Chu et al. EURASIP Journal on Wireless Communications and Networking (2020) 2020:52
https://doi.org/10.1186/s13638-020-01671-y

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Identifying correctness data scheme for aggregating data in cluster heads of wireless sensor network based on naive Bayes classification

Shu-Chuan Chu[1], Thi-Kien Dao[2*], Jeng-Shyang Pan[1,2] and Trong-The Nguyen[2,3]

Abstract

Wireless sensor network (WSN) has been paid more attention by scholars due to the practical communication of a system of devices to transfer information gathered from a monitored field through wireless links. Precise and accurate data of aggregating messages from sensor nodes is a vital demand for a success WSN application. This paper proposes a new scheme of identifying the correctness data scheme for aggregating data in cluster heads in hierarchical WSN based on naive Bayes classification. The collecting environmental information includes temperature, humidity, sound, and pollution levels, from sensor nodes to cluster heads that classify data fault and aggregate and transfer them to the base station. The collecting data is classified based on the classifier to aggregate in the cluster head of WSN. Compared with some existing methods, the proposed method offers an effective way of forwarding the correct data in WSN applications.

Keyword: Wireless sensor network, Naive Bayes, Fault detection, Classification

1 Introduction

Wireless sensor network (WSN) refers to a set of spatially dispersed and dedicated sensors for recording and monitoring the physical conditions and organizing the data collected at a central location [1, 2]. WSNs are a low-cost network that is widely used in numerous applications [3–5]. The needed environmental information like temperature, light, sound, humidity, wind, air, and water pollution levels could be captured and measured by sensor nodes of WSN [6–9]. A good designed and employed WSN often follows the clustering fashion as efficient ways of saving energy networks that is to generate clusters by arranging the sensor nodes into groups [10]. The power consumption of WSNs is affected directly by the clustering criterion problem [11]. The cluster composes node members (NM) and cluster head (CH). CH is selected from among NM. The functions of CHs are not only to collect the information from the NM but also to aggregate captured data that forwarded to the BS [12, 13]. Figure 1 shows an example of the clustering fashion of WSN.

Clustering provides various advantages like energy efficiency, prolonging lifetime, scalability, and less delay. However, the clustered WSNs also have got the drawback such as the aggregated data fault problem at CHs that caused the network reliability of the monitoring and predicting applications decreased [10, 14].

This paper considers the correct data to transfer from CHs to the base station (BS) of WSN. The decision function of the classifier is deployed in the CH for aggregating accuracy data. The cluster heads (CHs) in hierarchical WSN aggregate the captured data by MNs, then CHs send them to the BS or via hops (via other CHs). The method of classification is used to detect the faults based on the data learning model in making decisions by combining expert knowledge and statistical learning method. The accuracy of captured data has an essential role in successful ones for several applications such as

* Correspondence: 1101405123@nkust.edu.tw
[2]Fujian Provincial Key Lab of Big Data Mining and Applications, Fujian University of Technology, Fuzhou 350118, China
Full list of author information is available at the end of the article

