Annals of Telecommunications (2021) 76:321–329
https://doi.org/10.1007/s12243-020-00823-x

An energy-efficient data aggregation approach for cluster-based wireless sensor networks

Syed Rooh Ullah Jan[1] · Rahim Khan[1] · Mian Ahmad Jan[1]

Received: 12 August 2020 / Accepted: 2 November 2020 / Published online: 20 November 2020

Abstract

In wireless sensor networks (WSNs), data redundancy is a challenging issue that not only introduces network congestion but also consumes considerable sensor node resources. Data redundancy occurs due to the spatial and temporal correlations among the data gathered by the neighboring nodes. Data aggregation is a prominent technique that performs in-network filtering of the redundant data and accelerates knowledge extraction by eliminating the correlated data. However, most data aggregation techniques have low accuracy because they do not consider the presence of erroneous data from faulty nodes, which represents an open research challenge. To address this challenge, we have proposed a novel, lightweight, and energy-efficient function-based data aggregation approach for a cluster-based hierarchical WSN. Our proposed approach works at two levels: the node level and the cluster head level. At the node level, the data aggregation is performed using the exponential moving average (EMA), and a threshold-based mechanism is adopted to detect any outliers to improve the accuracy of data aggregation. At the cluster head level, we have employed a modified version of the Euclidean distance function to provide highly refined aggregated data to the base station. Our experimental results show that our approach reduces the communication cost, transmission cost, and energy consumption at the nodes and cluster heads and delivers highly refined, fused data to the base station.

Keywords Wireless sensor network · Data aggregation · Energy efficiency · Accuracy · Outlier detection

1 Introduction

In WSNs, a large number of nodes sense the environment, collect the raw data, and forward it to a centralized base station for further analysis [1]. In these networks, the nodes are densely deployed to generate spatially and temporally correlated data streams [2]. The transmission and processing of these streams consume a considerable amount of the available network resources. All of the network entities suffer considerably while processing and transmitting these streams. Forwarding these redundant data streams exposes the network to various challenges such as energy depletion, bandwidth consumption, and numerous overhead costs that are associated with data communication, processing, and storage [3]. Furthermore, in these networks, the performance of resource-constrained miniature nodes is adversely affected by various in-node operations such as data communication, computation, and sensing. Among these operations, data communication consumes considerably more energy. Transmitting higher volumes of raw redundant streams rapidly depletes the resources of these nodes [4]. Thus, the limited energy of the nodes and the large consumption of energy associated with communication create an imbalance in the network.

To overcome the aforementioned challenges in WSNs, data aggregation is one of the most promising approaches [5–7]. Data aggregation refers to local in-network data summarization from different source nodes while those data are traveling toward the base station. It manages data redundancy by filtering out redundant data and transmits highly refined data upstream toward the base station. This approach not only reduces communication cost, but also maximizes the energy conservation. In literature, the existing data aggregation techniques conserve

✉ Rahim Khan
rahimkhan@awkum.edu.pk

[1] Department of Computer Science, Abdul Wali Khan University, Mardan, KPK, Pakistan

either computation or communication costs, but not both at the same time [8]. In addition, the data aggregation techniques based on duplicate-insensitive functions suffer considerably at the time of eliminating the redundant data [9]. These techniques are inherently insensitive to duplicate, redundant, and erroneous readings from faulty nodes in the network.

A number of data aggregation techniques have been presented in the literature [10, 11]. However, they are designed specifically for resource-efficient nodes, e.g., cluster heads or aggregators. These techniques result in high computational and spatial complexities and are thus infeasible for tiny, resource-constrained sensor nodes [12]. Moreover, accuracy is a major concern with respect to these techniques [13]. Therefore, energy-efficient data aggregation techniques remain an open research challenge. In this work, we propose a lightweight, and energy- and memory-efficient data aggregation technique with improved quality of service (QoS) and accuracy. Our technique enhances the network lifetime by reducing the communication, computation, and storage overhead. The main objectives of our research are as follows.

1. Our proposed technique performs local data aggregation with improved accuracy for resource-constrained sensor nodes. Local data aggregation is performed using simple functions for the *Min* and *Max* along with the exponential moving average (EMA). Data aggregation techniques based on duplicate-insensitive functions suffer from low accuracy due to their inherent insensitivity to outliers. Unlike previous techniques, our proposed novel threshold-based mechanism is designed to detect outliers and minimize their effect on data accuracy.
2. After a predefined time interval, only a fraction of the aggregated data is retained within the memory of each node. In our technique, only eight out of sixty highly refined captured readings are retained. This makes our proposed technique extremely lightweight in terms of memory utilization.
3. Our approach lowers the amount of data comparison at the cluster head level to conserve computational energy. The partially aggregated data streams from the sensor nodes are received at the cluster head, where they are further aggregated. For energy preservation and network lifetime enhancement, only the data streams of neighboring nodes are compared. This is because the neighboring nodes have relatively higher chances of spatial and temporal correlation among each other. This reduces the number of comparisons to a greater extent, which conserves not only the transmission energy, but also the processing energy.

The rest of this paper is organized as follows. In Section 2, a detailed literature review is provided. In Section 3, we discuss our proposed approach for enhancing data aggregation. In Section 4, a comprehensive set of experimental results is provided. Finally, the paper is concluded and future research directions are provided in Section 5.

