Wan et al. EURASIP Journal on Wireless Communications and Networking (2019) 2019:59
https://doi.org/10.1186/s13638-019-1374-8

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Similarity-aware data aggregation using fuzzy c-means approach for wireless sensor networks

Runze Wan[1], Naixue Xiong[2*], Qinghui Hu[3], Haijun Wang[1] and Jun Shang[1]

Abstract

For resource-constrained IoT systems, data collection is one of the fundamental operations to reduce the energy dissipation of sensor nodes and improve the network lifetime. However, an anomaly or deviation will exert a great influence on the quality of data collected, especially for a data aggregation scheme. By taking into account data-aware clustering and detection of anomalous events, a similarity-aware data aggregation using a fuzzy c-means approach for wireless sensor networks is proposed. Firstly, by using a fuzzy c-means approach, the clustering process can be performed to organize sensors into clusters based on data similarity. Next, an effective support degree function is defined for further outlier diagnosis. Afterwards, the appropriate weight of valid data can be obtained by taking advantage of the probability distribution characteristics of normal samples within a certain period. Finally, the aggregation result in the cluster can be estimated. Practical database-based simulations have confirmed that the proposed data aggregation method can achieve better performance than traditional methods in terms of data outlier detection accuracy and relative recovery error.

Keywords: Fuzzy c-means, Data similarity, Aggregation, Wireless sensor networks

1 Introduction

Wireless sensor networks (WSNs) are typically composed of many small and low-cost sensor nodes with resource constraints, such as low memory capacity, less computational complexity, low communication bandwidth, and limited power. This new type of network demonstrates the characteristics of low cost, wide distribution, small volume, and flexible self-organizing [1]. With the rapid development, it has been successfully applied in the consumer electronics market and more and more widely used in the fields of target tracking, intelligent transportation, health prognosis, industrial automation, and so on. However, due to the WSN's imperfect nature, the sensor nodes need to be deployed densely to compensate for the quality of data collected [2, 3]. Nonetheless, for process-monitoring applications, high frequent sensing and the transmission of readings result in a large number of redundant samples, which may lead to the waste of the node's energy and bandwidth resource as well as the reduction of the network lifetime. Therefore, how to employ spatiotemporal correlation of the readings between sensor nodes and develop efficient data redundancy reduction for saving the energy of the sensors are urgent problems.

Data aggregation is an effective method to solve the above problems [4]. The basic idea is to aggregate the samples of multi-sensors with a certain degree of redundancy rather than transmit raw data. It means that some nodes will act as aggregator to eliminate redundant data received from other sensor nodes and achieve desirable results for data accuracy. In practical application, the monitoring indicators, such as temperature, humidity, flow rate, or pressure, will demonstrate smooth and steady change in the majority of cases [5]. Once a sudden event occurs, the surrounding sensor nodes are generally able to detect the situation and obtain the readings synchronously. Therefore, the samples with large deviations from individual nodes may have a greater impact on the overall fusion results and influence the quality of data collected [6]. In this paper, we

* Correspondence: xiongnaixue@gmail.com
[2]Department of Mathematics and Computer Science, Northeastern State University, Tahlequah, OK, USA
Full list of author information is available at the end of the article

focus on spatio-temporal correlation of the readings in cluster-based WSNs. In particular, we cope with data-aware clustering and detection of anomalous events, and we use fuzzy c-means approach to organize sensors into clusters based on data similarity.

2 Methods

This study originates from the need for detecting spatial outliers in terms of the spatial correlations among neighboring sensor reading, which can get more accurate fusion results. Our approach uses the spatial temporal correlations of sensor's samples to detect outliers locally.

Compared with previous works, our contributions are presented as follow:

- We propose a novel similarity-aware data aggregation using fuzzy c-means approach for wireless sensor networks.
- We propose a theoretical analysis to determine the optimization of cluster formation.
- We conduct extensive simulations to demonstrate the performance of the algorithms. Simulation results show that our proposed method can achieve better performance than traditional methods in terms of data outlier detection accuracy and relative recovery error.

3 Related work

The traditional methods of data aggregation can be classified into two major categories: random theory-based and artificial intelligence-based approaches [7, 8]. The former includes the weighted average method, least square method, the Bayesian estimation, D-S evidence theory, and so on. The latter uses artificial neural network, fuzzy reasoning, or rough set to eliminate the anomalous data.

Izadi et al. [9] presented a fuzzy-based data fusion approach for WSNs to mitigate redundant data and reduce energy consumption. The authors utilized a fuzzy logic controller to obtain the confidence factor, and then the true value is distinguished and transmitted to the cluster head (CH) for multi-sensor data fusion. Fu [10] proposed double CHs model for secure and accurate data fusion, in which each cluster maintains dual CHs according to the reputation evaluation. All CHs make data fusion and transmit the results to the base station (BS), and the dissimilarity coefficient can be obtained by BS according to the fusion results. If the dissimilarity coefficient exceeds the threshold, the CH will be put into the blacklist and rotate the CH selection immediately. Xiang et al. [11] proposed a data aggregation method based on the compressive sensing theory. Particularly, they adopted diffusion wavelets to make the raw sensor data sparse to decrease the communication overload as well as the computational complexity.

Furthermore, there are several strategies proposed in order to mitigate the energy hole problem. Sun et al. [12] proposed a data aggregation method of wireless sensor networks using artificial neural networks. The data fusion tree is established to reduce the packets flow and can update the leaf nodes dynamically. Aikaraki et al. [13] introduced a joint design of data aggregation with the routing technology, and presented a grid-based routing and aggregator selection scheme to achieve low energy dissipation and low latency without sacrificing quality. By investigating data fusion with communication constraint between the fusion center and each sensor, Xu et al. [14] presented a data fusion mechanism for target tracking in wireless sensor networks based on quantized innovations and Kalman filtering. By adding some delay time, all the data collected by relay node can be fused at one time so as to reduce the energy consumption. Aiming to ensure the data quality, Li et al. [15] proposed various metrics for QoS (quality of service) in the process of data aggregation, including lifetime, data delay, and retransmission rate. Also, the approach is discussed to ensure above QoS metrics in details.