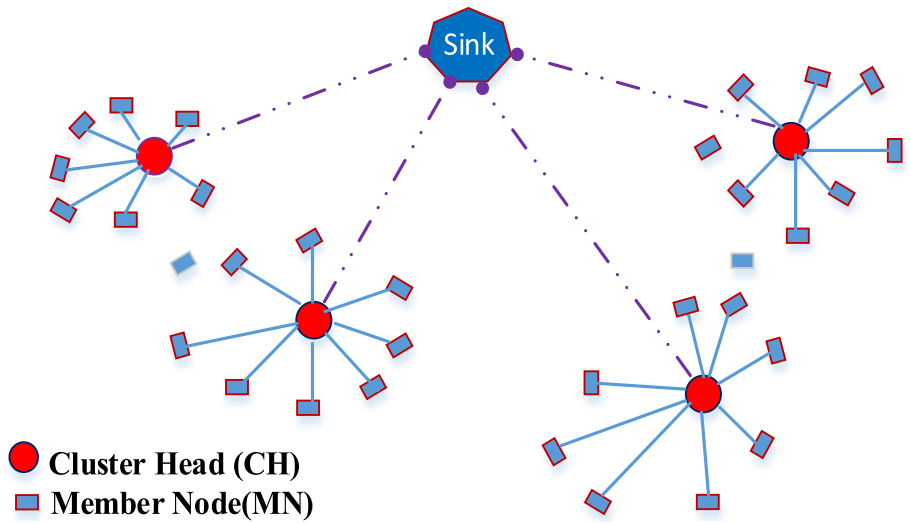


Fig. 1 A clustering scheme of a wireless sensor network. The clustering in WSN composes Node members (NM) and Cluster head (CH). CH is selected from among NM. The functions of CHs are not only to collect the information from the NM but also to aggregate captured data that forwarded to the BS. Clustering provides various advantages like energy efficiency, prolonging lifetime, scalability, and less delay. However, the clustered WSNs also have got the drawback such as the aggregated data fault problem at CHs that caused the network reliability of the monitoring, and predicting applications decreased.

weather prediction, military monitoring, traffic monitoring, seismic activity prediction, and healthcare monitoring [6, 15, 16].

The faults of WSN occur commonly due to its characteristic that is a network through wireless links of devices on ubiquitous and deployed in uncertain and hazardous areas, e.g., battlefields, forest, healthcare, volcanos, highways. Thus, how to get reliable data to transfer from CH to the BS for further processing will be an urgent requirement to guarantee proper functioning applications [17]. The distinction between normal and faulty data must be determined correctly. The detection of fault data also should be rapid and precise. Identifying the data fault which occurs spontaneously is difficult as those faults may cause continuous failures [18]. The faults will cause the WSN application to increase data network traffic and wastes battery power [19]. Data fault detection is a promising way to enhance bandwidth, integrity, and reliability. The applied classification is one of the favors of solutions for identifying faults in WSN [20].

In this paper, the decision function of the classification technique is deployed in CHs of the hierarchical WSN to classify sensing data from MNs and to detect its faults for the next steps of processing, e.g., aggregating data. The CHs should aggregate data of the sensed environmental information from sensor nodes and transfer them to the base station. Naive Bayes classifier is applied to detect data fault in CHs to enhance the integrity and reliability of the WSN application.

The rest of this paper is organized as follows: Section 2 describes the common types of faulty in WSN and related work. Section 3 introduces the proposed method. In Section 4, several experiments are carried on the scenario to evaluate the performance of the proposed method. Finally, the conclusion is discussed in Section 5.

2 Related works

2.1 Fault data issues in WSNs

As the WSN is a network through wireless links of devices on ubiquitous that are deployed in uncertain and hazardous areas, e.g., battlefields, forest, healthcare, volcanos, highways [21–23]. The electronic components in the sensor node are also easy to break down. Some models, such as centralized, distributed, and hybrid fault detection, have been introduced for solving the WSN failures [24]. The frequent failures happen of WSNs are classified into some types, such as hardware failure, software failure, and communication failure [25]. Figure 2 depicts some types of common faults that occur in WSNs.

Hardware failures occur due to the negligence of sensing capability, power (battery failure) location, and processing units of sensors—for example, the battery failure causes the impairment of sensors; software failures, e.g., the fusion and aggregation that occurs due to problems in sensor programs; and communication failures, e.g., a transceiver that disrupts the sending and receiving data from the sensors. Data faults might occur either separately or simultaneously together and also might happen over a while or instantly. Defects in WSN also can be categorized based on two aspects according to the time of the error and location of the fault. The time span of the failure indicates the duration of the fault. The