2 Literature review

One of the fundamental and challenging issues in WSNs is data redundancy. This mainly occurs due to the spatial and temporal correlations among the data captured by these densely deployed sensor nodes [14]. Such highly correlated data consume substantial network resources and cause congestion that negatively affects the QoS of the underlying network. To date, various techniques have been proposed: however, data aggregation is one of the most prominent techniques to reduce redundancy, save energy, and prolong network life [5–7, 15, 16]. In [5], the authors proposed a simple data aggregation technique that operates in two phases. During the first phase, local aggregation is performed at the node level and global aggregation is performed at the cluster head level. During local data aggregation, a similar function is used to aggregate periodically captured data from the sensing field. Next, this partially aggregated data is forwarded toward the cluster heads. During global data aggregation, the cluster heads search for and calculate dissimilarities between the datasets using set similarity functions, the one-way ANOVA model with statistical tests, and various distance-based functions. The proposed approach is designed for period-based WSNs, and as such, its suitability needs to be tested in a network where continuous and query-based data communication occurs. In [6], the authors proposed a novel data aggregation technique based on a k-means algorithm, i.e., the enhanced K-mean (EK-mean). The proposed approach operates at two levels: the sensor level and the aggregator level. First, the Euclidean distance is used at the node level to calculate dissimilarities among the captured datasets. Second, the EK-means clustering algorithm is used at the aggregator to group together all pairs of similar data generated by member nodes in the network. This approach removes the redundant data and significantly reduces data communication across the network. However, the proposed technique does not devise any data comparison mechanism at the cluster head, and thus incurs a higher computational cost. The proposed technique uses a function similar to that used in [5] and as such suffers from low accuracy. Furthermore, this technique operates at the expense of increased energy consumption, which adversely affects the resources of sensor nodes.

In [7], the author proposed a distributed data fusion algorithm, i.e., data summarization in the node by parameters (DSNP). The proposed algorithm performs local data aggregation using various metrics such as Average, Min, and Max that are provided by a remotely installed IoT application server. DSNP performs Bollinger analysis to decide whether or not to write the aggregated data back to the server. The main benefit of DSNP is the quality of locally aggregated data. However, the proposed algorithm operates on the basis of the previous mean value and its deviation from the current value. The absence of such values can adversely affect the performance of DSNP. Moreover, DSNP uses moving averages, and as such does not consider the presence of erroneous readings from the faulty nodes. A dynamical message list-based data aggregation (DMLDA) approach was proposed for cluster-based WSNs in [15]. DMLDA aimed at designing a computationally lightweight, real-time, and resource-efficient filtering scheme for distributed network architectures. The proposed approach defined a special data structure at each aggregator node that contains the history of the data prior to its transmission. This data structure is used for real-time data aggregation by filtering irrelevant, repeated, and redundant data from the captured raw data. The proposed approach achieves data aggregation at cluster heads only, and, as such, lacks data aggregation at the basic level, i.e., at the member nodes. As a result, the cluster heads drain their energy much faster. Moreover, the proposed technique does not take into account the number of packets traversing the network. Since DMLDA uses a list for data filtering, it becomes inefficient, particularly if the list is incomplete. Moreover, DMLDA does not devise any strategy in the absence of a list used for data aggregation. Due to real-time data aggregation, this method is infeasible for networks where periodic, continuous, or query-based data communications occur. The authors in [16] have proposed two novel data aggregation techniques for WSN, deployed in the field of agriculture, to sense various parameters such as temperature, air humidity, and soil humidity. The first technique employs a simple moving average (SMA) mechanism, particularly at the local level. The captured data are divided into equal slots, and the SMA of each slot is calculated to remove the correlated data. The same is then forwarded toward the final destination. In the second technique, a modified version of the threshold-sensitive energy-efficient sensor network (TEEN) protocol [17] was employed, particularly at those nodes that have a single sensor board, i.e., performing only one task, as opposed to the previous scheme where each sensor performs multiple tasks. TEEN is a reactive protocol [18] which is feasible for real-time applications. Because it avoids transmitting similar values in an energy-efficient manner, TEEN extends the network lifetime. Simulation results have proven that both techniques of [16] reduced the energy consumption. However, these techniques lack outlier detection, which adversely affects data accuracy. Most of the existing average based approaches result in the final aggregated value being negatively affected by the erroneous values captured by the sensor nodes, thus raising the question of the accuracy of these techniques. Moreover, applicability of the proposed techniques must be evaluated in large-scale WSNs, since the performance of reactive protocols decreases as the size of the network increases.

Although many data aggregation techniques exist in the literature, most of them are not computationally lightweight and have lower accuracy. The QoS of these techniques requires serious attention from both academia and industry in the field. More importantly, these techniques achieve data aggregation at the expense of increased energy consumption at the node, as well as at the cluster head level.

3 Proposed approach

We consider a network with a total of T_n nodes, where $T_n = \{1, 2, 3, ..., n\}$. These T_n are grouped together into equal-sized clusters C, where the total clusters $T_C = \{C_1, C_2, C_3....C_n\}$, similarly to the balanced clustering algorithm [19, 20]. In this paper, we consider five clusters, i.e., $T_C = 5$. Upon cluster formation, one-hop neighbor node discovery is initiated. As a result, each node within a cluster obtains its immediate one-hop neighbor information and retains it within its buffer. This information is later used at the cluster head for data comparison among the nearby neighboring nodes. This step reduces the number of comparisons to a greater extent by saving valuable energy of the network. The objective of our proposed approach is twofold: local and global data aggregation, as discussed in this section.

3.1 Local data aggregation

After network setup, the member nodes start gathering sensor readings from the deployed field. However, these readings contain highly correlated data due to their dense deployment in the field. The transmission of such spatially and temporally correlated data to the cluster heads not only wastes valuable energy and bandwidth, but also degrades the QoS and lifetime of the network [21–23]. Thus, during local aggregation, we aim for partial elimination of correlated data while preserving energy of the network at the same time.

