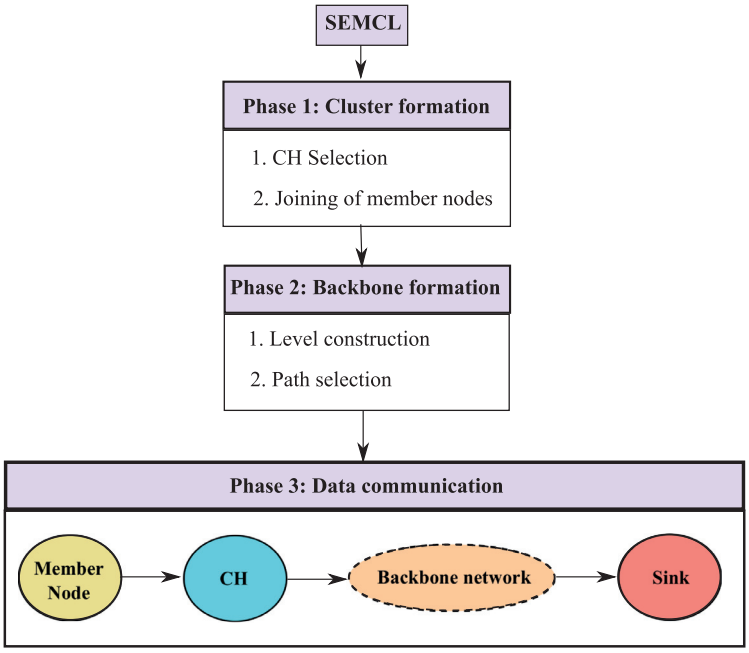


Fig. 2. SEMCL method.

Table 1
Packet description.

Packet Name	Packet Description	Transmission Mode	Content
$Pk_i(1)$	Probable CH announcement by node i	Broadcast	Source Id, Residual Energy, Regression Coefficients
$Pk_{i,j}(2)$	Packet from member node i to the selected CH_j	Unicast	Source Id, Destination Id, Seq No, Regression Coefficients
$Pk_i(3)$	Packet from CH_i for backbone formation	Broadcast	Source Id, Residual Energy, Backbone level
$Ack_{i,j}$	Packet from CH_i to CH_j	Unicast	Source ID, Destination Id, Ack Info
$Pk_{i,j}(4)$	Aggregated data Packet from CH_i to next hop node j	Unicast	Source Id, Destination Id, Aggregated Data

and the REWLS estimate is

$$\boldsymbol{\beta}_{1p} = \begin{cases} (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} & \text{if } \sigma_p > 0 \\ \boldsymbol{\beta}_{0p} & \text{if } \sigma_p = 0 \end{cases} \tag{8}$$

where $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_p})^T$, $\boldsymbol{W} = diag(w_1, \ldots, w_p)$ and $\boldsymbol{y} = (y_1, \ldots, y_p)^T$.

For the base estimates ($\boldsymbol{\beta}_{0p}$) in REWLS estimator, we use LMS estimator in our proposed prediction based communication method so that the proposed method performs better than the LMS based prediction methods especially in the presence of outliers in the sensed data.

The problem of true outlier detection using REWLS is described in the next section.

3.2.1. *Problem of REWLS estimator in detecting true outliers*

Identifying the causes and sources of outliers is essential to decide on the rejection of the detected outliers [25]. For example, if the detected outlier is a noisy data (error), it should be discarded from the sensed data and should not send to the sink to ensure high quality and accuracy of data. On the other hand, if the outlier is caused by an event (e.g., fire or chemical spills), elimination of the outlier will lead to the loss of crucial hidden information about the event. REWLS estimator is unable to distinguish the noisy data or error (we call it as the true outlier) and the outlier caused by an event (we call it as the false outlier). REWLS estimator treats all the observations which differ a large from the other similar observations in sequence as the true outlier, and thus it discards all these outliers from the future trend of the sensed data sequence (predicted data in Fig. 1). In Fig. 1, the false outliers (red circled observations) are the real humidity fluctuations due to some events in nature, and thus, these observations should not be considered as errors. However, REWLS detects those points as outliers (errors). The spatial correlation among observations of the neighboring nodes can help to distinguish between events and errors. Event measurements (true observations) are likely to be spatially correlated, while noisy measurements (erroneous observations) and sensor faults are likely to be stochastically unrelated [26]. Based on this insight, to overcome the inability of the REWLS method to detect proper outlier, a true outlier detection method named *TOD* has also been proposed (See Section 4.3.4).

4. Proposed method

4.1. Overview

The proposed SEMCL method consists of three phases. In the first phase, energy-efficient clusters are formed from the spatial correlation of data among the neighbour nodes. The robust regression method REWLS is used to find the spatially correlated neighbour nodes. Each non-CH node is associated with the CH node, which possesses high residual energy and shows high data similarity with the non-CH node. After the formation of clusters, in the second phase, a reliable backbone network structure is formed among all the CH nodes to communicate with the static sink. The backbone helps to choose the most energy-efficient path with

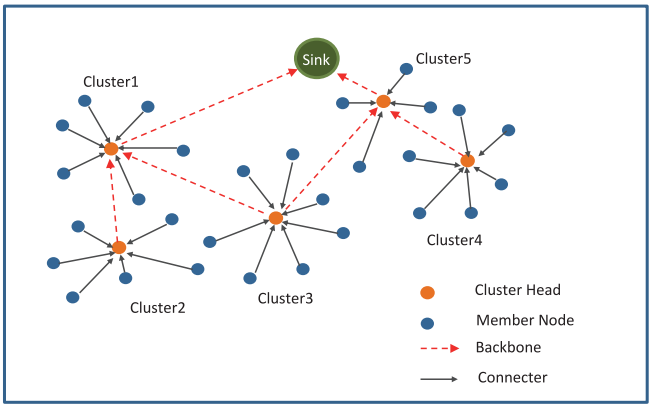


Fig. 3. Distributed clustered network with backbone.

strong link quality between a CH and the sink. In the third phase, intra-cluster and inter-cluster communication methods have been proposed for data transmission. REWLS based prediction method is employed to reduce intra-cluster data communication. Instead of receiving every sensed data from the cluster member nodes, each CH predicts sensed data of its associated member nodes. Then, each CH performs data aggregation by averaging all predicted data and sends the aggregated data to the sink through the inter-cluster backbone path. Fig. 2 describes the phases of SEMCL method. Before describing the phases of SEMCL method, the energy model is depicted.

4.2. Energy model

We use the reference of working configuration sheet of MicaZ 2.4GHz IEEE 802.15.4 motes [27] for the estimation of the expended energy in the WSN.

Energy Consumption for transmitting k bytes of data is obtained as

$$E_{tx}(k) = P_{tx} \times T_{tx}(k)$$

where $P_{tx} = Volt \times Ampere$ is the expended energy (in Joule/sec) during packet transmission and T_{tx} is the transmission duration (in Seconds).