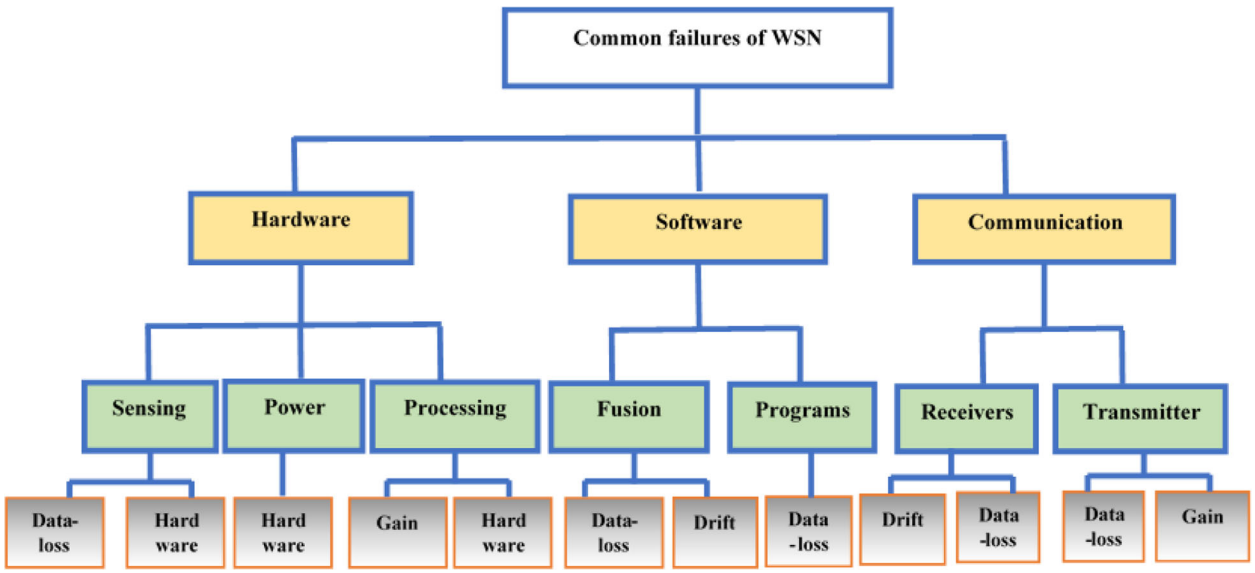


Fig. 2 The common faults occur in wireless sensor network

location of the fault suggests the environment where the fault occurs [26]. The period the defects can be categorized into persistent faults and transient faults. Persistent errors are permanent faults that can be solved when the system recovery made. Transient faults are temporary faults that occur due to network congestion or climatic condition. The location of responsibility is broadly categorized into data-centric and system-centric. Data-centric failure consists of the offset fault, gain fault, drift fault, data loss fault, hard over fault, spike fault, and fusion fault. System-centric fault includes the calibration fault, battery failure, and hardware failure that can cause the malfunctioning of the entire network. The classification is one of the effective solutions for identifying faults in WSN [26–28].

2.2 Classification methods

Definition of classification is a method of identifying new data belonging to which of a set of data (subset), based on a training set of data containing observations (or instances) whose category membership is known [29]. The most popular methods of classification include supervised learning, unsupervised learning, and semi-supervised learning [30]. Supervised learning is a paradigm that is trained on a predefined set of training examples using which the accurate result can be obtained when the new data is given. The paradigm works based on labeled data. Unsupervised learning is a paradigm that gives a bunch of data from which the pattern has to be obtained [31]. The method executes based on unlabeled data. Semi-supervised learning is a paradigm that its knowledge uses both labeled and unlabeled data [32].

The different affecting factors of sensing data from various sensor nodes in WSN, such as uncertain and hazardous areas of battlefields, forest, volcanos, or highways, are aggregated to model for classification. There are often two phases in the processing of classification. The aspect of learning data allows combining knowledge expertise with the statistical method. The different attributes affecting valid data in WSNs can be verified based on experiences knowledge of the experts. The applied pattern with the probability in the learning data based on the expertise for decision function is carried out to clarify the new capturing data [33].

Figure 3 shows a classification model from the data in the learning methods. Data is a set of data capture from the sensor nodes, i.e., the attributes of sending raw data (also called examples, instances, or cases) described by k attributes: A_1, A_2, ...A_k. Each attribute is considered as a class that is labeled with a predefined course. The goal is to learn a classification model from the data that can be applied to predict the classes of new data for aggregating and forwarding to the base station of WSN.

The essential contributions introduced recently by the community research engaging in fault detection in WSNs are reviewed in this subsection. A method of dealing with detecting failures in WSNs has been developed based on a probabilistic scheme that knows as naive Bayes classifier [34]. This method worked out with sensing data as well as remaining energy in the network. Bayesian classification represented a supervised learning method that provides practical learning algorithms, and prior knowledge and observed data could be combined. It assumes an underlying probabilistic model, and it allows us to capture uncertainty about the model in a