Initially, at the start of time period P, we assume that each member node has captured a single temperature reading t_r/s (second). Thus, the total number of readings $T_r = 60$. Next, these T_r are grouped into s slots, where each slot contains an equal number of captured readings, i.e., $s_i = 10$, where $s_i \in s$. Therefore, the total number of slots $T_s = [1, \cdots, 6]$. At the beginning of time period P, the default value assigned to *Min* is ∞, and the default assigned to *Max* is 0. As soon as a new reading, i.e., t_{r_i}, is captured for any slot s_j, where $s_j \in s$, the reading may either be normal or an outlier. For instance, if t_{r_i} is normal, i.e, within our predefined threshold range, then the value is checked against the current *Min* and *Max* readings stored in the buffer of the node using Algorithm 1.

In this algorithm, A_i represents an array for storing captured readings T_r that range from 1 to 60. Here, the current value (cur-value) stores t_{r_1}, which has an initial value of 0. Our proposed approach repeatedly checks the aforementioned conditions to obtain refined data from the captured data.

Algorithm 1 Local data aggregation algorithm.

```
1.  Required: Received sensor reading from the sensing field
2.  Return: Aggregated sensor reading
3.  Min = $\infty$
4.  Max = 0
5.  $A_i = [10]$
6.  $EMA_1, EMA_2, EMA_3.....EMA_6$
7.  $cur - value = 0$
8.  for i = 1 to 60    Repeat
9.  if ($cur - value$ within ($B_{lower}$ $or$ $B_{upper}$)) then
10. if $Value_{current} \in \{B_{lower}, B_{upper}\}$
11. $A_i = cur - value$
12. switch to step. 25
13. else
14. Algorithm. 2 (cur-value)
15. $A_i = refined - value$
16. go to step 24
17. end if
18. if   (i == 10) then
19. SMA = $\sum_{i=1}^{n} \frac{i}{n}$
20. $A_i = 0$                    ►empty $A_i$
21. else if (i = 20 $or$ 30 $or$ 40 $or$ 50 $or$ 60)   then
22. $EMA = (R \times K) + PreviousEMA \times (1 - K)$
23. $A_i = 0$                    ►empty $A_i$
24. end if
25. if   ($Curr - value$ < Min)  then
26. Min == $Curr - value$
27. else if ($Curr - value$ > Max)  then
28. Max == $Curr - value$
29. else
30. do nothing
31. end if
32. end for
```

To remove outliers and improve the accuracy of our proposed approach, we have designed a simple and novel threshold-based mechanism with predefined lower and upper bounds, i.e., B_{lower} and B_{upper}. These threshold bounds are based on the historical temperature data of a particular location. Whenever a newly captured reading crosses these ranges, an outlier is detected and refined using Algorithm 2. Our proposed outlier detection mechanism differs for the first two outliers of the first slot, i.e., $t_{r_1}, t_{r_2} \in s_1$, and for every first outlier of the remaining slots, i.e., $t_{r_1} \in \{s_2, ..., s_6\}$. Furthermore, it remains the same for the rest of the outliers within these remaining slots based on the following conditions.

1. **For the first two outliers** $[t_{r_1}, t_{r_2} \in s_1]$: At the start of time period P, assume that a newly captured reading t_{r_1} for s_1 is an outlier, i.e., $[t_{r_1} < B_{lower} || t_{r_1} > B_{upper}]$. Then, t_{r_1} is refined by replacing it with the standard temperature value of a specific location, i.e. $tr_1 == B_{lower}$ or $tr_1 == B_{upper}$. Similarly, for t_{r_2} of s_1, if $[t_{r_2} < B_{lower} || t_{r_2} > B_{upper}]$, then t_{r_2} is replaced by previously stored value t_{r_1}, i.e. $[t_{r_2} == t_{r_1}]$.
2. **For every first outlier, i.e.,** $[t_{r_1} \in \{s_2, ..., s_6\}]$: If $t_{r_1} \in \{s_2, ..., s_6\}$ is an outlier, i.e., if $[t_{r_1} < B_{lower} || t_{r_1} > B_{upper}]$, it is refined by replacing it with the EMA of the preceding slot, such that t_{r_1} == EMA. At this stage, we consider EMA instead of the average, due to the absence of sufficient values in a slot to calculate it.
3. **For all other outliers, i.e.,** $[t_{r_3} \cdots t_{r_{60}} \in \{s_2, ..., s_6\}]$: If $[t_{r_3} \cdots t_{r_{60}} < B_{lower} || t_{r_3} \cdots t_{r_{60}} > B_{upper}]$, then the current value is replaced by the average of previously stored values (preferably consecutive readings), i.e., $T_{r_i} = Average(T_{r_{i-1}}, T_{r_{i-2}})$. This process is repeatedly performed for all outliers from the captured readings to obtain refined data, as demonstrated in Algorithm 2.

