TELKOMNIKA, Vol.15, No.3, September 2017, pp. 1477~1484
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v15i3.7217

Enhanced Payload Data Reduction Approach for Cluster Head (CH) Nodes

N. A. M. Alduais*, J. Abdullah, A. Jamil
Wireless and Radio Science Center (WARAS), Faculty of Electrical and Electronic Engineering,
Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, Batu Pahat, Johor, Malaysia
* Corresponding author, e-mail: naifalduais@gmail.com

Abstract

In this paper, we suggested two approaches to minimizing the CH packet size by considering the accuracy of prediction of sensed data at the base station. The proposed coding schemes based relative difference (CS-RD) and based the factor of precision (CS-FP) instead of the absolute change method that has been used in recent work. The aim is to enhance the accuracy of prediction data at the base station. Therefore, the performance metric was evaluated in term of the accuracy of prediction data at the base station. Simulation results showed that the proposed approaches performed better in term of the accuracy of prediction data at the base station. Specifically, the distortion percentage and average Absolut error in the CS-RD and CS-FP method decreased by 50% and 88% better than the current new aggregation method (ADATDC). However, our proposed CS-FP showed a low reduction ratio for some states.

Keywords: *WSN; Data Collection, Accuracy, CH Payload packet size, Coding Scheme*

1. Introduction

Recently, WSN has become an enabler technology for the IOT applications, thus extending the physical reach of the monitoring capability. WSN, as it is, possesses several constraints such as a limited energy availability, a low memory size, and a low processing speed that are the principal obstacles to designing effective management protocols for WSNs [1]. This also concerns WSN-IOT integration. In IOT based WSN, the basic issues are concerned with the mechanism to reduce the energy consumption of the CH nodes that will result in prolonging the lifetime of the CH nodes. It is a very significant consideration since the energy consumed by transmitting one bit via WSN is higher than running numerous microcontroller instructions [2]. There are various assumptions in modeling energy consumption of sensor nodes. Most of this modeling focuses on calculating energy consumption while considering the cost of energy consumption in transmitting and receive modes, ignoring the cost of the process because the cost of transmitting one bit via WSN is higher than running numerous instructions in terms of energy consumption.

CH node packed size of the sensed data aggregation through clustering is the most common issue. Moreover, knowing the number of nodes that transmit sensed data to the CH still affects the CH payload data size. Furthermore, the cluster packet size is limited, where the aggregation data from the nodes must be equal or less than the payload data size. Reducing the packet size will also decrease energy consumption by the CH as well as prolong its lifetime. Recent work has proposed an approach to reducing the packet size for the CH node based on a coding scheme. Therefore, the aim of work reported in this paper is to enhance the payload data reduction approach for cluster head nodes.

2. Related Works

Collecting and reducing data at a CH node in the network is dependent on the Collecting and reducing data at the CH node in a given network is dependent on the performance of aggregation methods. That method is classified as a method used for aggregation without reducing the data and aggregation with reducing data. When the packet size for the CH node is assumed to be large enough to accommodate a large aggregate of data, the aggregation without reducing packet size approach is used. However, when the CH packet

Received May 29, 2017; Revised August 18, 2017; Accepted August 30, 2017

size is limited, we need to reduce the collected data using some other techniques, such as taking the mean, median, and min/max values of samples of collected data as these approaches require a minimum amount of bits. The idea behind this method of using one value to represent the set of collected data is that these techniques can give the lowest accuracy as some of the individual sensors need to sense the data.

In previous studies [3-7], the authors assumed that the CH was selected only for highly spatially correlation among the member nodes for the same CH. The idea behind this was to send only the CH node sensed data as a representative of all cluster member sensed data.

In [8], the authors developed an adaptive method of data aggregation with the aim of exploiting the spatial correlation between sensor nodes (ADATDC). ADATDC is a method that is applied on the cluster head node level with univariate data and each cluster member node supports a single sensor. They also used two bits for representing the sign (+/-) change value between the sensor node and the median of sensed data from all cluster members.