Energy Consumption for receiving k bytes of data is obtained as

$$E_{rx}(k) = P_{rx} \times T_{rx}(k)$$

where $P_{rx} = Volt \times Ampere$ is the expended energy (in Joule/sec) during packet receiving and T_{rx} is the duration of receiving k bytes of data.

If the data transmission rate is $R\,kbps$, then the time duration of transmitting or receiving k bytes of data can be calculated as $(\frac{k/1000}{R})$ seconds.

In the case of the clustering process, backbone formation, and data communication, various packets are transmitted over the network. A brief description of each packet is described in Table 1.

4.3. SEMCL Method

The cluster head selection procedure of the proposed SEMCL method is described below:

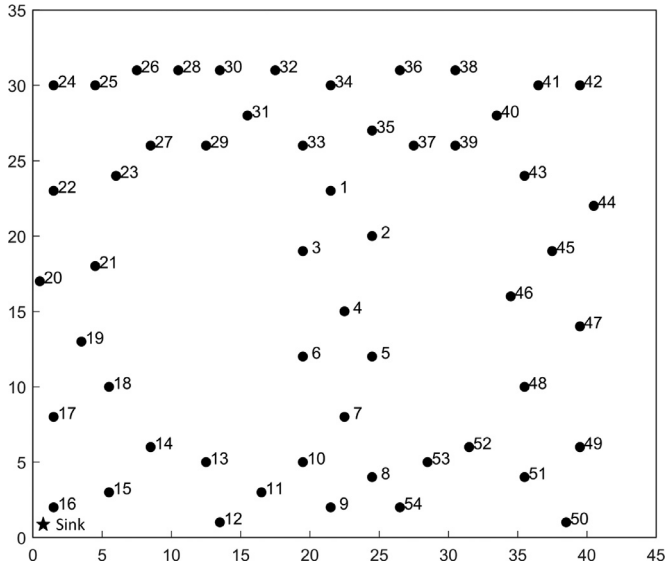


Fig. 4. Sensor node deployment in the Intel Lab.

4.3.1. Cluster head selection

Initially, all the nodes have non-uniform residual energy. All the nodes sense F_L number of data from the environment and find regression coefficients. The steps are as follows:

(i) Higher energy nodes announce themselves as CH after waiting a specific time based on residual energy of the nodes. This specific time is measured as the ratio of a time constant (TC), and *Residual Energy* of the announcing node say j. TC is the maximum transmission time of the announcement message $Pk_j(1)$ between two nodes.

(ii) If a node i receives $Pk_j(1)$ before it announces, then it stops announcing. Non-CH node i stores all received $Pk(1)$ packets in $CH_i^{Probable}$ list.

(iii) If a non-CH node i receives a single announcement packet $Pk_j(1)$, then it selects j as its final CH.

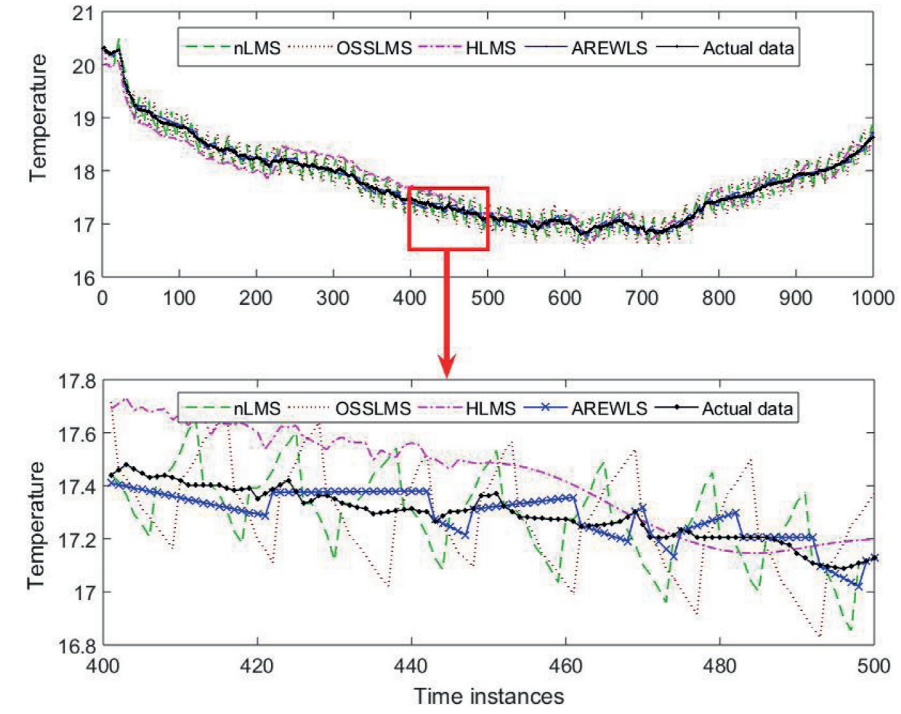
(iv) If a non-CH node receives multiple packets, then it performs two steps elimination to select its final CH. In the first step, the data dissimilarity is checked. Dissimilarity value ($DS_{i,j}$) is the Euclidean distance of regression coefficients between the announcing CH_j and the node i. If $\boldsymbol{\beta}_i = (w_{i0}, w_{i1}, w_{i2}, \ldots, w_{ip})$ and $\boldsymbol{\beta}_j = (w_{j0}, w_{j1}, w_{j2}, \ldots, w_{jp})$ (p is the number of sensed data) are the regression coefficient vectors of the nodes CH_j and i respectively, then the dissimilarity value between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ is calculated as
$DS_{i,j} = |(w_{i0} - w_{j0}) + (w_{i1} - w_{j1}) + (w_{i2} - w_{j2}) + \cdots + (w_{ip} - w_{jp})|$
If the dissimilarity value exceeds the predefined dissimilarity threshold (Dis_Thes), then CH_j is discarded from the $CH_i^{Probable}$ list. In the second step, determination factor ($DF_{i,j}$) between CH_j and node i is calculated as

$$DF_{i,j} = (Residual\ Energy\ of\ CH_j / Residual\ Energy\ of\ node\ i) \times (Dis\_Thes/DS_{i,j}) \quad (9)$$