Algorithm 2 Proposed outlier detection algorithm.

```
1.  Require: Received outliers from the sensing field
2.  Return: correct sensor reading
3.  if $cur - value$ outside ($B_{lower}$  or  $B_{upper}$)
4.  then
5.  if (cur-reading = 1st value of $S_1$) then
6.   $A_i$= standard temperature $C^0$.
7.  else if (cur-reading = 1st and 2nd value of $S_2 - S_6$)
    then
8.   $A_i$ = previous EMA.
9.  else if (cur-reading = 2nd value of $S_1$)     then
10.  $A_i = A_{i-1}$
11. else
12.  $A_i$ = avg ($A_{i-1}, A_{i-2}$)
13. end if
14.  return refined data
```

Once an outlier is removed based on Algorithm 2, an exponential moving average (EMA) of each slot is calculated. However, for slot s_1, a simple moving average (SMA) is calculated because the EMA computation is not possible [16] at this stage. The SMA is calculated using Eq. 1.

$$SMA = \sum_{i=1}^{10} \frac{i}{n}. \tag{1}$$

In this equation, n refers to the number of temperature readings, where each reading is represented by t_r. There are 10 readings in each slot, and hence n is equal to 10. Once SMA of this specific slot s_1 is calculated, then all readings within this particular slot are discarded except the SMA value, an updated Min and an updated Max. These three readings are retained in the memory of each node. Apart from s_1, for all other slots $\{s_2, ..., s_6\}$, the aforementioned procedure is repeatedly applied with a slight modification, i.e., EMA is calculated instead of SMA. To calculate EMA, we use Eq. 2.

$$EMA = (R \times K) + P_{EMA} \times (1 - K) \tag{2}$$

In this equation, R refers to the current temperature reading, and K refers to the weighting factor and is equivalent to $\frac{2}{W+1}$. Here, W represents the sliding window and is equal to 10 in our case. P_{EMA} refers to the EMA of the previous slot.

At the end of time period P, we have six EMA_s, updated Min and Max readings, and the neighboring node information. Thus, we retain only eight highly refined readings for T_r within the buffer of each node, as depicted by Fig. 1.

Our proposed approach is energy-efficient because it transmits only a fraction of the captured data to the cluster head. Hence, our approach is resilient for the elimination of data redundancy and achieves enhanced QoS for the underlying network at the same time.

There are eight readings because of the updated Min, Max and 6 EMA values.

3.2 Global data aggregation at the CH

At the end of each time period P, multiple aggregated datasets are received by each cluster head from their associated member nodes. These datasets were partially aggregated at the local level, and as such, they still possess redundant data. During global data aggregation, our approach further reduces redundancy before forwarding fine-grained data toward the base station. During data aggregation, we strive for data aggregation similar to that proposed by [24], as the latter has numerous advantages associated with it. The main drawback associated with [5, 24], and [6] is that these schemes compare the data among all member nodes at the cluster head. However, our proposed approach compares the data of only nearest neighbors, as shown in Fig. 2.

To retain highly refined data at the cluster heads in an energy-efficient manner, we compare the data collected from only nearby member nodes, since such data has a higher probability of similarity and redundancy. In this figure, data from node 3 is compared against the data of node 2. The resultant is then compared with the data from node 1, which is then forwarded toward the base station. A similar comparison is made for other neighboring nodes that are located within close vicinity of each other. Our proposed approach reduces the number of data comparisons, which not only conserves energy but also decreases data transmission by improving the QoS of the network. We have employed the Euclidean distance function [25–28] for the elimination of redundancy at the cluster head (as depicted

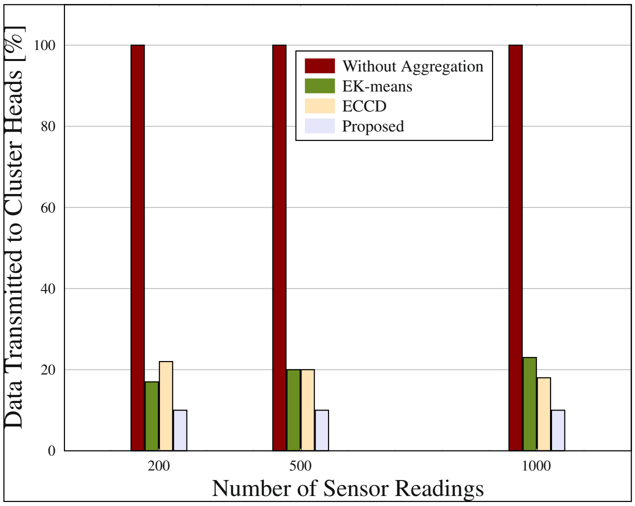

Fig. 1 Local data aggregation

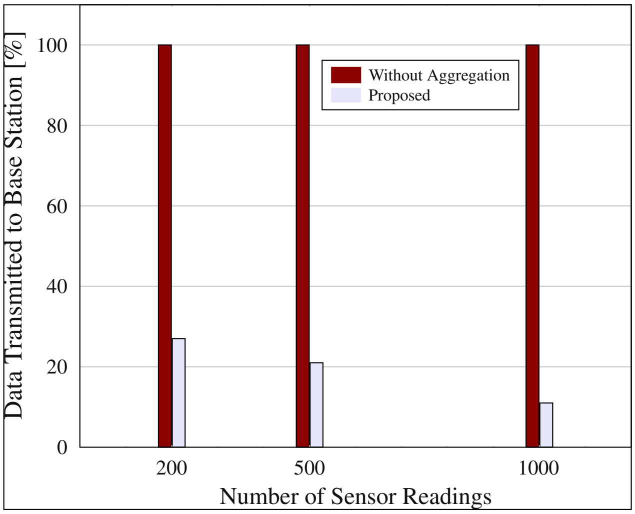

Fig. 2 Data comparison at the cluster head (CH)

by Fig. 2). In WSNs, the Euclidean distance is used for sensor localization during the deployment phase [29–31]. However, in this work, we have employed it for inter-data-set distance estimation [5, 6, 24]. The Euclidean distance (*Ed*) between two datasets S_i and S_j is computed using Eqs. 3 and 4, respectively.

$$Ed = \sqrt{\sum_{n=1}^{8}(S_i n - S_j n)^2} \leq t_d \tag{3}$$

$$Ed = \sqrt{\sum_{n=1}^{8}(S_i n - S_j n)^2} > t_d \tag{4}$$

Next, we compare the resultant values from Eqs. 3 and 4 with our predefined threshold value (t_d), where $t_d = 0.5$, using the following conditions.