Moreover, data outliers give rise to a very important impact on the correctness of data fusion results and the efficiency of IoT systems. In order to ensure the correctness of fusion results, the data outliers caused by such as software defects, occasionally failed communication, low battery, or malfunction on hardware should be excluded to avoid impact on the aggregation results. Actually, most of the monitoring targets or the occurrence of external events usually will be random and unexpected. With regard to the data outliers from anomalous events, the readings should be identified exactly. Krishnamachari et al. [16] proposed a distributed algorithm for fault-tolerant event region detection in wireless sensor networks, which can determine whether a node is abnormal. Besides, by exploiting the anomaly probability from adjacent nodes, only a few bit messages are sufficient to achieve fault-tolerant localization as events occurred. Tan et al. [17] presented a prediction model of data flow based on linear autoregressive analysis and further proposed a real-time detection algorithm for outliers identification and compression processing. Fernandes et al. [18] propose an autonomous profile-based anomaly detection system using principal component analysis and flow analysis to mitigate the impact of false data injection. By making inference of end-to-end measurements collected by relay nodes, Zheng et al. [19] proposed a trust-assisted framework for detecting and localizing network anomalies in a hierarchical sensor network, which also can obtain a flexible tradeoff between inference accuracy and probing overhead. Hu et al. [20]

presented outlier detection methods based on a neural network for WSNs, which exploited historical data to train the neural network to determine whether the actual measured value into the prediction interval so as to distinguish the data anomalies.

4 Network model and cluster formation

4.1 Network model

We consider a cluster-based architecture for a wireless sensor network, where all sensor nodes can monitor the given condition and periodically send its collected data to its CH. Most researches demonstrate that clustering is considered as an efficient topology control method in WSN to improve the scalability and lifetime of the whole system [21]. By dividing the network, sensor nodes will be grouped into different clusters based on certain rules and each cluster has a cluster head [22]. CH is responsible for managing the cluster and receiving the set of collected data from its member node during a certain period. Also, in order to improve the efficiency of data fusion, CH should have the ability to employ statistical detection based on the sensor readings. It can detect spatial outliers that deviate from normal data, thus ensuring the accuracy of data fusion.

In addition, the following assumptions about the network's topology are suggested:

1. At each period, the sensor nodes acquire the monitoring readings at a fixed sampling rate with m measures.
2. The original attribute information collected from the sensor nodes can be fuzzified into a set of membership functions.
3. In each cluster, the member nodes collect data in a periodic manner. Subsequently, all member nodes will send their data to the appropriate CH for data aggregation at the end of a round.

4.2 Cluster formation algorithm

In this section, we discuss the cluster formation based on data similarity by using fuzzy c-means approach. Compared to other topologies, cluster-based network topology is recently considered to be more effective for aggregating data packets separately. In addition, most of the existing data aggregation techniques based on clustering topology are dedicated to an event-driven data model. Many hierarchical cluster formation algorithms focus on the distance between nodes, residual energy, geographic coverage, and so forth. In contrast, the main purpose of our proposed method is to clear and ameliorate the collected data and provide the best information to end users [23]. From a statistical point of view of the correlation, the perceived data of same time slot can demonstrate spatial-temporal correlation in the adjacent monitoring region. If the monitoring indicators of perceptual physical objects in the region do not show great fluctuation, there will be minimal deviation of the data collected by the sensor nodes with close geographical location [24]. Therefore, cluster formation algorithm can make use of the spatial-temporal correlated environmental data and partition the adjacent sensor nodes with similar data instances into one cluster and different to objects in other groups.

The fuzzy c-means (FCM) algorithm was proposed by Bezdek [25] and has been used in cluster analysis, pattern recognition, image processing, and so forth. FCM is a clustering method derived from unsupervised learning, which uses fuzzy theory to divide a set of data points into a set of fuzzy clusters according to certain partitioning criteria [26]. Suppose a WSN that consist of N-sensor nodes randomly distributed over an area of $S \times S$ meters. By using of the sensor's respective geographical location and collected data initially, the BS computes the cluster centers and allocates sensor nodes to the clusters by applying the FCM algorithm.

Each node is assigned a degree of membership u_{ij} to a cluster C_k rather than completely being a member of other clusters. According to [27], to achieve adequate coverage rate, the optimal number of K clusters should be determined by

$$K = \left\lceil \frac{\ln(1-\delta)}{\ln\left(1-3\sqrt{3}R^2/2S^2\right)} \right\rceil \tag{1}$$

where S represents the side length of square region, R represents the sensor's communication radius, and δ denotes the coverage rate to be assured.

Assuming that sensor node s_i and s_j locate in the same cluster, X_i and X_j represent their collected data sets during a fixed period. x_{ij} denotes the measure generated by the sensor s_i at the time slot j, and m is the number of samples in the fixed period. In the periodic data collection model, in order to minimize data redundancy and still guarantee the accuracy of fusion results, studying the variance between measurements is an analytical way to choose appropriate nodes to form clusters. According to information entropy theory, the entropy value of all samples at time j can be obtained by

$$e_j = -\frac{1}{\ln m}\sum_{i=1}^{m} x_{ij}\ \ln x_{ij} \tag{2}$$

where $0 \le e_j \le 1$.