In this paper, we suggested two approaches to minimizing the CH packet size by considering the accuracy of prediction of sensed data at the base station. The proposed coding schemes are the based relative difference (CS-RD) and the based factor of precision (CS-FP) instead of the absolute change method that used in ADATDC. The relative difference/change has been employed as an update data strategy (sensor node level) with a fixed threshold used in recent work [9]. Therefore, in this study, we employed the CS-RD and CS-FPinstead of the absolute change to enhance the accuracy of prediction of data at the base station.

3. Proposal of Novel Approaches to Minimize the Packet Size for CH Node

3.1. CS-RD

In the section, we explain the CS-RD approach to reducing the packet size for the cluster head.

Problem Formulation: It is well known that the number of nodes $S_N(n)$ that transmit sensed data to the CH has an effect on the CH payload data size (P_{DS}). In addition, the cluster payload size is very dependent on the number of nodes in the CH. Furthermore, the cluster packet size is limited, where the aggregation data from the nodes must be equal or less than the payload data size$D_{agg} \leq P_{DS}$.

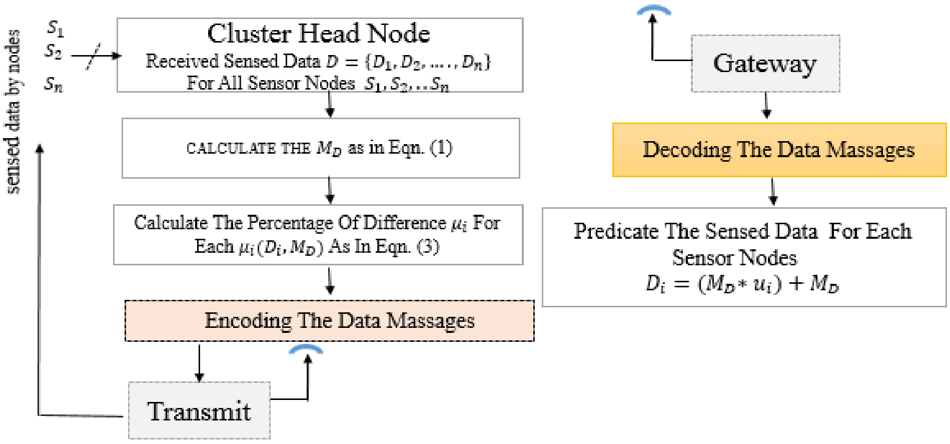
sensed data by nodes
S_1
S_2
S_n
Cluster Head Node
Received Sensed Data $D = \{D_1, D_2, ..., D_n\}$
For All Sensor Nodes $S_1, S_2, ..S_n$
CALCULATE THE M_D as in Eqn. (1)
Calculate The Percentage Of Difference μ_i For Each $\mu_i(D_i, M_D)$ As In Eqn. (3)
Encoding The Data Massages
Transmit
Gateway
Decoding The Data Massages
Predicate The Sensed Data For Each Sensor Nodes
$D_i = (M_D * u_i) + M_D$

Figure 1. The Proposed method flowchart

Algorithm Description: We divide the algorithm into two distinct phases: the CH node level, and the Gateway level as shown in Figure 1.

CHNode Level: In this phase, the steps taken to reduce the CH node payload packet size is described.

Step 1: Let us consider that the number of sensor nodes in CH is $S = \{S_1, S_2, ..., S_n\}$; S_i is i-th sensor, then the sensed data by the sensor nodes are $D = \{D_1, D_2, ..., D_n\}$; S_i is the i-th sensor node and its sensed data D_i i-th respectively, $i = \{1,2,\dots,n\}$, and (n) the number of sensor nodes. The median of the sensed data (M_D) is defined as in Equation (1).

$$\boldsymbol{M_D} = \begin{cases} \frac{\boldsymbol{D_{(n-1)}}}{\boldsymbol{2}} \boldsymbol{for\ (n)\ isodd} \\ \frac{\boldsymbol{D_{((n*0.5)+1)} + D_{(n/2)}}}{\boldsymbol{2}} \boldsymbol{for\ (n)\ iseven} \end{cases} \tag{1}$$