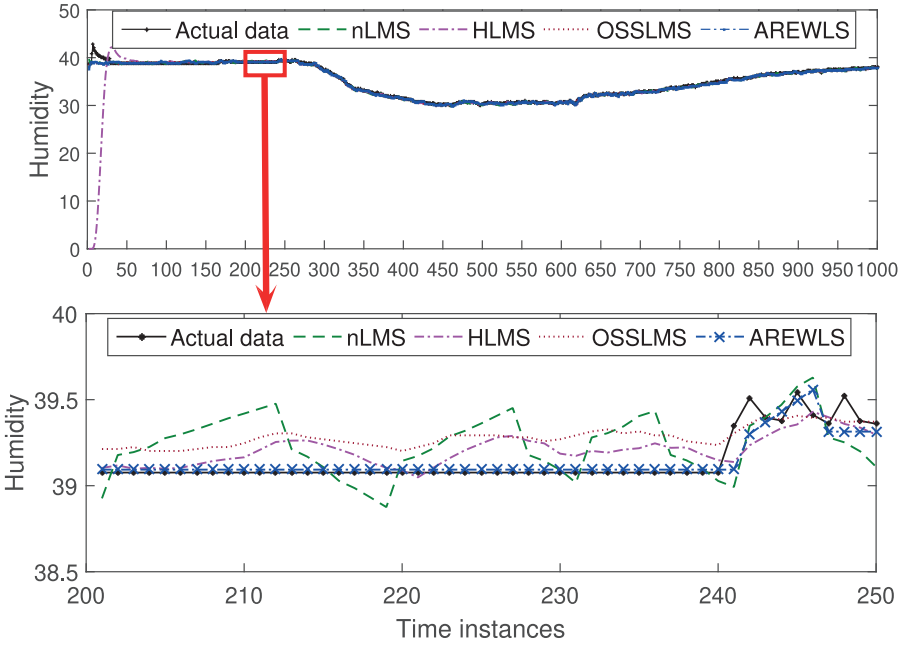
(v) Node i selects the CH that has highest DF value among all the probable CHs.

(vi) Node i joins the CH_j by sending the packet $Pk_{i,j}(2)$ containing node id and regression coefficients to CH_j.

(vii) After receiving a packet $Pk(2)$, a CH stores the node-id and regression coefficients of the associated member node.



(a)



(b)

Fig. 5. Predicted data vs actual data (a) Temperature (b) Humidity.

After the final selection of the CHs, the backbone network is formed among all the selected CHs. The procedure of the backbone formation is described next.

4.3.2. Backbone formation

For inter-cluster communication, a backbone structure is formed among the selected CHs according to the quality of the link between two CHs. In this work, the quality of link ($linkQ_{i,j}$) between two CHs i and j is measured on the basis of the link quality indicator ($LQI_{i,j}$) received by the destination node j for the link {i,j} and the residual energies of node i and node j. The value of $LQI_{i,j}$ is represented by a normalized parameter which varies with the inverse of the square of the distance between CHs i and j. The quality of the link ($linkQ_{i,j}$) between two nodes i and j is measured as

$$linkQ_{i,j} = \frac{Residual\ energy\ of\ i}{Residual\ energy\ of\ j} \times LQI_{i,j} \tag{10}$$

Each CH sets its backbone level to a very large number (∞). Backbone level of the sink is assumed to be zero. The sink starts the

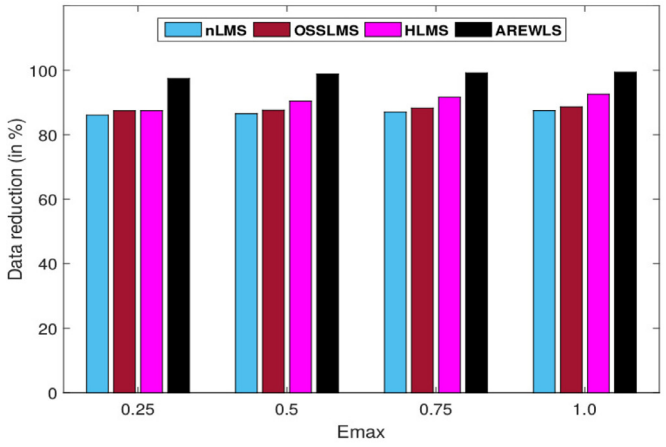


Fig. 6. (a) Data reduction percentage varying with *Emax* (b) RMSE varying with *Emax*.

formation of the backbone by broadcasting the $Pk(3)$ packet. If a CH node j with level $level_j$ receives a single packet $Pk_i(3)$ from the node i with level $level_i$ where $level_i < level_j$, then the CH node j stores the value $LQI_{i,j}$. Also, the CH node j sets node i as a predecessor (*Next_Hop*) and sets its new level by incrementing the level of the node i by 1. Then the CH node j broadcasts the packet $Pk_j(3)$ containing its new level and residual energy. If CH node j with level $level_j$ receives multiple $Pk(3)$ packets with the same level ($level_n$) where $level_n < level_j$, then the node j calculates *linkQ* value for each source CH node and selects the CH node with higher *linkQ* value as its *Next_Hop*.

Fig. 3 describes the multi-hop relay based backbone formation among CHs. Data aggregation is done at a CH by averaging all predicted data of its member nodes and its sensed data. Then the CH sends aggregated data to the sink directly if the sink is within the range of the CH (see Cluster 1 and Cluster 5 in Fig. 3). Otherwise, the CH sends aggregated data through other CH in the backbone (see Cluster 2, Cluster 3, and Cluster 4 in Fig. 3). In Fig. 3, Cluster 2 and Cluster 4 forward their aggregated data packet through Cluster 1 and cluster 5, respectively, as no other alternative path is available to Cluster 2 and Cluster 4. Cluster 3 has two available paths to forward its data (through Cluster 1 or Cluster 5). Cluster 3 selects the path which possesses a higher *linkQ* value at the time of data forwarding.

4.3.3. Communication protocol between CH and its member nodes

In the proposed scheme, each non-CH node i calculates the regression coefficients ($\boldsymbol{\beta}_i$) from the sample data using the REWLS method. Then each non-CH node sends the regression coefficients to the CH node. The CH node starts prediction for each associated non-CH node from the received regression coefficients.

The proposed intra cluster communication protocol is described in Algorithm 1 as follows.

Algorithm 1: Communication Protocol.

Input : N //Set of member nodes
$ListCH$ //Set of cluster heads
F_L //Number of sensed data for regression coefficient determination
$n = F_L$ //Time sequence in terms of rounds
Set $Emax$ //Prediction Error Threshold
Output: Aggregated data from each CH to the Next_Hop node