1. If Ed is between S_i and $S_j \leq t_d$, then the elements in the two sets are similar; as a result, the elements of a single set are retained, while those of the other are discarded.
2. If Ed is between S_i and $S_j > t_d$, then the elements in the two sets are dissimilar; therefore, the elements of both datasets are retained, as shown in Fig. 3.

We have formulated two disjoint theorems for this purpose.

Theorem 1 *Assume that we have two similar datasets S_i and S_j and that the Ed between them is less than t_d, i.e., $Ed\ (S_i, S_j) \leq t_d.Then,$*

$$Ed = \sqrt{\sum_{n=1}^{8}(S_i n - S_j n)^2} \leq \frac{t_d}{\sqrt{2}} \tag{5}$$

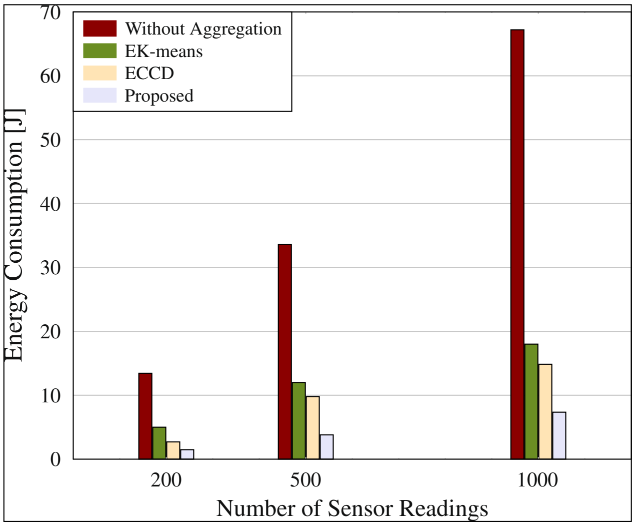

Fig. 3 Global data aggregation

Proof Consider two similar datasets, S_i and S_j, where $s_i \in S_i$ and $s_j \in S_j$. If $Ed\ (S_i, S_j) \leq t_d, then$

$$Ed = \sqrt{\sum_{n=1}^{8}(S_i n - S_j n)^2} \leq t_d$$

$$= \sum_{n=1}^{8}(S_i n - S_j n)^2 \leq {t_d}^2$$

$$= \sum_{n=1}^{8}(S_i n - S_j n)^2 \leq \frac{t_d^2}{2}$$

Hence, it is proven that

$$Ed = \sqrt{\sum_{n=1}^{8}(S_i n - S_j n)^2} \leq \frac{t_d}{\sqrt{2}}$$

Thus, S_i and S_j are considered similar if the Euclidean distance among them is $\leq t_d$, where t_d is the threshold value that we have defined in our proposed approach, which can vary from one application to another. □

Theorem 2 *Assume that we have two datasets S_i and S_j, and E_d between them is $> t_d$, i.e., $Ed\ (S_i, S_j) > t_d$*

$$Ed = \sqrt{\sum_{n=1}^{8}(S_i n - S_j n)^2} > \frac{t_d}{\sqrt{2}} \tag{6}$$

Proof The proof remains the same as Theorem 1. The only difference is that $\leq$ is replaced by $>$. Consider two similar datasets, where each dataset contains six readings collected at a predefined time period "t" with $s_i \in S_i =$ Slot 1 data and $s_j \in S_j =$ Slot 2 data of $Ed(S_i, S_j) > t_d$, then;

$$= \sqrt{\sum_{n=0}^{y_2-1}(S_i n - Sjn)^2} > t_d$$

$$= \sum_{n=0}^{y_2-1}(S_i n - Sjn)^2 > {t_d}^2$$

$$= \sum_{n=0}^{y_2-1}(S_i n - Sjn)^2 > \frac{t_d^2}{2}$$

Hence, it is proven that

$$Ed = \sqrt{\sum_{n=1}^{8}(S_i n - S_j n)^2} > \frac{t_d}{\sqrt{2}}$$

□

4 Experimental results and evaluation

We have performed extensive simulations to evaluate the efficiency and robustness of our data aggregation approach. We simulated our algorithms on an HP Corei7 using MATLAB R2018a. For comparison, we evaluated our approach against those in the latest literature, i.e., ECCD [32] and EK-means [6]. Moreover, we also evaluated the efficiency of our approach in the absence of data aggregation at the local and global levels. For flexibility, we tested our approach for varying numbers of deployed nodes, sensed readings, and threshold values.

Figure 4 shows the percentages of aggregated data transmitted to the cluster heads and base station. In Fig. 4a, local data aggregation for a varying number of sensed readings is computed. In our approach, the sensor nodes rely on B_{lower} and B_{upper} for the elimination of redundant data and outlier detection. When the value of sensor readings (τ) increases from 200 to 1000, the percentage of aggregated data transmitted to the cluster heads decreases. At each sampling interval P, the amount of data collected and transmitted by each node is reduced by at least 90% for τ equal to 200, and by up to 93.5% for τ equal to 1000. For a larger value of τ, the number of redundant readings increases in $|B_{upper} - B_{lower}|$, and hence the probability of redundancy increases. In comparison to our approach, the existing approaches achieve lower data aggregation at the node level. ECCD achieves 77%, 80%, and 82% reductions in transmitted data for 200, 500, and 1000 sensed readings, respectively. EK-Means, on the other hand, achieves 83%, 80%, and 78% reductions in transmitted data for similar values of τ. In Fig. 4b, global data aggregation performed at the cluster head is shown for varying values of τ. The efficiency of cluster heads increases with an increase in the values of τ. In our approach, 27% of the data is transmitted to the base station for τ equal to 200, which plummets to 11% when τ reaches 1000. Without global aggregation, i.e., in the absence of our approach, the cluster heads will transmit all of the received data from the nodes to the base station. This figure does not reflect ECCD and EK-means because these schemes do not employ global data aggregation.