The correlation between the sensed date is D_i by a sensor node S_i and the median sensed data for all sensor nodes M_D is defined as in Equation (2), where $r_i \in [1,0]$, when r_i =1, which means that the sensed data D_i has a high correlation. Conversely, r_i = 0 means that the sensed data D_i is weak or has no correlation with $M_D.\beta$, and the allowed amount of the difference is selected by the sink, where its value is totally dependent on the application error tolerance.

$$\boldsymbol{r_i} = \begin{cases} \boldsymbol{1}, & \boldsymbol{\sqrt{(D_i - M_D)^2} \leq \beta} \\ \boldsymbol{0}, & \boldsymbol{otherwise} \end{cases}, \tag{2}$$

This study focuses on the accuracy of the predicted data by the sink and the number of bits coding. For this reason, we assume all aggregated data D are correlated.

Step 2: Estimations of the percentage of the difference μ_i between the sensor node value $D_i$and the median sensed data for the CH node samples M_D are as defined in Equation (3).

$$\boldsymbol{\mu_i} = \left(\frac{\boldsymbol{D_i - M_D}}{\boldsymbol{(D_i + M_D)0.5}}\right) \times \boldsymbol{100} \tag{3}$$

Step 3: $S_b(i)$ isa one bit for the sign(+/-) based on the difference between the data D_i, and $M_D$as shown in Table 1.

Table 1. One-bit coding for the sing Data Message D_i

If	$Code(S_b(i))$
$D_i \geq M_D$	0
$D_i < M_D$	1

Step 4: Representing the percentage of similarity $\mu_i(D_i, M_D)$in binary format based Decimal to binary conversion (BCD) b_{u_i} is required for $|\mu_i|$ as shown in Table 2. From Equation (4) we can estimate the number of bits required to send both $D_i$and M_D value in bits' form b_{D_i} and b_{M_D} respectively.

$$\boldsymbol{P_{DS} \geq b_{M_D} + n\ (sing_{bit} + b_u)} \tag{4}$$

Table 2. $(2)^{n-1\ \mathrm{bits}}$ coding for the value of D_i

$\boldsymbol{D_i\ Code}$	…..	$\boldsymbol{(2)^2}$	$\boldsymbol{(2)^1}$	$\boldsymbol{(2)^0}$	$\boldsymbol{\mu_i}$
0/1		0	0	0	**0%**
0/1	….	0	0	1	**1%**
0/1	….	0	1	0	**2%**
⋮	⋮	⋮	⋮	⋮	⋮
0/1	…	1	1	1	**7%**

Gateway level: The base station can predict the real sensed data for each sensor node S_i ∈ {CH} as defined in Eqnuation (5).

$$\boldsymbol{D_i = (M_D * u_i) + M_D} \tag{5}$$

Algorithm 1 (CH node Level) Encoding the Sensed Data

```
Inputs:{ $D$},$n$  sensed data array for the sensor nodes {S} in the CH; number of sensor nodes in CH
Output: $\{u\}$ $P_{DS}$,: array of the difference, The CH payload data
Begin: //Arrange the sensed data array {D} from smaller to larger value
Calculate $M_D$  // as defined in Eqn. (1).
Convert $bM_D \leftarrow binary\ format\ (M_D)$
For $i = 1{:}n$ Do // i=1, 2,n ;
    Set $u(i) \leftarrow ((D(i) - M_D)/(D(i) + M_D) * 0.5)) * 100$/ as defined in Eqn. (3).
    IF $u(i) >= M_D$ Then
        $Sb(i) = 0$;  Else
        $Sb(i) = 1$; End if
    Set $u(i) \leftarrow abs(u(i))$; Set $bu(i) \leftarrow BCD(u(i))$ // Convert the u(i) from decimal to binary
    $SD(i) \leftarrow [Sb(i)\ bu(i)\ ]$ //
Next
Set $P_{DS} \leftarrow bM_D + \{SD\}$ // Sensed data” in bits’ format” for all sensor nodes in CH
Send ($P_{DS}$ ) To BS // Send the data to the base station (BS)
End Algorithm
```