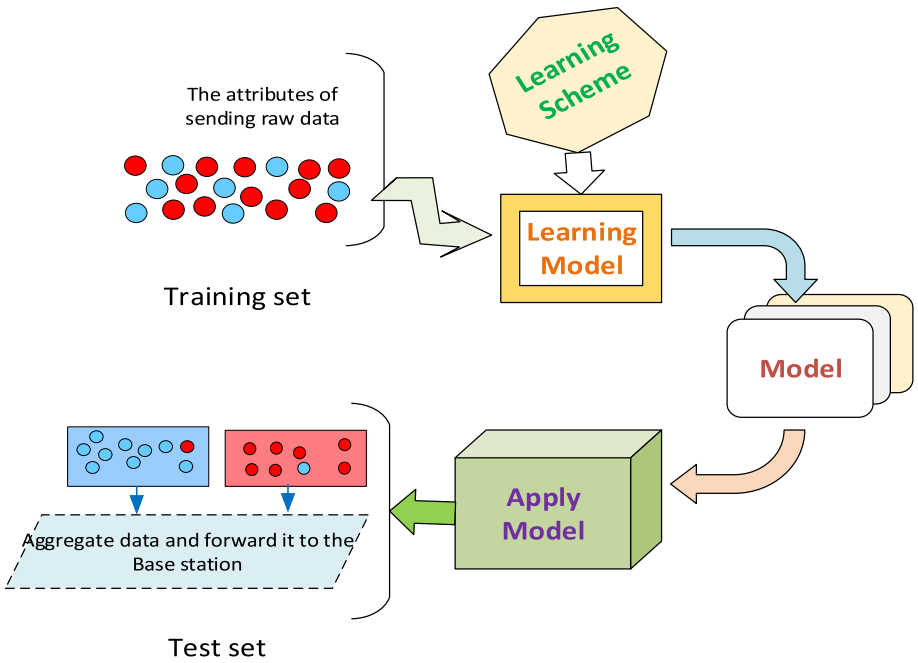


Fig. 3 A classification model from the data in the learning method

principled way by determining the probabilities of the outcomes. It can solve diagnostic and predictive problems. The Bayesian classification [35] provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis, and it is robust to noise in input data. The probability of the hypothesis is formulated as follows.

$$p(h|d) = \frac{P(d|h) \times P(h)}{P(d)} \tag{1}$$

where h and d are hypothesis and data vectors respectively; $P(h)$ is a probability of hypothesis as prior belief; $P(d|\ h)$ is a probability of the data if the hypothesis h is true; the marginal probability of data $P(d) = \sum_h P(d|h)P(h)$ is data evidence; and $p(h|\ d)$ is posterior.

Moreover, the decision tree learning algorithm is a greedy divide-and-conquer algorithm. Fault detection in WSN has introduced through decision trees (DT) [36]. Assume attributes are categorical now (continuous traits also can be handled). The tree is constructed in a top-down recursive manner. At the start, all the training examples are at the root. Instances are partitioned recursively based on selected attributes. Attributes are selected based on an impurity function (e.g., information gain). Conditions for stopping partitioning consist of one of the conditions as follows. All attributes for a given node belong to the same class. There are no remaining attributes for further partitioning, i.e., the majority class is the leaf, and there are no examples left.

Additionally, a method of fault detection in WSN based on the support vectors machine (SVM) with the Gaussian kernel is deployed for real-time data classification [26]. The SVM is a classifier that is the supervised learning model to identify the optimal boundary separating data of two classes. The Lagrangian coefficients are determined, and support vectors along with decision function are defined that are composed of a data preparation block.

The Gaussian kernel scheme is applied in the SVM to maximize the hyper margin plane allows the SVM to fit the maximum margin in a transformed feature space. The hyper-plane in the middle passes with the optimal condition is as $v_i(w \cdot u_i + b) > = 1$. The decision function of the SVM is modeled as follows.

$$S(u) = \sum_{i \in V} \alpha_i v_i(u_i, u) + b \tag{2}$$

The primary task of kernel function is to identify the general structure of relations in the dataset. Operating SVM with a kernel trick would convert the linear model into a non-linear model. The model will be saved from the input space to feature space.

$$K\left(u, u^{'}\right) = \exp^{\frac{-\left\|u-u^{'}\right\|}{2\sigma^2}} \tag{3}$$

where σ is the kernel width parameter. The modeled kernel width parameter in data-dependent obeyed the cross-validation technique. The techniques of distributed, centralized machine learning, or hybrid based on

neighbors, self-detection have been applied for data fault detection schemes in WSN [24]. As the same style of the solution, an approach detection errors in big data sets WSN was developed based on the cloud computing scheme (CLOUD) [37]. One of the centralized techniques uses a hidden Markov model (HMM) to identify offset fault, gain fault, and struck at fault [38]. A method of minimum redundancy maximum relevance (MRMR) proposed to find a subset of features that are used for fault prediction for feature selection [39]. MRMR reduces redundancy in features and selects the relevant features. The maximum relevance feature is found by calculating mutual information difference, and the minimum redundancy feature is found by calculating mutual information quotient. The cloud-based technique is the hybrid-based fault detection technique, where the data collected from sensors are stored in cloud storage. Mapreduce is used for parallelism of fault detection tasks.