In Fig. 5, a comparison of energy consumption between local and global data aggregation is made. In WSNs, the energy consumption mainly depends on the amounts of transmitted and received data. Figure 5a reflects the energy consumption for local data aggregation with varying values of τ. Local data aggregation alleviates the redundancy in the data sensed by the nodes, and at the same time conserves their energy by reducing the number of transmissions. Our approach consumes approximately 1.48 J of energy at sensor nodes for aggregating 200 readings. In contrast, EK-means consumes 5 J for aggregating the same number of readings. In the absence of local data aggregation, the nodes consume 13.44 mJ of energy for aggregating 200 readings. For aggregating 500 readings, our approach consumes 3.8 mJ, much lower than the energy consumption of the existing approaches. Moreover, our approach achieves much better results when τ reaches 1000. In Fig. 5b, the energy consumed at the cluster head level is computed in the presence or absence of our approach. In comparison to no aggregation at the cluster head, our approach reduces energy consumption by 6% when τ is equal to 200 and by 54% when τ is 1000.

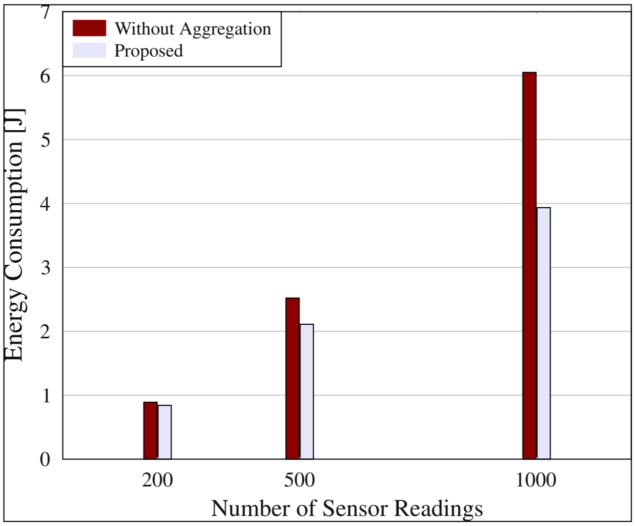

Fig. 4 Data transmission for varying sensor readings

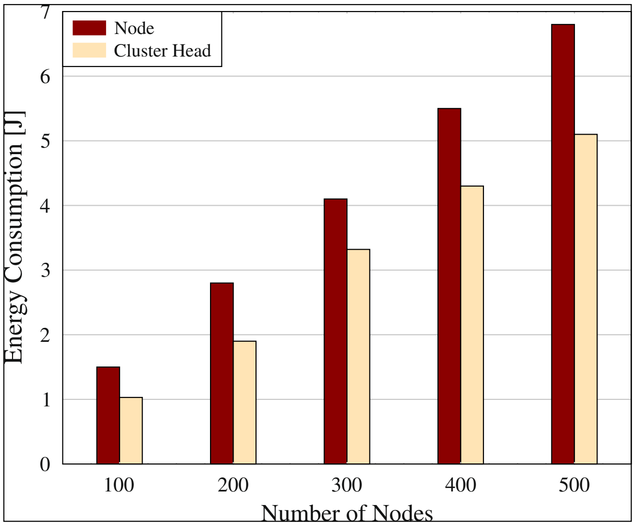

Fig. 5 Energy consumption for varying sensor readings

In Fig. 6, energy consumption for varying numbers of nodes (n) is shown in the presence and absence of data aggregation. The results are computed for 200 sensed readings having a threshold value (t_d) of 0.5. For n equal to 100, the energy consumed is 1.5 J at the node level and 1.03 J at the cluster head level, as shown in Fig. 6a. In comparison, the nodes consume 1.5 times more energy without aggregating the data, and their associated cluster heads consume 12 times more energy, as shown in Fig. 6b. The energy consumption varies significantly when the number of nodes increases. For n equal to 500, our approach consumes 6.8 J at the node level and 5.1 J at the cluster head level. In comparison, the same nodes and cluster heads consume 15.78 J and 67.2 J in the absence of our approach,