Algorithm 2 (Sink Level) Prediction the sensed Data

```
Inputs:{ $P_{Ds}$}  sensed data packet for the sensor nodes {S} in the CH
Output: $\{RD\}$: Received array for sensed data array by the sensor nodes {S} in the CH
Begin: $M_D \leftarrow Binary\ To\ Decimal\ (bM_D)$
//Estimate Sb(i) bu(i) bits from the data for each sensor separately
For $i = 1{:}n$ Do  // i=1,2,..,n ;
    $D(i) \leftarrow Binary\ To\ Decimal\ (bu(i))$// If Sb(i) ==1 Then
        Set $D(i) \leftarrow D(i) * -1$;
        Set $RD(i) \leftarrow (D(i) * M_D) + M_D$; // as defined in Eqn. (5)
    Else
        Set $RD(i) \leftarrow (D(i) * M_D) + M_D$;
    End If  // Next
End Algorithm
```

3.2. CS-FP

CS-FP is a method to enhancing the accuracy of predicted sensed data at the BS. In this part, we discuss another solution based on the factor of precision (ω) by modifying the main proposed CS-RD algorithm as illustrated in the following (i) Equation (3) and replacing Equation(5) by Equation (6) and Equation (7), respectively.

$$\boldsymbol{u_i = \lceil |D_i - M_D| \times \omega \rceil} \tag{6}$$

$$\boldsymbol{D_i = \left(\frac{u_i}{\omega}\right) + M_D} \tag{7}$$

For example, if u =0.5 by applying ADATDC Round (0.5) becomes u=1 and by applying CS-FP (0.5× ω) becomes u=5 where ω=10, and at the BS, we return it again by dividing 5/ ω. More details about benefits and cons of CS-FP analysis compared to the main proposed method (CS-RD), and ADATDC as well as their performance will be assessed through simulations in the next section.

4. Performance Evaluation

The main differences between our proposed methods and the original work ADATDC are:

1. **ADATDC** represents the distance between M_D for CH samples, and all members are based on the absolute change with nearest the value to near integer decimal number $\lceil |D_i - M_D| \rceil$ defined as:

$$\boldsymbol{u_i = Round\ (\ D_i - M_D)} \tag{8}$$

For more accuracy, we proposed the percentage of difference formula or used the factor of precision (ω) instead of the absolute change to estimate the difference between M_D for CH samples and all members as defined in Equation (3) and Equation (6), respectively.

2. **ADATDC** uses a 2-bit format for representing the sensed bit (Sb), where Sb=00 when the $Sd_i = MD$ Sb=01 when the $Sd_i > MD$ and Sb=10 when $Sd_i < MD$ In our proposed methods, we only need one bit to represent the sign-bit as shown in Table 1.

As we mentioned early, the performance metric for the algorithm evaluation is the accuracy of data predicted at the sink. Hence, to test our proposed approaches and ADATDC accuracy, we applied the percentage of distortion equation and average absolute Error equation as defined in Equation (9) and Equation (10), respectively.

$$\textbf{Percentage of Distortion \%} = \frac{|SD_i - RD_i|}{SD_i} \times 100 \tag{9}$$

$$\textbf{Average absolute Error} = \frac{\sum_{i=1}^{n} |SD_i - RD_i|}{n} \tag{10}$$

4.1. Mathematical Evaluation And Analysis

To mathematically evaluate our proposed methods with those methods proposed in recent work, specifically in [8], both methods were applied for the same example in [8]. Suppose that there are 13 sensor nodes in a cluster sent to their cluster head with ID = #99, then, the data is measured by the nodes as shown in the 2nd column in Table 3. The distortion percentage between the real sensed data and the received data is defined in Eqn. (9). From Figure 2., it obvious that the proposed CS-RD and CS-FP methods show more accuracy than ADATDC, where the average distortion percentages between the real sensed data SD_i and the received data RD_i for all sensor nodes in the CH are 0.28 % and 0.00%. In contrast, the average distortion percentage by applying the ADATDC method is 0.53 %. But CS-FP has limitations in data reduction ratio, which is further discussed in the next section.