Since the utility value of the index is proportional to its impact on fusion results, the weight value of variable j can be defined as

$$\omega_j = \frac{1-e_j}{\sum_{j=1}^{m}(1-e_j)} \tag{3}$$

Next, the weighted Euclidean distance between the reading of node s_i and centroid point v_k of cluster C_k can be expressed as

$$\mathrm{Dis}(s_i, v_k) = \sqrt{\sum_{j=1}^{m} \omega_j \left(x_{ij} - v_{kj}\right)^2} \tag{4}$$

where v_{kj} indicates the reference value of the centroid point at time j in cluster C_k.

Next, the objective function should be proposed to enhance the quality of the clusters and allocate sensor nodes into their most appropriate one [28]. By using FCM, the objective function, which will operate with iterative procedures, can be formulated as follows:

$$\min J(X, U, C_1, C_2, \cdots, C_K) = \sum_{k=1}^{K}\sum_{i=1}^{N} u_{ki}^{v_0}[\mathrm{Dis}(s_i, v_k)]^2 \tag{5}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} u_{ki} = 1, 1 \le i \le N; \quad u_{ki} \in [0,1], \ 1 \le k \le K, \ 1 \le i \le N; \quad \sum_{i=1}^{N} u_{ki} \in [0, N], \ 1 \le k \le K.$$

where u_{ki} denotes the degree of membership being assigned to node s_i to join in the cluster C_k. U denotes the membership matrix of u_{ki}. v_0 is a weighting exponent on each fuzzy membership that determines the amount of fuzziness of the resulting classification, and is set to 2.

By using the Lagrange's multiplier to optimize formula (18), the problem is equivalent to find the minimum value of the Eq. (6).

$$F(U, \lambda) = \sum_{k=1}^{K}\sum_{i=1}^{N} w_i u_{ki}^{v_0}[Dis(s_i, v_k)]^2 + \sum_{i=1}^{N} \lambda \left(\sum_{k=1}^{K} u_{ki} - 1 \right) \tag{6}$$

where w_i indicates the weight value in process of data aggregation. It can be set as $1/N$ initially, and be updated by formula (22) at the end of each sampling period.

Note that the first order partial derivative should be equal to 0, i.e., $\partial F/\partial \lambda = 0$ and $\partial F/\partial u_{ki} = 0$. Then, we have

$$\partial F/\partial \lambda = 1 - \sum_{k=1}^{K} u_{ki} = 0 \tag{7}$$

$$u_{ki} = \frac{\lambda}{w_i[\mathrm{Dis}(s_i, v_k)]^2} \tag{8}$$

Thus, u_{ki} can be determined as

$$u_{\mathrm{ki}} = \frac{w_{\mathrm{i}}[Dis(s_{\mathrm{i}}, v_{\mathrm{k}})]^2}{\sum_{k=1}^{K} w_{\mathrm{i}}[Dis(s_{\mathrm{i}}, v_{\mathrm{k}})]^2} \tag{9}$$

Similarly, suppose that $\partial F/\partial v_k = 0$, the reference vector of the centroid point in cluster C_k can be given as

$$v_k = \frac{\sum_{i=1}^{N} w_i u_{ki}^2 X_i}{\sum_{i=1}^{N} w_i u_{ki}^2} \tag{10}$$

5 Data aggregation

5.1 Data outlier detection

Outliers are often known as anomaly or deviation, which can even mislead systems into unsafe conditions. Whether the quality of data collected by WSNs is reliable and accurate or not will influence the performance of the whole system [29]. Therefore, data outliers should be detected and isolated in time so as to ensure the validity of data aggregation result and fusion efficiency. For clustered WSN, it is impossible for CH to determine the validity of the data sent by its members. However, the geographical relationship between the readings of sensor nodes within a certain physical spatial range or cluster may be an effective means to identify outliers through credible tests of the masses. In this sub-section, an effective support degree function is defined for further outlier diagnosis is introduced, which is based on a standard statistical distribution model and makes use of the measures between neighboring nodes.

As mentioned above, due to the spatial-temporal correlation in the adjacent monitoring region, the measurements between node s_i and s_j at the same sampling period will show relatively small differences. Hence, the support degree from node s_j to s_i can be expressed as consistency between the samples X_i and X_j.

First, to eliminate the influence of the measurement scale, the normalization processing of raw data is introduced to make a relatively objective comparison between measurement sets collected by different sensor node. The original attribute of raw data may belong to positive index or negative index. First, linear transformation of raw data can be given by

$$\begin{cases} x'_{ij} = x_{ij}/\hat{x}_j, & \text{if } x_j \text{ is positive index} \\ x'_{ij} = \hat{x}_j/x_{ij}, & \text{if } x_j \text{ is negative index} \end{cases} \tag{11}$$

where $\hat{x}_j$ denotes the ideal value for the samples at the time slot j, and x'_{ij} is the proximity of x_{ij} to ideal values.

Then, the normalized value of sample x_{ij} can be represented as

$$y_{ij} = x'_{ij} / \sum_{i=1}^{n} x'_{ij} \tag{12}$$

In order to characterize that mutual support relationship between sensor nodes, the support degree can be defined as

$$p_{ij} = \frac{\sum_{k=1}^{m} \left(y_{ik} - \hat{y}_k\right)\left(y_{jk} - \hat{y}_k\right)}{\sqrt{\sum_{k=1}^{m} \left(y_{ik} - \hat{y}_k\right)^2 \sum_{k=1}^{m} \left(y_{jk} - \hat{y}_k\right)^2}} \tag{13}$$

where $\hat{y}_k$ represents the mean value.