The mentioned methods though several were archived successfully, there still exist disadvantages, e.g., clean data preparation and feature selection before classifying data are not considered. Without noise removed from the data, they would cause misclassification or maybe increase the fault rate. It has failed to organize data if the test data was small that is limited in data learning, so it also leads to misclassification.

3 Fault detection scheme methodology

The aim of our proposed scheme is to establish a decision function in the CHs that can be used in real time for aggregating the precise data that classified any new collecting data from MNs and forwarding them to the base station in WSN. The proposed scheme of fault detection consists of the steps of collecting data, preprocessing data, identifying faults, selecting features, and classifying data. Figure 4 shows a flowchart of the steps in designing scheme involves the collection of data, preprocessing of data, identifying the fault, selecting the features, and classification of data.

The detail of the steps of the designing scheme is presented as following subsections.

3.1 Data collection

A dataset with the noise of the collected data from various sources by the sensor nodes for a system input would cause misclassification data. The datasets should be "clean" data by removing the noisy before further processing data. One of the preferred techniques used for performing the data cleaning process is called Gaussian smoothing [40]. Gaussian smoothing is like doing a filter using the original function convolution with the Gaussian weight. Data cleaning starts with the initial step of the design scheme after collecting data from sensors; filtering noise out from datasets before processing classification is to improve its performance and accuracy. An effective smoothing technique with an approximating function attempts to capture important patterns in the data while filtering noise out of the data. A form of collecting data is a vector of the data-points. Let S_i be a smoothed value at position ith with k is kernel size. The full width at half maximum is used as an expression of the extent of function given by the difference between two extreme values of the independent variables at vector data is equal to half of its maximum value.

$$F(\sigma, u_i) = \frac{1}{2}\pi\sigma^2 \times \exp^{-\frac{1}{2}\left(\frac{u_i}{\sigma}\right)^2} \tag{4}$$

where σ is the width of the curve and u_i is the distance from the origin point i in the horizontal axis. If a current spike occurs in collecting data, that could be the noise value, and the smoothing cost will be calculated and substituted them [41]. A weighted average of smoothed value is for the new points that are calculated as follows.

$$S_i = \begin{cases} \frac{1}{2k+1}\sum_{j=-k}^{k} u_{i+j}, & \text{if (no spike)} \\ \sum_{j=-k}^{k}\left(w_j * u_{i+j}\right), & \text{otherwise} \end{cases} \tag{5}$$

where w_j is the weighting factor of smoothing. The advantage of the weighted average is that considered for a straightforward implementation that is for every attribute in the resulting data, and fewer values in the weighted sum.

3.2 Data preprocessing and attribute selection

3.2.1 Preprocessing data

There are some types of faults based on the data-centric in WSN for sensor readings such as due to hardware failure, software failure, and communication failure [42]. This model of data gathered consists of the triplet d (n, t,f(.)) where n denotes the sensor which senses the data, t means the time the information is detected, and f(.) denotes the function of sensed data at time t by node n. The selected features from the smooth dataset after identifying the fault are selected by applying a technique called minimum redundancy maximum relevance (MRMR) as an effective filter [39]. The less redundant and more irrelevant of selecting features are identified by the MRMR algorithm to find the relevance with mutual information difference and to find the redundancy with mutual information quotient. The calibration fault and battery fault are also a significant fault type that affects the data-centric fault system. The time data fault occurs when the sensor failure or due to environmental conditions. The function of fault identification is formulated as follows.