Table 3. Table on a CH to Collection Correlated Data

No.ID	Sensed Data	Received data at Sink			Distortion Percentage %		
	$M_D = 34.46$	ADATDC	CS-RD	CS-FP	ADATDC	CS-RD	CS-FP
99	35.36	35.46	35.49	35.36	0.28	0.37	0.000
11	33.44	33.46	33.43	33.46	0.06	0.03	0.001
22	33.44	33.46	33.43	33.46	0.06	0.03	0.001
33	34.46	34.46	34.46	34.46	0.00	0.00	0.000
47	32.9	32.46	32.74	32.86	1.34	0.49	0.001
49	34.3	34.46	34.46	34.26	0.47	0.47	0.001
62	34.98	35.46	34.8	34.96	1.37	0.51	0.001
75	34.98	35.46	34.8	34.96	1.37	0.51	0.001
81	34.46	34.46	34.46	34.46	0.00	0.00	0.000
83	35.36	35.46	35.49	35.36	0.28	0.37	0.000
85	35.36	35.46	35.49	35.36	0.28	0.37	0.000
93	34.46	34.46	34.46	34.46	0.00	0.00	0.000
94	32.9	32.46	32.74	32.86	1.34	0.49	0.001
Average Distortion Percentage %					**0.53**	**0.28**	**0.000**

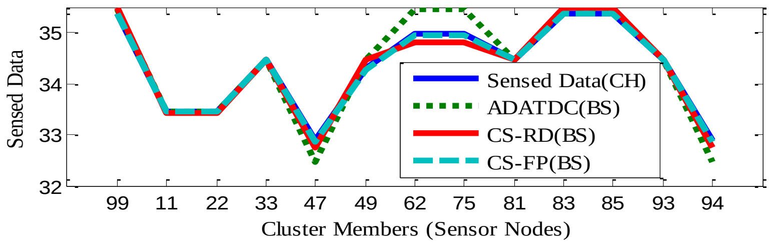

Figure 2. Performance Evaluation of the Prediction

4.2. Simulation And Analysis

The detailed simulation study that tested the performance and accuracy of our approach is presented in this section. The evaluation was performed based on numerous scenarios. The dataset used in this study was extracted from the WSN deployments for Intel Berkeley Research Lab (IBRL) [10]. Subsets of this dataset were selected for assessing the proposed algorithms. MATLAB software was used to write the code for evaluating the performance of the proposed methods.

From Figures 3, 4 and 5, we can observe that the proposed methods in this study show better performance than ADATDC in terms of accuracy, where the prediction sensed data at the sink by our methods is more accurate than that of the ADATDC. The average Absolute error (ABE) and the percentage of distortion were also estimated for each CH sample by applying the equations as defined in Eqn. (9) and Eqn. (10), respectively. The ABE for six nodes cluster members with 14 samples as an outcome of applying the ADATDC are {0.219 and 0.195} for temperature and humidity sensors, respectively. Conversely, the ABE after applying our methods are {0.103 and 0.067} for temperature and humidity sensors, respectiveThus, based on the results, it is clear that the performance of CS-RD and CS-FP is better by approximately 50% and 88% than that of the ADATDC. It is worth noting that the CS-FP method showed the best accuracy at all. But it has a low data reduction percentage as displayed in Figure 6. for the humidity sensor. To explain that, if u=3.5 by applying ADATDC Round (3.5) become u=4 and by applying propoesd2 (3.5× ω) becomes u=35 where ω=10 implies that the CS-FP showed a low reduction ratio. We can estimate that CS-FP is benefited merely when the (u) value was in the range [0- 0.7]. Otherwise, the main proposed (CS-RD) showed the best solution in terms of the accuracy and reduction ratio.

The result is very dependent on the method used, wherein the ADATDC is used to represent the distance between M_D for CH samples and all members are based on the absolute change with nearest the value to the near integer decimal number $[|D_i - M_D|]$, which means that if the distance between (D_i, M_D) = {0.5, 0.4}, it will become {1,0} respectively. In this case, if the originally sensed data value is 35.5 and 35.4 the sink will receive it as 36 and 35, respectively. This problem is solved by the CS-RD and CS-FP methods based on the relative difference/ factor of precision(ω), thus increasing the accuracy as defined in Eqn. (3) and Eqn.(6), respectively. More details and example are discussed in 3.1, the "Mathematical Evaluation and Analysis" section in this study. In addition, for the advantage of our proposed methods, we used only one bit to represent the state of the data message. As shown in Table 1, the ADaTDC method used two bits for that purpose to show how it assists in decreasing the number of bits.