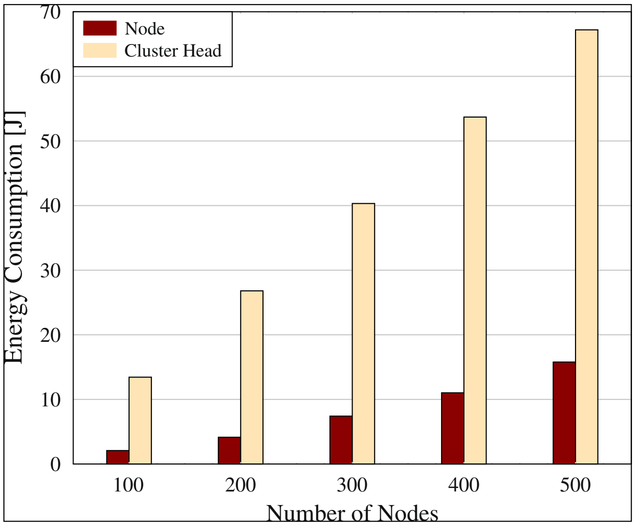

Fig. 6 Energy consumption for varying number of nodes

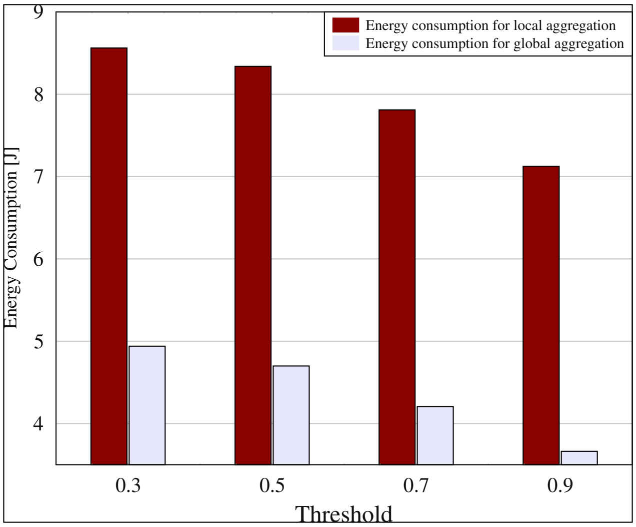

Fig. 8 Energy consumption for varying threshold values

i.e., without aggregation. Without data aggregation, the energy consumption of the clusters is significantly higher due to the distant transmission of raw redundant data to the base station. In the case of aggregation, the cluster heads send only highly refined limited data to the base station.

In Fig. 7, the amount of fused data for varying t_d values is computed at the cluster head level. The results in this graph are calculated for τ equal to 200. These results aim to retain and transmit highly refined aggregated data from the cluster heads to the base station. The use of Euclidean distance and its comparison with t_d determines whether to retain or discard the sensed readings at the cluster heads. Moreover, higher values of t_d indicate that a large number of data packets will be treated as redundant and ultimately discarded. For example, $|t_{r_i} - t_{r_j}| \leq t_d$, where $t_d = 0.5$, means that if the difference between t_{r_i} and t_{r_j} is less than or equal to 0.5, they will be treated as similar readings and only a single copy will be retained by the cluster head. In our approach, as the value of t_d increases, the number of readings transmitted by the cluster heads to the base station decreases because a high volume of redundant data within the threshold range is discarded. Without aggregation, the cluster heads transmit all of the readings received from the sensor nodes to the base station, as shown in this figure. In this case, highly redundant readings are also transmitted, which increases the percentage of delivered data.

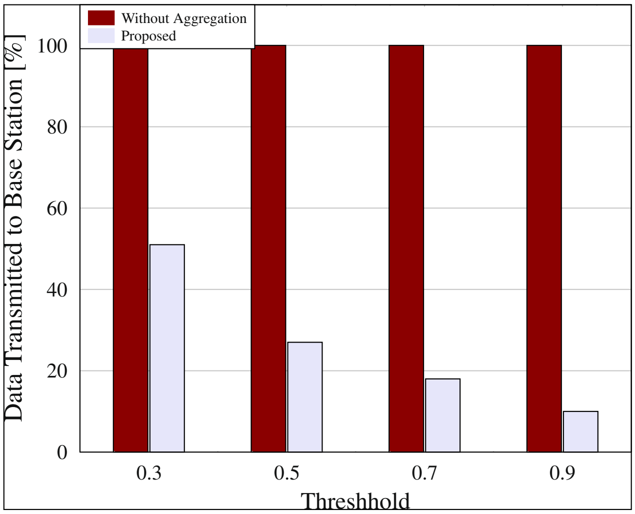

Fig. 7 Data fusion for varying threshold values

In Fig. 8, energy consumption is computed for local and global data aggregation with varying threshold values. At the local level, the SMA and EMA determine the transmission of data to the cluster heads, whereas at the global level, $|t_{r_i} - t_{r_j}| \leq t_d$ determines the transmission of data to the base station. In global data aggregation, as the value of t_d increases, more readings fall within the range $|t_{r_i} - t_{r_j}|$. As a result, a very large volume of redundant data is discarded, which ultimately decreases the energy consumed while transmitting the packets to the base station. Moreover, at the global level, the energy consumption is significantly reduced due to a lower amount of comparison involving only the adjusted neighboring nodes. In local data aggregation, on the other hand, increasing the threshold values has little impact on the number of transmitted packets, leading to higher energy consumption.

5 Limitations and future work

In this paper, we have proposed an application-dependent data aggregation approach for resource-constrained sensor networks. Our proposed approach performs local data aggregation at the sensor nodes using duplicate-sensitive functions such as SMA and EMA. However, the drawback of these functions is that the final outcome is severely affected by the occurrence of duplicate values in a sensed region. Moreover, erroneous reading from the faulty nodes is another challenging issue that adversely affects the accuracy of final outcomes. To address these challenges, we have devised a simple yet novel Euclidean distance–based threshold approach at the cluster heads. As a result, only a fraction of highly refined data are forwarded toward the base station. This approach not only saves the network energy, but also optimizes the memory and bandwidth usage. Our proposed approach faces certain challenges that need to be addressed in the future. For instance, reliability of the refined data transmitted across the network is a major concern in low-power and lossy networks. In the future, we aim to test the efficiency of our approach in other scenarios and network settings. We also aim to include an efficient security framework to safeguard our highly refined aggregated data from malevolent entities.