$$f(x) = \alpha + \beta \cdot x + \gamma \tag{6}$$

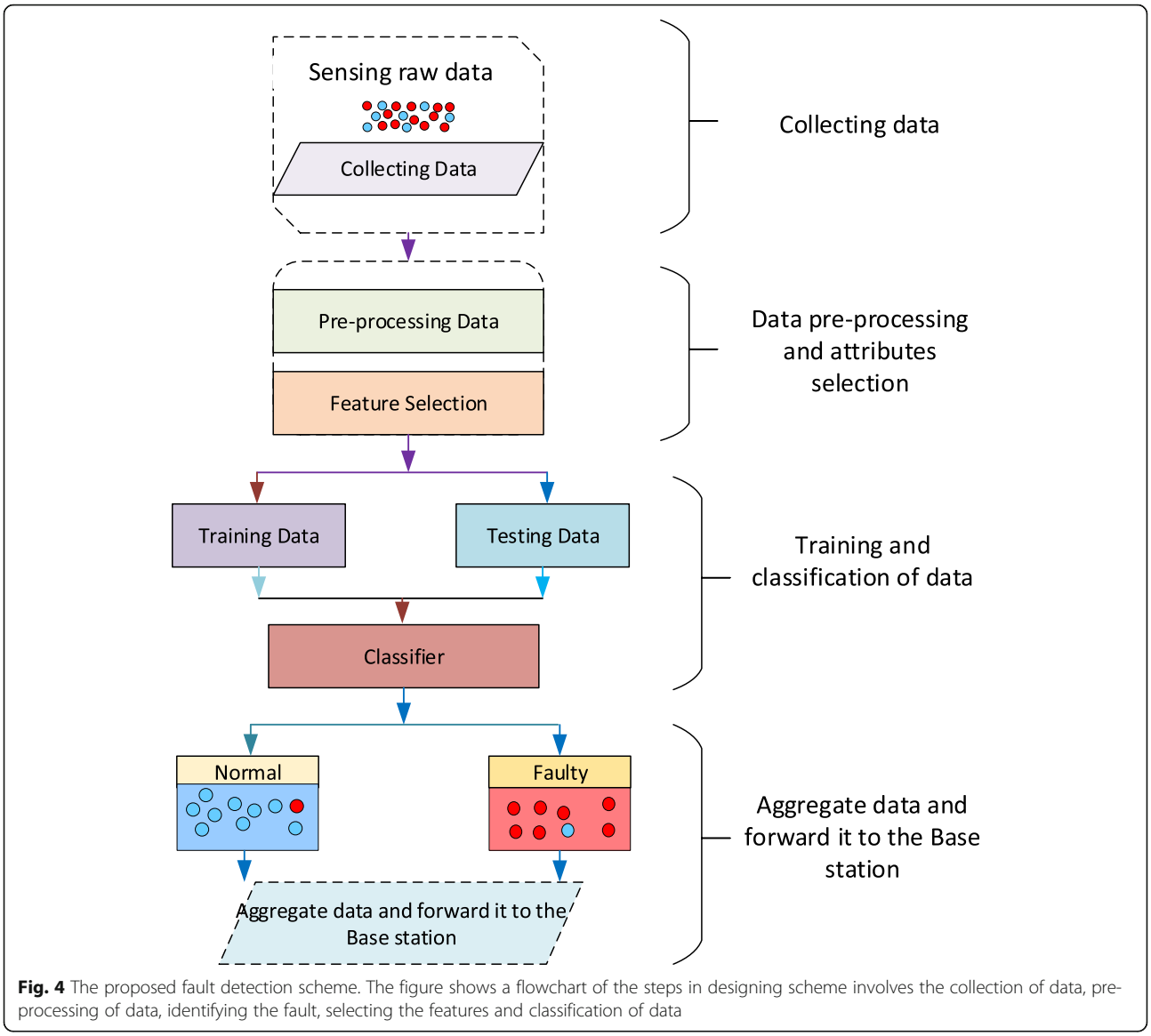


Fig. 4 The proposed fault detection scheme. The figure shows a flowchart of the steps in designing scheme involves the collection of data, pre-processing of data, identifying the fault, selecting the features and classification of data

where α, β, and x denote the offset, gain, and sensed data respectively, and γ denotes the noise which can be neglected. For example, it can be described as gain fault when the sensed data is different from expectation, so $f(x) = x$; offset fault is with $f(x) = \alpha + x$, where α is a constant value that is added to sensed data; hardware fault describes as sensed data that is null or 0, so $f(x) = \text{null}$; a hardcover fault is detected data increases above the maximum threshold, so $x \geq \theta$ where θ is the maximum threshold.