According to the above function, the support degree matrix P of the measures from all nodes in a cluster can be obtained

$$P = \begin{bmatrix} 1 & p_{12} & \cdots & p_{1n} \\ p_{21} & 1 & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & 1 \end{bmatrix} \tag{14}$$

where p_{i1}, p_{i2}, $\cdots$, p_{in} are support degree of the samples collected by s_i from the member nodes in the same cluster, and n indicates the number of member nodes. By integrating the evaluation of all neighboring nodes, the comprehensive support degree of X_i can be calculated as

$$T_i = \frac{1}{n} \sum_{j=1}^{n} p_{ij} \tag{15}$$

Since T_i is relative to the amount of support degree from other member nodes or its nearest local neighbors, it indicates normal level compared to the majority of sensor readings. When a sensor sends abnormal data due to noise errors or malicious attacks, the readings will obviously deviate from the measures of other sensors. As a result, its comprehensive support is very small. Unless a large area of intra-cluster nodes fail simultaneously, the probability of that exceptional case will be very low and can be neglected.

Suppose that the comprehensive support of data X_i is T_i, if $T_i \geq \zeta$, X_i is determined as normal data. Otherwise, the data will be regarded as outliers. Among them, the parameter ζ is set as the availability threshold value. When the value of T_i is less than the threshold value ζ, the corresponding readings X_i will be processed to mitigate the influence on the aggregation result.

5.2 Data aggregation strategy

In this section, we present the data aggregation strategy to ensure the accuracy of the aggregation result. Before aggregation process, the data being collected from member nodes will be sent entirely to the cluster head, which can conduct outlier detection based on the centralized approach. If the data being received is valid, it will be put into data aggregation process. Otherwise, they should be rejected immediately. Therefore, the two types of memory buffers can be embedded in CH and corresponding parameter a and b is set to count the number of normal and outlier data uploaded by each member node. Under certain conditions, the probability distribution of normal and outlier data will be approximate to the posterior probability distribution with binomial model, which can obey the beta distribution. Therefore, the beta distribution characteristics can be employed to evaluate the data validity.

Set χ as the posterior probability of a random event, the probability of distribution can be obtained based on the Bayesian statistics as

$$\begin{aligned} P(\chi|a,b) &= \frac{P(\chi,a,b)}{P(a,b)} \\ &= \frac{\binom{a+b}{a}\chi^a(1-\chi)^b}{\int_0^1 \binom{a+b}{a}\chi^a(1-\chi)^b d\chi} \\ &= \frac{\chi^a(1-\chi)^b}{\int_0^1 \chi^a(1-\chi)^b d\chi} \end{aligned} \tag{16}$$

where $0 \leq \chi \leq 1$, $a > 0$, $b > 0$.

The probability density function of the parameter (a, b) can be expressed as

$$f(\chi|a,b) = \frac{\chi^{a-1}(1-\chi)^{b-1}}{\int_0^1 u^{a-1}(1-u)^{b-1}du} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\chi^{a-1}(1-\chi)^{b-1} \tag{17}$$

Thus, we have

$$P(\chi|a,b) = \frac{\chi^a(1-\chi)^b}{f(a+1,b+1)} \tag{18}$$

As a result, the reliability of the monitored samples of the determined node will obey the beta distribution with parameters $a + 1$ and $b + 1$.

$$P(\chi|a+1,b+1) = \begin{cases} \dfrac{\chi^a(1-\chi)^b}{B(a+1,b+1)}, & 0 < \chi < 1 \\ 0, \text{ otherwise} \end{cases} \tag{19}$$

Therefore, the mathematical expectation $E(\chi)$ of beta distribution can be given as

$$E(\chi) = \frac{a+1}{a+b+2} \tag{20}$$

Considering the effect of data retransmission caused by channel quality, the uncertainty of the readings being

collected by the determined node can be attenuated by introducing time attenuation factor. Accordingly, the time attenuation factor should be defined as

$$\theta = Td_i / Td_{\text{ave}} \tag{21}$$

where Td_i denotes the packet transmission delay of node s_i, and Td_{ave} denotes the average packet transmission delay of all member nodes.

Thus, formula (8) can be modified to:

$$E(\chi) = \frac{a+1}{a+\theta b+2} \tag{22}$$

The revised mathematical expectation can be defined as the weight value in process of data aggregation, and $w_i = E_i(\chi)$ will be allocated to member nodes.

Finally, the aggregation result in the cluster can be estimated by

$$X^* = \sum_{i=1}^{n} w_i X_i \tag{23}$$

6 Experiments

In this section, practical database-based simulations have been conducted to evaluate the performance of our method. Firstly, the datasets are derived from the real sensed data collected from 54 Mica2Dot sensors deployed in the Intel Berkeley Research Lab between February 28 and April 5, 2004 [30]. The sensed data included humidity, temperature, light, and voltage values collected. In the experiments, we first selected some measurements of temperature from the sensor nodes 36 until 43, for the time period from March 18, 2004, to March 20, 2004, corresponding to 2000 log rows. We do not take into account the other features (humidity, light, and voltage). The quantity of data is about 2.3 million readings; it was collected using the TinyDB in-network query processing system, built on the TinyOS platform. Based on the dataset, we add a given mass of outliers to simulate the occurrence of events, which can make the data fluctuate to a certain extent.

To evaluate the performance of our approach, we use the TOSSIM tool [31]. TOSSIM is a TinyOS simulation tool which simulates WSN physical and link layer features accurately. This allows validating the solution under realistic WSN deployment conditions. In each test, we repeat the simulation for 30 times and compute the mean of results. The key simulation parameters are summarized in Table 1.

In the experiment scenarios, outliers are simulated randomly, and 100 values of temperature are generated and then added to the dataset. In terms of the evaluation metrics, detection accuracy rate is defined as the ratio of outliers being detected to all outliers, and false alarm rate represents the ratio of normal data mistakenly detected as outliers.