References

1. Jan MA, Nanda P, He X, Liu RP (2014) Pasccc: priority-based application-specific congestion control clustering protocol. Comput Netw 74:92–102
2. Vuran MC, Akan ÖB, Akyildiz IF (2004) Spatio-temporal correlation: theory and applications for wireless sensor networks. Comput Netw 45(3):245–259
3. Wang Q, Lin D, Yang P, Zhang Z (2019) An energy-efficient compressive sensing-based clustering routing protocol for WSNs. IEEE Sens J 19(10):3950–3960
4. Roy NR, Chandra P (2020) Analysis of data aggregation techniques in WSN. In: International conference on innovative computing and communications. Springer, pp 571–581
5. Harb H, Makhoul A, Laiymani D, Jaber A (2017) A distance-based data aggregation technique for periodic sensor networks. ACM Trans Sens Netw (TOSN) 13(4):32
6. Rida M, Makhoul A, Harb H, Laiymani D, Barhamgi M (2019) EK-means: a new clustering approach for datasets classification in sensor networks. Ad Hoc Netw 84:158–169
7. Maschi LFC, Pinto ASR, Meneguette RI, Baldassin A (2018) Data summarization in the node by parameters (DSNP): local data fusion in an IoT environment. Sensors 18(3):799
8. Wei G, Ling Y, Guo B, Xiao B, Vasilakos AV (2011) Prediction-based data aggregation in wireless sensor networks: combining grey model and Kalman filter. Comput Commun 34(6):793–802
9. Kang B, Nguyen PKH, Zalyubovskiy V, Choo H (2017) A distributed delay-efficient data aggregation scheduling for duty-cycled WSNs. IEEE Sens J 17(11):3422–3437
10. Dhand G, Tyagi SS (2016) Data aggregation techniques in WSN: survey. Procedia Comput Sci 92:378–384
11. Sirsikar S, Anavatti S (2015) Issues of data aggregation methods in wireless sensor network: a survey. Procedia Comput Sci 49: 194–201
12. Sarangi K, Bhattacharya I (2019) A study on data aggregation techniques in wireless sensor network in static and dynamic scenarios. Innov Syst Softw Eng 15(1):3–16
13. Alghamdi W, Rezvani M, Wu H, Kanhere SS (2019) Routing-aware and malicious node detection in a concealed data aggregation for wsns. ACM Trans Sens Netw (TOSN) 15(2):1–20
14. Brahmi IH, Djahel S, Magoni D, Murphy J (2015) A spatial correlation aware scheme for efficient data aggregation in wireless sensor networks. In: 2015 IEEE 40th local computer networks conference workshops (LCN workshops). IEEE, pp 847–854
15. Du T, Qu Z, Guo Q, Qu S (2015) A high efficient and real time data aggregation scheme for WSNs. Int J Distrib Sens Netw 11(6):261381
16. Fajar M, Litan J, Munir A, Halid A et al (2017) Energy efficiency using data filtering approach on agricultural wireless sensor network. Int J Comput Eng Inf Technol 9(9):192
17. Manjeshwar A, Agrawal DP (2001) Teen: arouting protocol for enhanced efficiency in wireless sensor networks. In: ipdps, vol 1, p 189
18. Mohammed IY (2019) Comparative analysis of proactive & reactive protocols for cluster based routing algorithms in WSNs. World Sci News 124(2):131–142
19. Jan MA, Usman M, He X, Rehman AU (2018) Sams: a seamless and authorized multimedia streaming framework for wmsn-based iomt. IEEE Internet Things J 6(2):1576–1583
20. Nikolidakis SA, Kandris D, Vergados DD, Douligeris C (2013) Energy efficient routing in wireless sensor networks through balanced clustering. Algorithms 6(1):29–42
21. Harb H, Makhoul A, Laiymani D, Jaber A, Tawil R (2014) K-means based clustering approach for data aggregation in periodic sensor networks. In: 2014 IEEE 10th international conference on wireless and mobile computing, networking and communications (WiMob). IEEE, pp 434–441
22. Harb H, Makhoul A, Tawil R, Jaber A (2014) Energy-efficient data aggregation and transfer in periodic sensor networks. IET Wirel Sens Syst 4(4):149–158
23. Makhoul A, Harb H, Laiymani D (2015) Residual energy-based adaptive data collection approach for periodic sensor networks. Ad Hoc Netw 35:149–160
24. Jan SRU, Jan MA, Khan R, Ullah H, Alam M, Usman M (2018) An energy-efficient and congestion control data-driven approach for cluster-based sensor network. Mob Netw Appl 24(4):1–11
25. Bhowmik T, Banerjee I, Bhattacharya A (2020) A novel fuzzy based hybrid psogsa algorithm in wsns. In: Proceedings of the 21st international conference on distributed computing and networking, pp 1–5
26. Elmir Y, Khelifi N (2019) Secured biometric identification: hybrid fusion of fingerprint and finger veins. International Journal of Information Technology and Computer Science 11(5):30–39
27. Elappila M, Chinara S, Parhi DR (2018) Survivable path routing in wsn for iot applications. Pervasive Mob Comput 43:49–63
28. Zhang J, Lin Z, Tsai P-W, Xu L (2020) Entropy-driven data aggregation method for energy-efficient wireless sensor networks. Inf Fusion 56:103–113
29. Fliege J, Qi H-D, Xiu N (2019) Euclidean distance matrix optimization for sensor network localization. In, Cooperative Localization and Navigation: Theory, Research and Practice. CRC.
30. Liu D, Mansour H, Boufounos PT et al (2018) Robust sensor localization based on Euclidean distance matrix. In: IGARSS 2018–2018 IEEE international geoscience and remote sensing symposium. IEEE, pp 7998–8001
31. Ullah I, Chen J, Su X, Esposito C, Choi C (2019) Localization and detection of targets in underwater wireless sensor using distance and angle based algorithms. IEEE Access 7:45693–45704
32. Jan SRU, Jan MA, Khan R, Ullah H, Alam M, Usman M (2019) An energy-efficient and congestion control data-driven approach for cluster-based sensor network. Mob Netw Appl 24(4):1295–1305

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.