3.2.2 Attribute selection

Attribute selection is called variable selection, or feature selection, or variable subset selection that is a process of selecting the subset of relevant attributes concerning the output category [43–46]. The data training time and the over-fitting could be reduced due to applying attribute selection. The models are easy to interpret if attribute selection is figured out. Attribute selection on time-series data is crucial as the redundancy among the features. Attribute selection is distinguished into three main categories, which included the filter, wrapper, and embedded methods. The filter approach uses statistical methods to set the score for each attribute. The rating ranks the characteristics, and the elements with the highest rank are considered, and the rest is ignored. Some of the filter methods include chi-squared tests, information gain, and correlation coefficient scores. The wrapper approach uses a stochastic process in which a subset of attributes is evaluated and compared with other combinations. These scores are assigned based on the accuracy of the model. The heuristics of the features

are also considered. Cross-validation is done to improve feature selection. Characteristics are passed forward and backward to add or remove attributes. Some of the wrapper methods include recursive attribute elimination algorithm and back-forward algorithm. The embedded approach uses built-in feature selection methods. Inbuilt functions are used to reduce over-fitting. Examples of embedded methods include regularized trees and mimetic algorithms.

A filter approach as the MRMR is used in this paper for connecting to output with maximum relevance and left out elements that are redundant as minimum redundancy. The attributes are selected by applying an objective function as a greedy search function of relevance and redundancy. Two types of objective functions commonly are mutual information difference (MID) criterion to calculate relevance and mutual information quotient (MIQ) criterion to calculate excess. The MID measures how much the information is shared between two independent variables. Two variables u and v of the probability distribution are independent, relating mutual information is formulated as follows.

$$A(U,V) = \sum_{v\in V}\sum_{u\in U} p(u,v) * \log\left(\frac{p(u,v)}{p(u)*p(v)}\right) \tag{7}$$

where $A(U, V)$ is the relating mutual information with probability distribution; $p(u, v)$ is the joint probability of U and V; $p(u)$ and $p(v)$ are the marginal probability of U and V respectively; $p(U = u|\ V = v)$ is the conditional probability of U and V.

3.3 Training and classifying data

3.3.1 Training dataset

There are two phases of the fault detection solution in the proposed scheme: the data learning and real-time decision-making stages. In the data learning phase, the essential elements of the data are respected and maintained in the process. The data learning stage also uses a statistical learning method. The needed experience from the expertise to resolve different problems affecting WSNs is applied in learning data phrase because the classification based on data learning allows using expertise in making decisions. The classifiers and the decision function are implemented in the cluster head in WSN for aggregating accuracy data. The aim is to establish a real-time decision function in CHs to classify any new collecting data from sensor node members.

A labeled dataset is used as a learning database. It is composed of a set of regular data and a set of erroneous data. Figure 5 shows two phases in a model of the fault detection solution.

In the decision-making phase, a new data vector is constructed with blocks of data measurements V_t that included three measures (V_t, V_{t-1}, V_{t-2}). The decision function will makeover the new data vector. If its result is positive, it belongs to the standard case (a class of normal data functionality); otherwise, it is considered as a faulty case. The computationally inexpensive by applying a simple decision function used in the cluster head that makes the proposed scheme very efficient with sensors as limited resource nodes.

3.3.2 Classifying data

A part of a dataset with the selected attributes would be used to train the dataset by applying the model learning, the kernel function, the cross-validation, and expert's experience values. After the dataset is trained, the classifier model would be applied to test data for classification. Figure 6 shows a model of class attributes of fault detection solution. The decision function for fault detection can decide whether the attribute data belong to a class or not.

Let $x_0, \ldots, x_n$ and $c_0, \ldots, c_k$ be attributes and classes respectively, where n and k are the numbers of attributes and classes involved. The probability of the attributes occurring in each category is determined and returned to the most likely category. $P(c_i)$ is the proportion of the dataset that falls in category c_i.

The naive Bayes probability model is an independent attribute model that can derive constructing a classifier. The classifier combines the model with a decision rule. The probability of a collected data is estimated character belongs to a specific category that can process classifying with naive Bayes. The most substantial likelihood received data point's attributes are to pick the c_i by naive Bayes. Each class, $P(c_i\ |x_0, \ldots, x_n)$ is calculated for classification, and a faulty detection strategy is to predict as policy follows.

$$z = \arg\ \max_{c_i} P(c_i)\prod_{j=1}^{n} P(x_j|c_i) \tag{8}$$

One common practice is to pick the most probable hypothesis; this is called the maximum a posteriori. The corresponding Bayes classifier is the function that assigns a class label c_i. In the classification, an attribute can be replaced with a new adjusted set of attributes as the category c_i in collecting data.