Table 1 Simulation parameters

Parameter	Value
Area size	500 × 500 m
Number of sensor nodes N	81
Node's communication range	40 m
Transmission channel	Wireless channel
Propagation model	Normal path loss model
Data packet size	32 bytes
Bandwidth	200 kilobytes per second
Radio layer	CC2420 radio layer
Queue size	50 packets
δ	0.6~0.9
ζ	0.1~0.9
Outlier probability	5%, 10%, 15%
υ_0	2
m	50

First, our objective is to study the detection of abnormal data in accordance to our proposed method. Figure 1 shows the detection accuracy rate when varying δ and ζ. Since the support degree can differentiate the normal data and outliers, it can effectively guarantee the data accuracy when handling the aggregation process. This truth is clearly shown in Fig. 1 when the sensor node applies the aggregation phase and when δ and ζ increases. In addition, it can be noticed that the detection accuracy rate can be significantly improved as ζ increases. The reason is that once the node's support degree cannot be satisfied with the threshold's requirement, its readings will not be preserved and submitted to CH for data fusion. It is beneficial to the accuracy of the final fusion results.

Figure 2 shows the false alarm rate when varying δ and ζ. Indeed, we can find that the overall trend of false alarm rate is opposite to that of the detection accuracy rate, especially when δ is larger, the fluctuation is more obvious. The higher value of δ means that more clusters will be distributed in the monitoring region. For fixed number of total sensor nodes, it will cause the reduction of the number of members in a single cluster and make the determination of outliers more stringent. In addition, the increase of threshold ζ leads to higher level for support degree, and the samples with large deviation can easily be judged to be invalid. In the case of a monitoring indicator that suddenly changes dramatically and only a few nodes perceived it, their readings will be treated as anomalous data owing to low support degree. Thus, the detection accuracy rate will be increased.

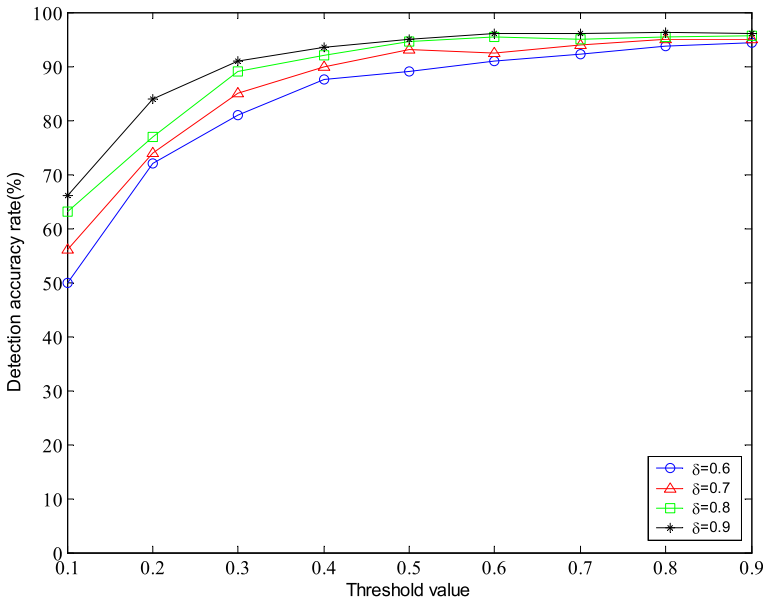

Fig. 1 Detection accuracy rate. The detection accuracy rate reflects the ratio of outliers being detected to all outliers

Figure 3 shows the deviation degree of aggregation result as a function of threshold ζ. Based on the obtained results, we can notice that ζ moves from the extreme values of 0.1 and 0.9 towards the optimal value 0.6. When the threshold is either too low or too high, the final fusion result is not ideal. This is due to the high accuracy and low false alarm rate of differentiating normal data and outliers in terms of appropriate availability threshold. Low threshold setting may result in low detection rate and thus affecting the accuracy of the final fusion result. Conversely, excessive value of specified threshold will make the criteria more rigorous, and the normal measurement will be identified as outlier. Moreover, if such a situation occurs in a continuous time, it will inevitably lead to a sharp decline in the weight of these valid data in data fusion processing and thus affecting the data accuracy.

Next, we further study the impact of different outlier probability on relative recovery error and make a comparison with KPFF [32] and DSADC [33]. Recovery accuracy

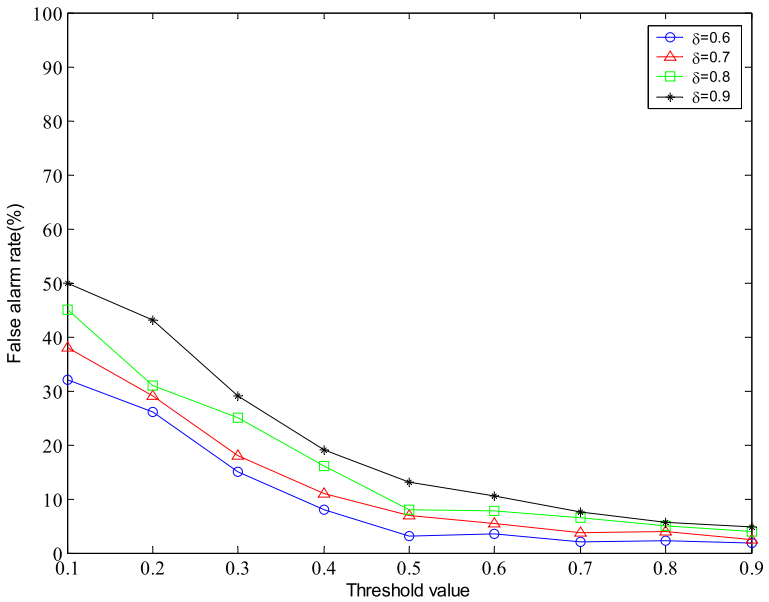

Fig. 2 False alarm rate. The false alarm rate demonstrates the ratio of normal data mistakenly detected as outliers