1 **At each member node** $i \in N$**:**
2 **begin**
3 Initialise F_L number of sensed data ($\boldsymbol{x}_i$ = [x(n),x(n-1),x(n-2),... x(n-F_L+1)])
4 **Regression Coefficient Determination State:**
5 Calculate regression coefficients $\boldsymbol{\beta}_i$ using REWLS method
6 Transmit packet $Pk_{i,c}(2)$ containing ($\boldsymbol{\beta}_i$, $\boldsymbol{x}_i$) to its cluster head c
7 **Prediction State:**
8 **for do**
9 n=n+1
10 Sense data x(n) at time t(n)
11 $y(n) = \boldsymbol{x}_i^T \times \boldsymbol{\beta}_i$
12 e(n)=x(n)-y(n)
13 **if** $|e(n)| \leq Emax$ **then**
14 $\boldsymbol{x}_i$ = [y(n),x(n-1),x(n-2),... x(n-F_L+1)]
15 n=n+1
16 **end**
17 **else**
18 $\boldsymbol{x}_i$ = [x(n),x(n-1),x(n-2),... x(n-F_L+1)]
19 Go to **Regression Coefficient Determination State**
20 **end**
21 **end**
22 **end**
23 **At each cluster head** $c \in ListCH$**:**
24 **begin**
25 **Initialization State:**
26 Initialise $\hat{\boldsymbol{\beta}}_c$ with the received values ($\boldsymbol{\beta}_i, \boldsymbol{x}_i$) from each member node i during cluster formation
27 **Prediction State:**
28 **for do**
29 n=n+1
30 **if** *any* $Pk_{i,c}(2)$ *packet received where* $i \in N$ **then**
31 Update $\hat{\boldsymbol{\beta}}_c(i)$ with received ($\boldsymbol{\beta}_i, \boldsymbol{x}_i$)
32 //Predict values $y(i), \forall i \in N$
33 $y(i) = \boldsymbol{x}_i^T \times \boldsymbol{\beta}_i$
34 //Check for true outliers
35 $TOD(y, N)$ //*Algorithm2*
36 **if** $y(i)$ *is true outlier where* $i \in N$ **then**
37 Discard new $\boldsymbol{\beta}_i$
38 **end**
39 **end**
40 //Predict values for each node $iy_i(n) = \boldsymbol{x}_i^T \times \boldsymbol{\beta}_i$
41 n=n+1
42 Aggregate predicted data $[y_i(n)], \forall i \in N$
43 Transmit packet $Pk_{c,j}(4)$ to *Next_Hop* node
44 **end**
45 **end**

Lines 3 to 6 of Algorithm 1 initialize F_L number of sensed data $\boldsymbol{x}_i$ for each non-CH node i. Then each node i generates regression coefficients ($\boldsymbol{\beta}_i$) by using the REWLS method and sends those coefficients and data set ($\boldsymbol{\beta}_i$, $\boldsymbol{x}_i$) to the CH node. Then the non-CH node i starts prediction. Line numbers 7 to 21 describes the prediction state of node i. At a particular time instant $t(n)$, the node i predicts the value $y(n)$ and also senses the value $x(n)$. Then it verifies if the error in prediction $e(n)$ exceeds the maximum tolerable error limit $Emax$. If $|e(n)|$ exceeds $Emax$ then the node i runs the

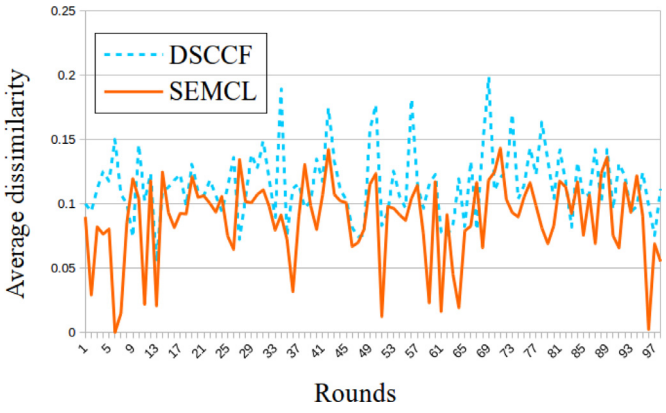


Fig. 7. Average dissimilarity.

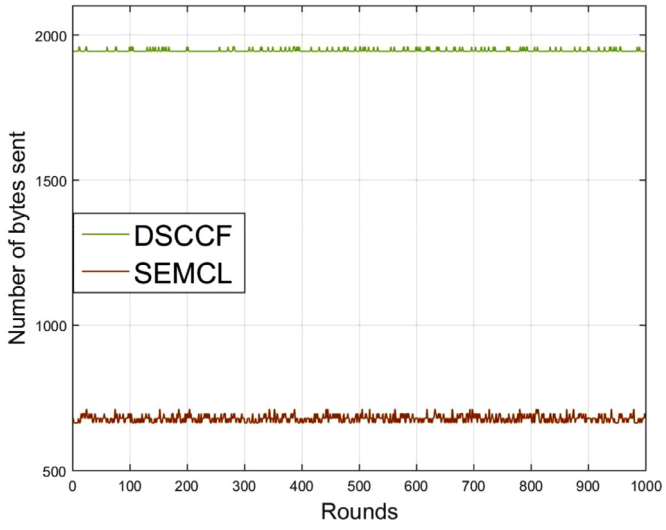


Fig. 8. Clustering overhead (in bytes).

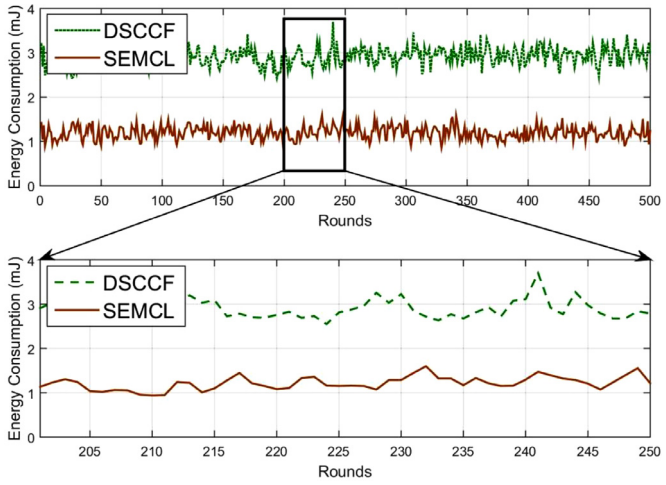


Fig. 9. Energy consumption in clustering for whole network.

REWLS method to adjust the regression coefficients with the help of the stored last F_L number of data ($\boldsymbol{x}_i$) including the last actual reading $x(n)$. In this way, the REWLS method is adaptively used to adjust the regression coefficients dynamically for better prediction.

On the other hand, at each time instant $t(n)$ before predicting values, the CH node waits for a certain amount of time to accept new regression coefficients, if any (line 30). If the CH node receives new coefficients and data set ($\boldsymbol{\beta}_i$, $\boldsymbol{x}_i$) from any non-CH node i, it updates the corresponding coefficients temporally. Now, the CH checks whether the changes in coefficients are due to real changes in sensed data caused by events. Because REWLS method cannot detect the true outliers as discussed in Section 3.2.1. For this purpose, the CH predicts all values $y(i)$ using all changed regression coefficients (Line 33). Then it runs the *TOD* method (Algorithm 2) to detect true outliers. If the coefficients are changed due to the real changes in the environment, then all the spatially correlated values $y(i)$ should not possess any outlier. If *TOD* detects any true outlier, then the corresponding $\boldsymbol{\beta}_i$ is discarded from the updated list of coefficients, and the corresponding previous regression coefficients are retained (See lines 34 to 37).