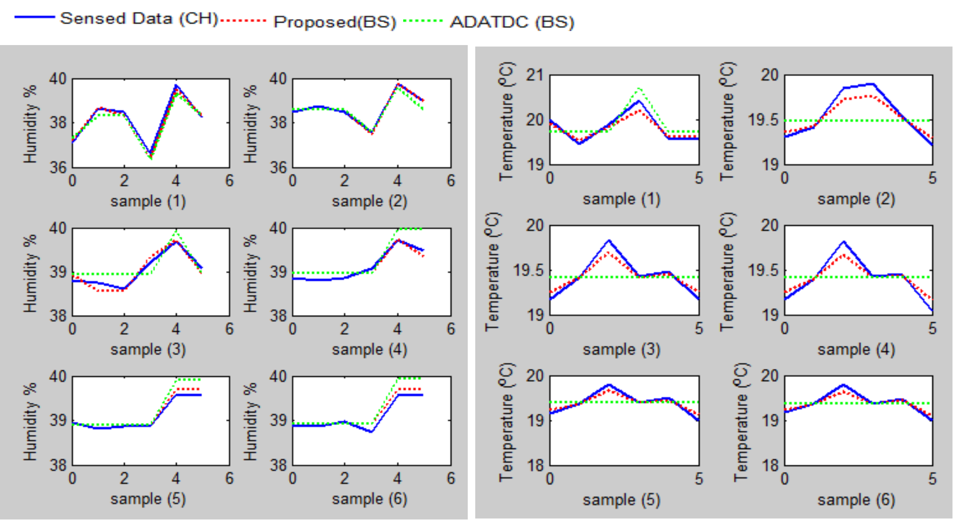

Figure 3. Performance Evaluation of the Prediction sensed data for temperature and humidity sensors