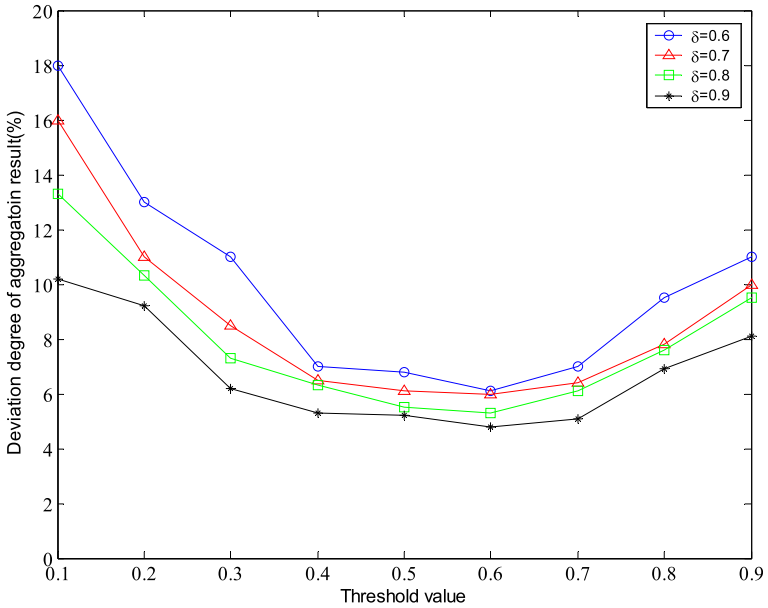

Fig. 3 Deviation degree of aggregation result. Appropriate threshold can be beneficial to detect outliers locally based on the spatial temporal correlations of the sensor's samples

is a normally used metric to evaluate the quality of data aggregation algorithms [34]. In this paper, recovery accuracy is mathematically defined by relative recovery error (RRE), which is the relative difference between original and recovered data matrices. Figures 4, 5 and 6 demonstrate the instant RRE along the timeline.

The obtained results show clearly that applied support degree in such a way is very effective. It also maintains adaptability with different outlier probability. From the experiment results, it can be seen that the RRE curves of both KPFF and DSADC algorithms fluctuate dramatically. But we can still observe that nearly 90% RRE values of similarity-aware data aggregation using fuzzy c-means approach (SDAF) are below those of DSADC. The error of the fusion results obtained by SDAF is smaller than other methods especially as outlier probability increases. In the process of data aggregation, outlier samples can be identified effectively by diagnosis mechanism in

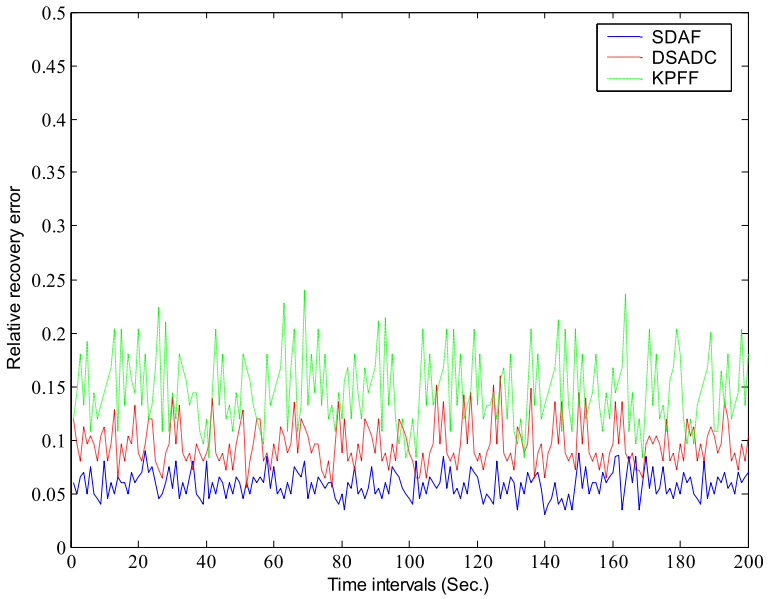

Fig. 4 RRE with outlier probability (5%). Recovery accuracy is a normally used metric to evaluate the quality of data aggregation algorithms

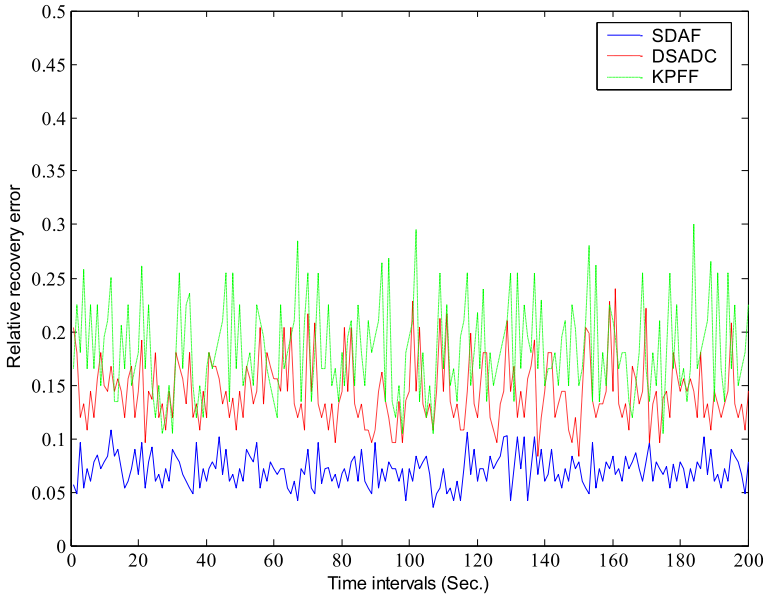

Fig. 5 RRE with outlier probability (10%). RRE can reflect the effectiveness to the aggregation result by data outliers

SDAF, and the outlier-free readings are further aggregated and transmitted to the CH. Therefore, it can reduce the effectiveness to the aggregation result by data outliers and avoid the possibility of misleading systems into unsafe conditions.