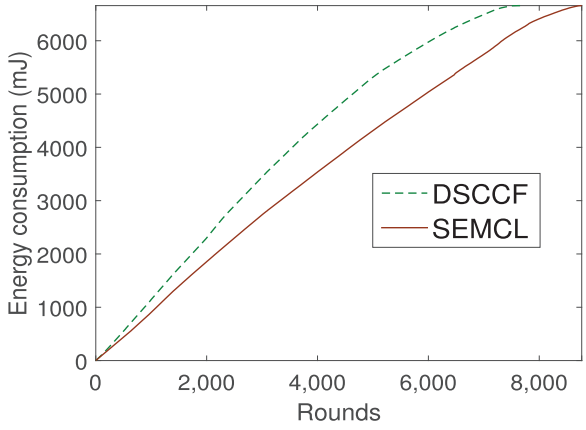


Fig. 10. Network energy consumption per round.

Algorithm 2: True Outlier Detection (TOD).

```
function TOD(Y, N)
Y: Set of values,
N: Number of items in Y,

Y_1 = sort_ascending(Y).
Calculate median.
Calculate the 1st quartile (Q1).
Calculate the 3rd quartile (Q3).
Calculate interquartile range (α)=Q3-Q1.
Calculate l_1 = Q3 + α × 3.
Calculate l_2 = Q1 − α × 3.

for i=1 to N do
    if Y[i] > l_1 or Y[i] < l_2 then
        Y[i] is true_outlier.
    end
end
end function
```

Now, each CH node predicts data for its all member nodes, aggregates all predicted data by averaging and sends to the next hop (*Next_Hop*) CH node. The *Next_Hop* CH node is determined at the backbone formation phase, as explained in Section 4.3.2.

4.3.4. *True outlier detection:*

For true outlier detection, we use interquartile range statistics [28]. The interquartile range of a dataset describes the spread of the middle fifty percent of the distribution. The interquartile range is useful to describe data sets where a few extreme values (outliers) may exist because the interquartile range is not sensitive to extreme values of the dataset. The true outlier detection method is described in Algorithm 2. The *TOD* algorithm first finds the median of the sorted values of the given dataset and then with the help of the median, it calculates the first and third quartiles ($Q1$ and $Q3$) [28]. Then the interquartile range (α) is calculated.

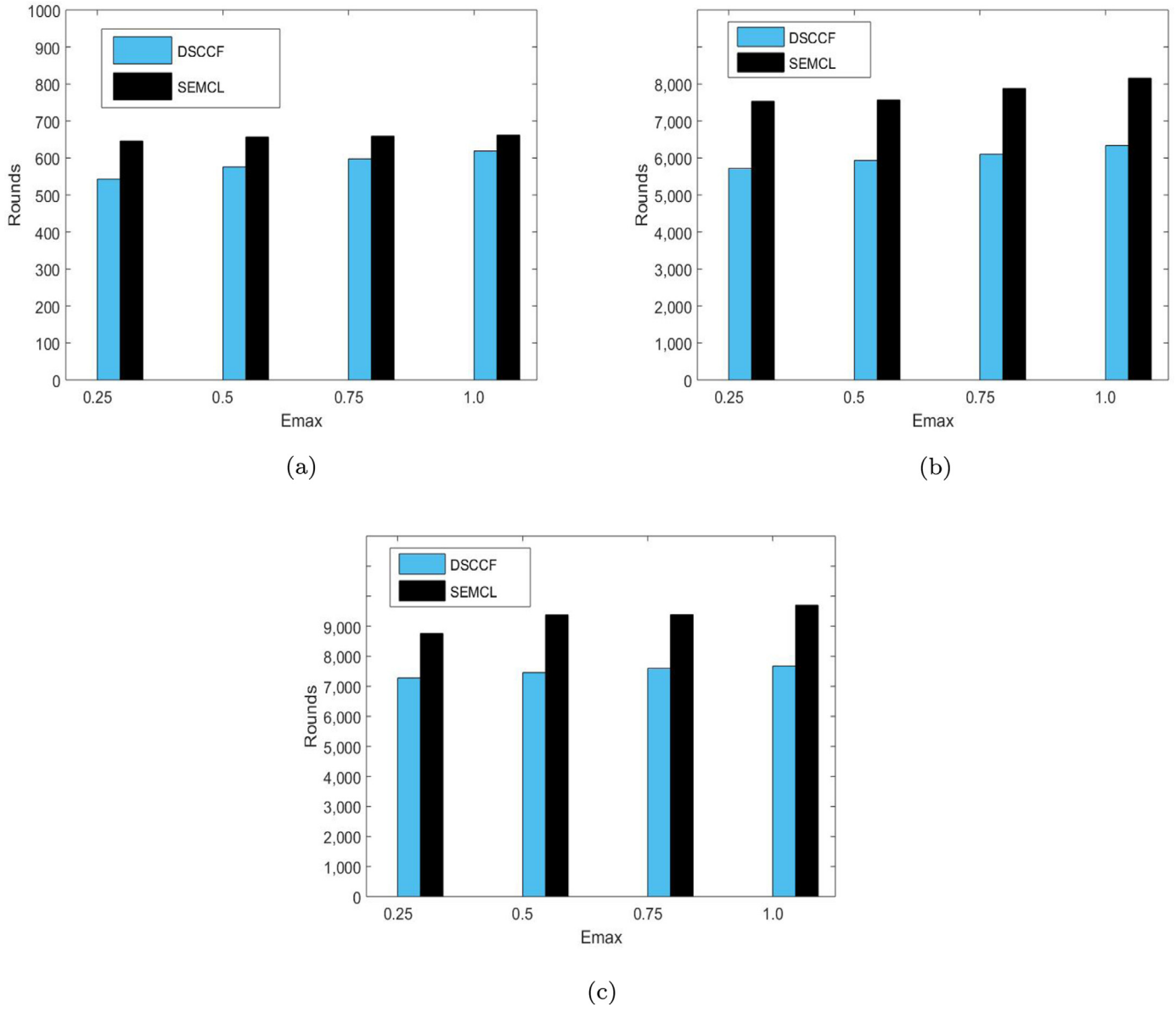


Fig. 11. Node death varying with *Emax* (a) First Node Death (b) Half Node Death (c) Last Node Death.

Table 2
Simulation parameters.

Parameters	Values
Number of nodes	54
Simulation area	45m × 45m
Transmission range	20m-30m
Filter length (F_L)	4
Emax	0.25, 0.5, 0.75, 1.0
Initial energy of a node (*IE*)	0.125J, 0.25J, 0.5J, 0.75J, 1J
Current draw in receive mode [27]	19.7*mA*
Current draw in transmit mode [27]	11*mA*
Battery power (in Volt) [27]	3*V*
Data transmission rate [27]	250*kbps*
Data packet size [6]	4000 bits
Control packet size [6]	100 bits

To detect outliers, *TOD* calculates two thresholds I_1 and I_2 respectively over the third quartile and below the first quartile. If any value ($\boldsymbol{Y}[i]$) of the dataset lies outside of these thresholds, then it is treated as an outlier.