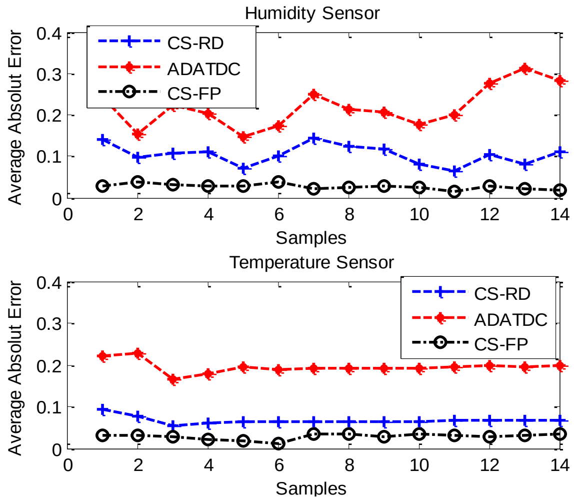

Figure 4. Average Absolute Error

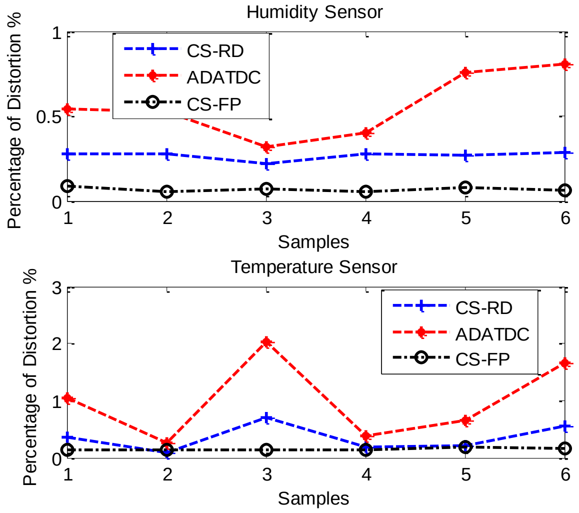

Figure 5. Percentage of Distortion %

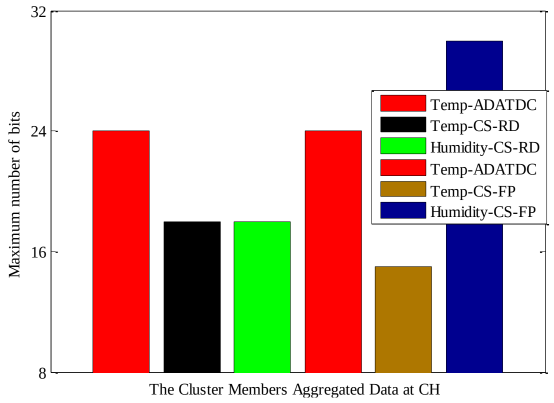

Figure 6. Maximum number of bit requiring for cluster member aggregated data at CH per epoch

5. Conclusion

In this study, we proposed two approaches to reducing the CH packet size by considering the accuracy of prediction of sensed data at the base station. The proposed coding schemes, namely; CS-RD and CS-FP were used in this study instead of the absolute change method used in recent work where ADATDC improves the accuracy of approximation data at the BS. Therefore, the performance metric was evaluated in terms of the accuracy of prediction data at the BS. Simulated results showed that the proposed approaches performed better in term of the accuracy of prediction of data at the base station, where the distortion percentage and average Absolut error in the CS-RD and CS-FP method decreased by 50% and 88% than those of the ADATDC approach. However, the CS-FP showed the lowest reduction ratio in some states.

Acknowledgement

The authors would like to thank the staff of Wireless and Radio Science Center (WARAS) of Universiti Tun Hussein Onn Malaysia (UTHM) for support. This research is supported by the Fundamental Research Grant Scheme(FRGS) vot number 1532 from the Ministry of Education Malaysia.

References

[1] Alduais NAM, Audah L, Jamil A, Abdullah J. *Performance evaluation of different logical topologies and their respective protocols for wireless sensor networks*. *ARPN J Eng Appl* Sci. 2015; 10(19): 8625-8634.

[2] Bispo KA, Rosa NS, Cunha PR. *A semantic solution for saving energy in wireless sensor networks. InComputers and Communications (ISCC*), 2012 IEEE Symposium on. IEEE. 2012; 000492-000499.

[3] Karjee J, Jamadagni HS. Data accuracy model for distributed clustering algorithm based on spatial data correlation in wireless sensor networks. arXiv preprint arXiv:1108.2644. 2011.

[4] Bahrami S, Yousefi H, Movaghar A. *Daca: data-aware clustering and aggregation in query-driven wireless sensor networks.* In2012 21st International Conference on Computer Communications and Networks (ICCCN). IEEE. 2012; 1-7.

[5] Ma Y, Guo Y, Tian X, Ghanem M. Distributed clustering-based aggregation algorithm for spatial correlated sensor networks*. IEEE Sensors Journal*. 2011; 11(3): 641-8.

[6] Cho CY, Lin CL, Hsiao YH, Wang JS, Yang KC. *Data aggregation with spatially correlated grouping technique on cluster-based WSNs*. InSensor Technologies and Applications (SENSORCOMM). 2010 Fourth International Conference on. IEEE. 2010; 18: 584-589.

[7] Kasirajan P, Larsen C, Jagannathan S. *A new data aggregation scheme via adaptive compression for wireless sensor networks.* ACM Transactions on Sensor Networks (TOSN). 2012; 9(1): 5.

[8] Enam RN, Qureshi R. *An adaptive data aggregation technique for dynamic cluster-based wireless sensor networks*. In2014 23rd International Conference on Computer Communication and Networks (ICCCN). IEEE. 2014; 1-7.

[9] Alduais NAM, Abdullah J, Jamil A, Audah L. *An efficient data collection and dissemination for IOT based WSN*. In Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE 7th Annual. IEEE. 2016: 1-6.

[10] Intel Lab Data. Db.csail.mit.edu, 2016. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html. [Accessed: 01- Apr- 2016].