7 Conclusions

To minimize the energy consumption by redundant data and reduce the expense of transmissions to the sink, data aggregation technology is very essential for WSNs. Data anomaly or deviation will exert a great influence on the quality of aggregated results. In this paper, we have proposed a similarity-aware data aggregation using a fuzzy c-means approach in clustered WSNs. By investigating the spatio-temporal correlations of sensor data and local detection of anomalous events, we presented a cluster formation algorithm based on fuzzy c-means approach. Then, we define an effective support degree function for further outlier diagnosis. Finally, based on statistical analysis of the outlier or outlier-free sensor

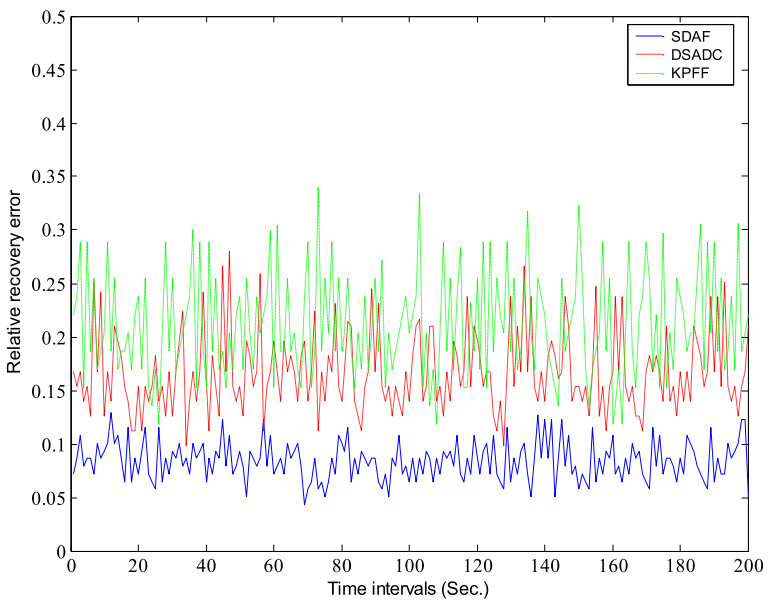

Fig. 6 RRE with outlier probability (15%). The results demonstrate adaptability with different outlier probability

data, the readings aggregation is conducted. Overall, the simulation results show that the proposed method can achieve better performance than traditional methods in terms of data outlier detection accuracy and relative recovery error.

In our future work, we plan to conduct the research on the analysis of outlier detection in terms of characteristics like the multi-dimension, detection mode, architectural structure, and correlation extraction.

Abbreviations

BS: Base station; CH: Cluster head; FCM: Fuzzy c-means; QoS: Quality of service; RRE: Relative recovery error; SDAF: Similarity-aware data aggregation using fuzzy c-means approach; WSNs: Wireless sensor networks

Acknowledgements

The authors acknowledged the anonymous reviewers and editors for their efforts in valuable comments and suggestions.

Funding

This research was supported in part by the Hubei Provincial Educational Science Program (Grant No. 2018GB073) and the Guangxi Nature Science Fund (Grant No. 2016GXNSFAA380226).

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Authors' contributions

WR proposes the innovation ideas and theoretical analysis, and XN carries out experiments and data analysis. HQ also wrote parts of the manuscript. SJ and WH participated in the coordination of the study and reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Hubei Co-Innovation Center of Information Technology Service for Elementary Education, Hubei University of Education, Wuhan, China. [2]Department of Mathematics and Computer Science, Northeastern State University, Tahlequah, OK, USA. [3]School of Computer Science & Engineering, Guilin University of Aerospace Technology, Guilin, China.

Received: 13 September 2018 Accepted: 21 February 2019
Published online: 12 March 2019