5. Simulation results and performance analysis

The software simulation of the proposed scheme has been performed in MATLAB [29]. We evaluate and analyze the performance of our proposed scheme for data reduction and data accuracy in transmission, data similarity in semantic clustering, overhead, energy consumption and network lifetime and compare the results with the existing works DSCCF [6], OSSLMS [10] and HLMS [11]. For simulations of the algorithms, a network which consists of spatially correlated sensor nodes is considered. For this purpose, we use publicly available real data sheet provided by the Intel Berkeley Lab [30]. Fig. 4 shows the deployment of the sensor nodes in the Intel Lab, where the sink is located at [0.5, 0.5]. Energy consumption of 600μJ is considered for prediction [6]. The simulation parameters are shown in Table 2.

5.1. Performance evaluation of REWLS based prediction

We discuss the performance of the proposed REWLS based prediction method (In figures, we refer it as *AREWLS*) and compare

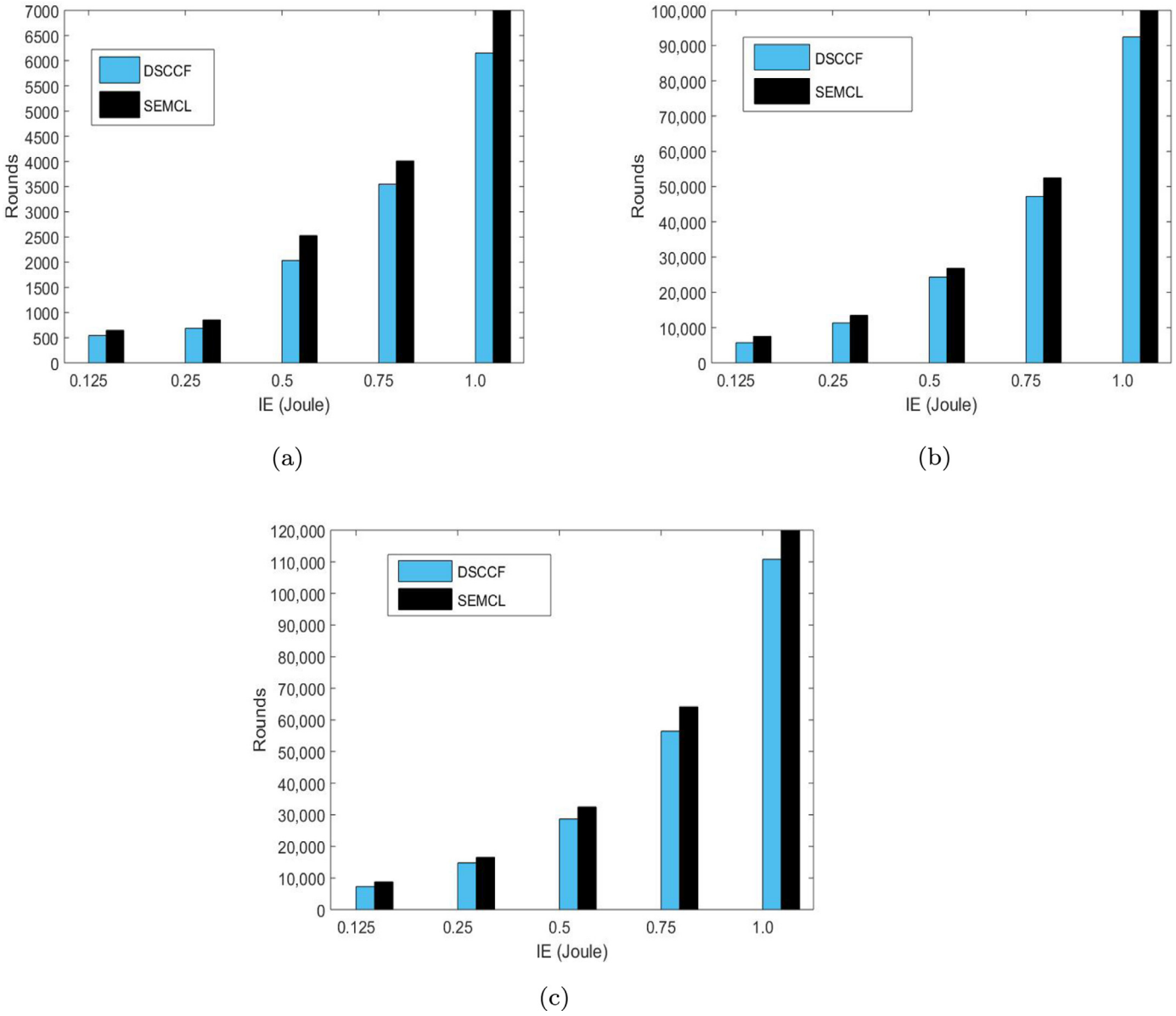


Fig. 12. Node death varying with Initial energy(*IE*) (a) First Node Death (b) Half Node Death (c) Last Node Death.

it with three existing LMS filter based prediction methods namely nLMS [6], OSSLMS [10] and HLMS [11]. Randomly we have selected the sensor node 20 for the evaluation. The prediction error threshold (*Emax*) is set to 0.25. Fig. 5(a) describes the closeness between the sensed data (real data) and the predicted data for 1000 sample data. The closeness is measured by the difference between the sensed data and predicted data. It can be seen in the figure that the predicted data of the REWLS based prediction method are very close to the real data when compared with other existing prediction methods. It leads to good prediction accuracy and high data reduction in communication, which is verified in Fig. 6.

Fig. 6 (a) describes the percentage of data reduction of the whole network with varying *Emax*. We consider 2000 samples for each node in the network. The values set for *Emax* are 0.25, 0.5, 0.75 and 1. From the figure, it is observed that the REWLS based prediction method yields high data reduction (97.4% to 99.5%) while other existing algorithms manage 86.1% to 92.5% data reduction. Thus the REWLS based prediction method has achieved 7% to 11.3% improvement in data reduction compared with the existing algorithms.

Now, we evaluate the data prediction accuracy of all algorithms by measuring the Root Mean Square Error (RMSE) [31]. RMSE measures the average deviation of the predicted values from the actual sensed data and is calculated as given in equation (11).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e(n)^2}{t}} \tag{11}$$

where $e(n)$ is the error in prediction at the nth time instance. It is seen from Fig. 6(b) that REWLS based prediction method yields low RMSE value in all the cases as compared with the existing algorithms. Thus REWLS based prediction method produces minimum average deviation when compared with nLMS, OSSLMS, and HLMS.