References

1. P. Zhong, Y.T. Li, W.R. Liu, Joint mobile data collection and wireless energy transfer in wireless rechargeable sensor networks. Sensors. **17**(8), 1–23 (2017).
2. Y.Y. Zeng, C.J. Sreenan, L. Sitanayah, An emergency-adaptive routing scheme for wireless sensor networks for building fire hazard monitoring. Sensors. **11**(3), 2899–2919 (2011).
3. C. Lin, N. Xiong, J.H. Park, T. Kim, Dynamic power management in new architecture of wireless sensor networks. Int. J. Commun. Syst. **22**(6), 671–693 (2009).
4. H.J. Cheng, Y.Z. Chen, N.X. Xiong, Layer-based data aggregation and performance analysis in wireless sensor networks. J. Applied Mathematics. **2013**, 1–12 (2013).
5. C. Zhou, S. Huang, N. Xiong, S.H. Yang, Design and analysis of multimodel-based anomaly intrusion detection systems in industrial process automation. IEEE Transactions on Systems, Man, and Cybernetics: Systems. **45**(10), 1345–1360 (2015).
6. L. Shu, Y. Zhang, Z. Yu, L.T. Yang, M. Hauswirth, Context-aware cross-layer optimized video streaming in wireless multimedia sensor networks. J. Supercomput. **54**(1), 94–121 (2010).
7. H.J. Cheng, D.Y. Feng, X.B. Shi, Data quality analysis and cleaning strategy for wireless sensor networks. EURASIP J. Wirel. Commun. Netw. **61**, 1–11 (2018).
8. K. Lin, J.P.C. Rodrigues, N. Xiong, H.W. Ge, Energy efficiency QoS assurance routing in wireless multimedia sensor networks. IEEE System Journal. **5**(4), 495–505 (2011).
9. I. Davood, J.H. Abawajy, S. Ghanavati, A data fusion method in wireless sensor networks. Sensors. **15**(2), 2964–2979 (2015).
10. J.S. Fu, Y. Liu, Double cluster heads model for secure and accurate data fusion in wireless sensor networks. Sensors. **15**(1), 2021–2040 (2015).
11. L. Xiang, J. Luo, C. Rosenberg, Compressed data aggregation: energy-efficient and high fidelity data collection. IEEE/ACM Trans. Netw. **21**(6), 1722–1735 (2013).
12. L.Y. Sun, X.X. Huang, W. Cai, Data aggregation of wireless sensor networks using artificial neural networks. Chinese Journal of Sensors and Actuators. **24**(1), 122–127 (2011).
13. J.N. Aikaraki, R. Uimustafa, A.E. Kamal, Data aggregation and routing in wireless sensor networks: optimal and heuristic algorithms. Comput. Netw. **53**(7), 945–960 (2009).
14. J. Xu, J.X. Li, S. Xu, Data fusion for target tracking in wireless sensor networks using quantized innovations and Kalman filtering. Science China: Information science edition. **55**(3), 530–544 (2012).
15. H. Li, H.Y. Yu, Research on data aggregation supporting QoS in wireless sensor networks. Application Research of Computers. **25**(1), 64–67 (2008).
16. B. Krishnamachari, S. Iyengar, Distributed Bayesian algorithms for fault-tolerant event region detection in wireless sensor networks. IEEE Trans. Comput. **53**(3), 241–250 (2004).
17. Y.H. Tan, Y.P. Lin, T. Dong, Real-time detection algorithm for anomaly data in sensor networks. Journal of System Simulation. **19**(18), 4335–4341 (2007).
18. G.J. Fernandes, J.P.C. Rodrigues, M.L. Proença, Autonomous profile-based anomaly detection system using principal component analysis and flow analysis. Appl. Soft Comput. **34**(9), 513–525 (2015).
19. S. Zheng, J.S. Baras, in *8th IEEE Communications Society Conference on Sensor, Mesh and ad hoc Communications and Networks(SECON)*. Trust-assisted anomaly detection and localization in wireless sensor networks (2011), pp. 386–394.
20. S. Hu, G.H. Li, W.W. Lu, Outlier detection methods based on neural network in wireless sensor networks. Computer Science. **41**(11), 208–211 (2014).
21. R. Kumar, N. Singh, A survey on data aggregation and clustering schemes in underwater sensor networks. Int. J. Grid Distrib. Comput. **7**(6), 29–52 (2014).
22. W. Guo, N. Xiong, A.V. Vasilakos, G. Chen, H. Cheng, Multi-source temporal data aggregation in wireless sensor networks. Wirel. Pers. Commun. **56**(3), 359–370 (2011).
23. V. Chatzigiannakis, S. Papavassiliou, Diagnosing anomalies and identifying faulty nodes in sensor networks. IEEE Sensors J. **7**(5), 637–645 (2007).
24. X. Wang, Q. Li, N. Xiong, Y. Pan, in *International Conference on Wireless Algorithms, Systems, and Applications (WASA 2018)*. Ant colony optimization-based location-aware routing for wireless sensor networks (2018), pp. 109–120.
25. F. Herrera, Genetic fuzzy systems: status, critical considerations and future directions. Int. J. Comput. Intell. Res. **5**, 59–67 (2005).
26. N. Goyal, M. Dave, A.K. Verma, in *Int. Conf. Electron. Commun. Syst. (ICECS)*. Fuzzy based clustering and aggregation technique for under water wireless sensor networks (2014), pp. 1–5.
27. W.B. Heinzelman, A.P. Chandrakasan, H. Balakrishnan, An application-specific protocol architecture for wireless micro-sensor networks. IEEE Trans. Wirel. Commun. **1**(4), 660–670 (2009).
28. J.S. Lee, W.L. Cheng, Fuzzy-logic-based clustering approach for wireless sensor networks using energy predication. IEEE Sensors J. **12**, 2891–2897 (2012).
29. S.M. Reda, M. Abdelhamid, S. Hadj, A. Amar, Performance evaluation of network lifetime spatial–temporal distribution for WSN routing protocols. J. Netw. Comput. Appl. **35**(4), 1317–1328 (2012).
30. Intel lab data home page. http://db.lcs.mit.edu/labdata/labdata.html. March 20, 2014.
31. P. Levis, N. Lee, M. Welsh, D. Culler, in *Proc. of the 1st International Conference on Embedded Networked Sensor Systems*. TOSSIM: accurate and scalable simulation of entire TinyOS applications (ACM Digital Library, Los Angeles, California, 2003), pp. 126–137.

32. H. Harb, A. Makhoul, D. Laiymani, in *Proc. of the 10th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. K-means based clustering approach for data aggregation in periodic sensor networks (2014), pp. 434–441.
33. G. Fernando, J. Gentian, N. Michele, S. Aldri, Data similarity aware dynamic node clustering in wireless sensor networks. Ad Hoc Networks. **24**, 29–45 (2015).
34. Y. Sang, H. Shen, Y. Tan, N. Xiong, in *Proc. of International Conference on Information and Communications Security*. Efficient protocols for privacy preserving matching against distributed datasets (2006), pp. 210–227.