5.2. Performance evaluation of SEMCL

We evaluate the performance of the proposed clustering algorithm SEMCL and compare the results with an existing clustering algorithm DSCCF [6]. We do not consider OSSLMS [10], and HLMS [11] for this performance evaluation as those schemes have not proposed any clustering methods. We measure the clustering efficiency in terms of the average dissimilarity between a non-CH node and its associated CH node, clustering overhead, and energy dissipation in the clustering process. Fig. 7 shows average dissimilarity values per round where *Emax* is set to 0.25, and *Dis_Thes* is set to 1. Average dissimilarity value in a round is calculated by $\frac{\sum_{i=1}^{q} DS_{i,c}}{q}$, where $DS_{i,c}$ is the dissimilarity value between a non-CH node i and its corresponding CH node c, and q is the number of non-CH nodes. Low dissimilarity value implies a better similarity between the non-CH node and its CH. We observe that SEMCL possesses lower average dissimilarity than that of DSCCF.

The energy efficiency of a clustering process requires minimum overhead transmissions. In Fig. 8, we compare the overhead transmissions (in bytes) in clustering between SEMCL and DSCCF methods. From the figure, we see that SEMCL has significantly less overhead (in bytes) than that of DSCCF. Lower overhead cost invariably leads to lower energy consumption, as shown in Fig. 9. The energy consumption of the whole network per round during the cluster formation and the backbone construction is presented in Fig. 9. One can observe from the simulated results in both the figures that the energy consumption of SEMCL is approximately half than that of DSCCF. This less energy dissipation helps in extending the network lifetime as is depicted in Figs. 11 and 12.

From the simulation results, it is seen that the proposed SEMCL produces better results than DSCCF in terms of intra-cluster data similarity as well as energy consumption during cluster formation. In the next section, we evaluate the performance of the overall system.

5.3. Performance evaluation of the overall system

We evaluate the overall system performance of our proposed algorithm in terms of scalability, energy consumption, and network lifetime. Fig. 10 shows the cumulative network energy consumption per round. In the experiment, we set *IE* to 0.125J and *Dis_Thes* to 2. From the figure, it can be seen that the SEMCL achieves per round energy dissipation lower than that of the DSCCF.

Figs. 11 and 12 compare the proposed method SEMCL with DSCCF for the network lifetime in terms of first node death, half node death and last node death respectively. To prove the scalability of the proposed model, we evaluate the network lifetime varying with *IE* and *Emax* in Figs. 11 and 12 respectively. In all the cases of first, half and last node death, SEMCL runs for more number of rounds than DSCCF. We see that the number of rounds increases with the increase in *Emax* (Fig. 11); because as the upper bound of the error threshold increases, the algorithm has the scope to make a correct prediction within a wider error range. Thus, the number of successful predictions is increased, which in effect reduces the number of data communications of individual nodes. It is also seen that in the different scenarios with varying *IE* (Fig. 12), SEMCL outperforms DSCCF in terms of network lifetime. The overall percentage of improved network lifetime of SEMCL over DSCCF is 4% to 36%.

6. Conclusion

This work proposes a semantic clustering method (SEMCL) which exploits temporal and spatial correlations of data to form an efficient data collection framework in a sensor network. A communication protocol is also proposed for intra-cluster correspondence, which utilizes the REWLS based prediction method. Experimental results show that the REWLS based prediction method achieves high accuracy on data prediction (up to 99.5%); this greatly helps in reduction of data in intra-cluster data communication which in turn saves network energy to a great extent. This work also depicts an efficient method of inter-cluster communications by forming an LQI based backbone structure. The proposed work is compared with the existing works in various network settings. Through rigorous simulation, it is shown that the proposed work showcases better results in terms of QoS on data accuracy and reliability, data reduction, energy consumption, and network lifetime.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Luomala, I. Hakala, Effects of temperature and humidity on radio signal strength in outdoor wireless sensor networks, in: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), 2015, pp. 1247–1255, doi:10.15439/2015F241.

[2] M. Tornatore, J. André, P. Babarczi, T. Braun, E. Følstad, P. Heegaard, A. Hmaity, M. Furdek, L. Jorge, W. Kmiecik, C.M. Machuca, L. Martins, C. Medeiros, F. Musumeci, A. Pašić, J. Rak, S. Simpson, R. Travanca, A. Voyiatzis, A survey on network resiliency methodologies against weather-based disruptions, in: 2016 8th International Workshop on Resilient Networks Design and Modeling (RNDM), 2016, pp. 23–34, doi:10.1109/RNDM.2016.7608264.

[3] I. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, Wireless sensor networks: a survey, Comput. Netw. 38 (4) (2002) 393–422, doi:10.1016/S1389-1286(01)00302-4.

[4] F. Wang, S. Wu, K. Wang, X. Hu, Energy-efficient clustering using correlation and random update based on data change rate for wireless sensor networks, IEEE Sensors J. 16 (13) (2016) 5471–5480, doi:10.1109/JSEN.2016.2561283.

[5] A. Sinha, D. Lobiyal, Prediction models for energy efficient data aggregation in wireless sensor network, Wirel. Pers. Commun. 84 (2) (2015) 1325–1343.

[6] M. Arunraja, V. Malathi, E. Sakthivel, Distributed similarity based clustering and compressed forwarding for wireless sensor networks, ISA Trans. 59 (2015) 180–192.

[7] A. Jindal, K. Psounis, Modeling spatially correlated data in sensor networks, ACM Trans. Sensor Netw. (TOSN) 2 (4) (2006) 466–499.

[8] V.J. Yohai, High breakdown-point and high efficiency robust estimates for regression, Ann. Statist. 15 (2) (1987) 642–656, doi:10.1214/aos/1176350366.

[9] D. Gervini, V.J. Yohai, A class of robust and fully efficient regression estimators, Ann. Statist. 30 (2) (2002) 583–616, doi:10.1214/aos/1021379866.

[10] M. Wu, L. Tan, N. Xiong, Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications, Inf. Sci. 329 (2016) 800–818.

[11] L. Tan, M. Wu, Data reduction in wireless sensor networks: a hierarchical LMS prediction approach, IEEE Sensors J. 16 (6) (2016) 1708–1715.

[12] M.M. Afsar, M.-H. Tayarani-N, Clustering in sensor networks: a literature survey, J. Netw. Comput. Appl. 46 (2014) 198–226, doi:10.1016/j.jnca.2014.09.005.

[13] P.S. Mann, S. Singh, Energy efficient clustering protocol based on improved metaheuristic in wireless sensor networks, J. Netw. Comput. Appl. 83 (2017) 40–52, doi:10.1016/j.jnca.2017.01.031.

[14] M. Ashouri, H. Yousefi, J. Basiri, A.M.A. Hemmatyar, A. Movaghar, PDC: prediction-based data-aware clustering in wireless sensor networks, J. Parallel Distr. Comput. 81 (2015) 24–35.

[15] G. Wei, Y. Ling, B. Guo, B. Xiao, A.V. Vasilakos, Prediction-based data aggregation in wireless sensor networks: combining grey model and kalman filter, Comput. Commun. 34 (6) (2011) 793–802.

[16] Y. Huang, W. Yu, C. Osewold, A. Garcia-Ortiz, Analysis of pkf: a communication cost reduction scheme for wireless sensor networks, IEEE Trans. Wireless Commun. 15 (2) (2016) 843–856.

[17] S. Ozdemir, Y. Xiao, Polynomial regression based secure data aggregation for wireless sensor networks, in: Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE, IEEE, 2011, pp. 1–5.

[18] C. Carvalho, D.G. Gomes, N. Agoulmine, J.N. de Souza, Improving prediction accuracy for wsn data reduction by applying multivariate spatio-temporal correlation, Sensors 11 (11) (2011) 10010–10037.

[19] H. Jiang, S. Jin, C. Wang, Prediction or not? an energy-efficient framework for clustering-based data collection in wireless sensor networks, IEEE Trans. Parallel Distrib. Syst. 22 (6) (2011) 1064–1071.

[20] H. Firouzi, A.O. Hero, B. Rajaratnam, Two-stage sampling, prediction and adaptive regression via correlation screening, IEEE Trans. Inf. Theory 63 (1) (2017) 698–714, doi:10.1109/TIT.2016.2621111.

[21] A. Kalmuk, O. Granichin, O. Granichina, M. Ding, A dynamic threshold based algorithm for change detection in autonomous systems, IFAC-PapersOnLine 49 (13) (2016) 141–145, doi:10.1016/j.ifacol.2016.07.941. 12th IFAC Workshop on Adaptation and Learning in Control and Signal Processing ALCOSP 2016
[22] F.C. Pereira, C. Antoniou, J.A. Fargas, M. Ben-Akiva, A metamodel for estimating error bounds in real-time traffic prediction systems, IEEE Trans. Intell. Transp. Syst. 15 (3) (2014) 1310–1322, doi:10.1109/TITS.2014.2300103.
[23] S. Robben, G. Englebienne, B. Kröse, Delta features from ambient sensor data are good predictors of change in functional health, IEEE J. Biomed. Health Inform. 21 (4) (2017) 986–993, doi:10.1109/JBHI.2016.2593980.
[24] A.R. Rocha, L. Pirmez, F.C. Delicato, írico Lemos, I. Santos, D.G. Gomes, J.N. de Souza, {WSNs} Clustering based on semantic neighborhood relationships, Comput. Netw. 56 (5) (2012) 1627–1645, doi:10.1016/j.comnet.2012.01.014.
[25] M. Bahrepour, Y. Zhang, N. Meratnia, P.J.M. Havinga, Use of event detection approaches for outlier detection in wireless sensor networks, in: 2009 International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2009, pp. 439–444, doi:10.1109/ISSNIP.2009.5416749.
[26] X. Luo, M. Dong, Y. Huang, On distributed fault-tolerant detection in wireless sensor networks, IEEE Trans. Comput. 55 (1) (2006) 58–70, doi:10.1109/TC.2006.13.
[27] Micaz, (http://www.memsic.com/userfiles/files/Datasheets/WSN/micaz_datasheet-t.pdf).
[28] G. Upton, I. Cook, Understanding Statistics, Oxford University Press, 1996.
[29] Matlab, (http://in.mathworks.com/help/matlab/).
[30] Intel lab data, 2004, (http://db.csail.mit.edu/labdata/labdata.html).
[31] L. Zhu, Z. Huang, Y. Liu, C. Yue, B. Ci, The nonparametric bayesian dictionary learning based interpolation method for wsns missing data, AEU - Int. J. Electron. Commun. 79 (2017) 267–274, doi:10.1016/j.aeue.2017.06.005.

Srijit Chowdhury is currently working for the Ph.D. degree in Engineering at Indian Institute of Engineering Science and Technology, Shibpur (formerly Bengal Engineering and Science University, Shibpur). His current research is focused on efficient data gathering techniques in wireless sensor network. He received B.Sc. (Hons.) in Economics from Calcutta University, Master of Computer Application from Sikkim Manipal University, and Master of Technology in Information Technology from Bengal Engineering and Science University, Shibpur, India. He is a student member of IEEE, IEEE ComSoc and member of ISOC, India, Kolkata Chapter. He can be communicated through email at srijitc@it.iiests.ac.in and csrijitc@gmail.com.

Ambarish Roy is currently working as Research Scientist at National Remote Sensing Center (NRSC), Indian Space and Research Organisation (ISRO), under Dept. of Space, Govt. of India. A. Roy completed Master of Technology in Information Technology with specialisation in Information and Communication Engineering from Indian Institute of Engineering Science and Technology, Shibpur in 2017. He received Bachelor of Engineering in Computer Science and Engineering from University of Burdwan in 2015. His research interests include Artificial Intelligence, Natural Language Processing and Wireless Sensor Network. He can be communicated through email at ambarish268@gmail.com.

Abderrahim Benslimane received the B.S. degree from the University of Nancy in 1987, the DEA (M.S.) degree from the Franche-Comte University of Besançon in 1989, and the Ph.D. degree in 1993, all in computer science. He has been a Professor of computer science with Avignon University, France, since 2001. He was a Technical International Expert with the French Ministry of Foreign and European Affairs from 2012 to 2016. He served as a Coordinator with the Faculty of Engineering, French University, Egypt. He is Editor in Chief of Multimedia Intelligence and Security Inderscience Journal, Area Editor of Security IEEE IoT Journal, Area Editor of Security and Privacy Wiley Journal and EB member of several other journals. His research interests are in development of communication protocols with the use of graph theory for mobile and wireless networks. He was a recipient of the French Award of Doctoral Supervisions from 2017 to 2021 and attributed the French Award of Scientific Excellency from 2011 to 2014. He was a recipient of the title to supervise researches (HDR 2000) from the University of Cergy-Pontoise, France. He served as a Symposium Co-Chair/Leader in many IEEE international conferences such as ICC, GLOBECOM, AINA, and VTC. He is currently the Chair of the ComSoc TC of Communication and Information Security. He has been an Associate Professor with the University of Technology of Belfort-Montbéliard since 1994.

Chandan Giri received B.Tech degree in Computer Science & Engineering from Calcutta University, Kolkata, India in 2002 and subsequently Masters of Engineering (M.E) in Computer Science & Engineering from Jadavpur University, Kolkata, India in 2002 and the Ph.D degree from the Department of Electronics & Electrical Communication Engineering, Indian Institute of Technology, Kharagpur in 2008. He served as an Assistant Professor in the department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, India, from 2008 to 2018. He is currently serving as an Associate Professor in the department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, India. His research interests are Wireless Sensor Networks, testing and design-for-testability of integrated circuits (especially 3D and multicore chips), Micro-fluidic Biochip Design and Test. He served as a Conference/Symposium Co-Chair in many IEEE international conferences such as VLSI Design, VLSI Design and Test (VDAT), ISED, ISDCS, RC etc. He is a member of IEEE and ACM. He is currently one of the executive members of IEEE CAS Chapter, Kolkata, India. He can be contacted at: chandan@it.iiests.ac.in and chandangiri@gmail.com.