4 Experimental results and discussions

4.1 Setting parameters

Assumed an N-node topology of cluster-based WSN was deployed in scattering the area of $M \times M$ randomly, where N is a number of nodes that can be 100, 200, 300; M is a length measurement of the deployed area that can be 200, 300, 400 m. The network has a base station (BS) that

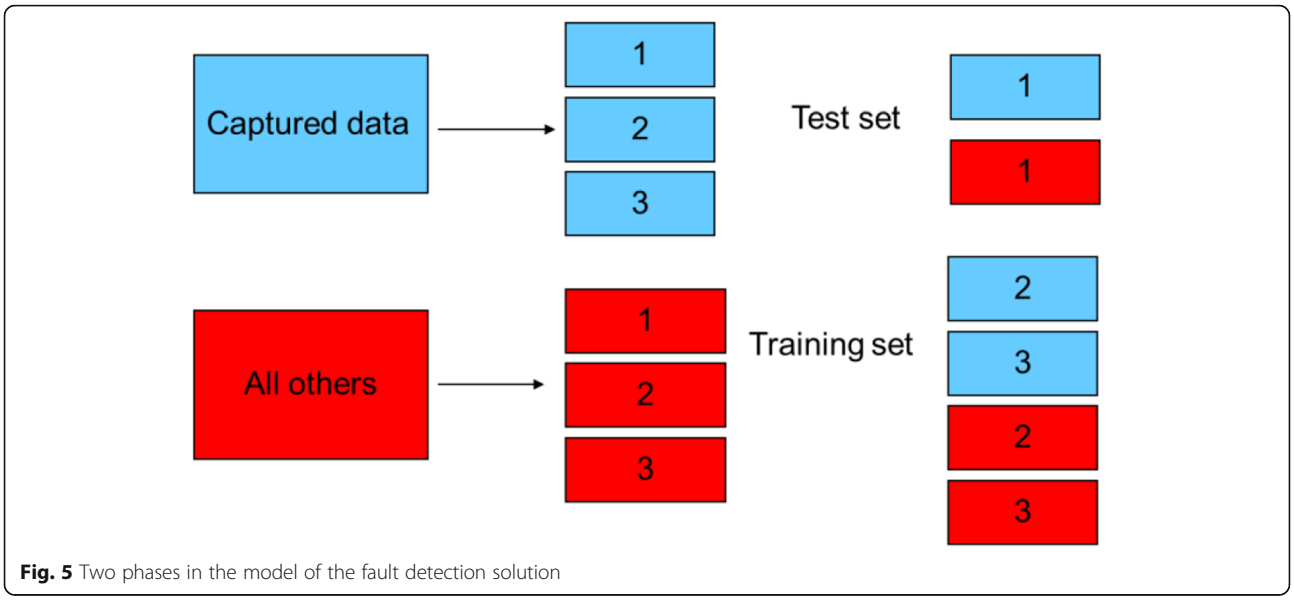


Fig. 5 Two phases in the model of the fault detection solution

operates with an unlimited power supply. BS receives the aggregated data from CHs. The characteristics of WSN operation were assumed that behaving like scheduling periods of packet transmission time. Table 1 lists the initial values for setting the parameters of the experiment.

The information of the packets such as data transmission time and source node IDs received by the CHs are aggregated and forwarded to the BS. Data were randomly picked up by member nodes and were forwarded to the CHs. Then, the CHs will pack the data to be packets and transmit them towards the BS node.

There are two circumstances: default and adjusted options of the cluster in WSN. The default option is the formed clusters without any adjustment from its outside. However, the customized option is applied to the unequal group for balancing energy in the WSN (mentioned in Section 4.4).

4.2 Visualization results with default option

The attributes, e.g., temperature, humidity, light, network status, could be "normal" or "faulty." For an example of the network status, if the attribute is class labeled "normal," the network consists of no defective sensor. Otherwise, the system includes at least one faulty sensor in both the training and testing phase. Figure 7 shows a visualization of training and testing set results. The collecting dataset is taken from outdoor data collection from multi-hop WSN with a total of 4001 samples that consist of temperature measurements for 5-min period of scheduling loop for packets data with a total of 60 min.

Different scenarios are simulated, such as some set faulty nodes within the network, i.e., increasing or decreasing temperature, humidity, and sensed data damaged in traffic congestion conditions. Under each setting conditions, no-fault network and different numbers of

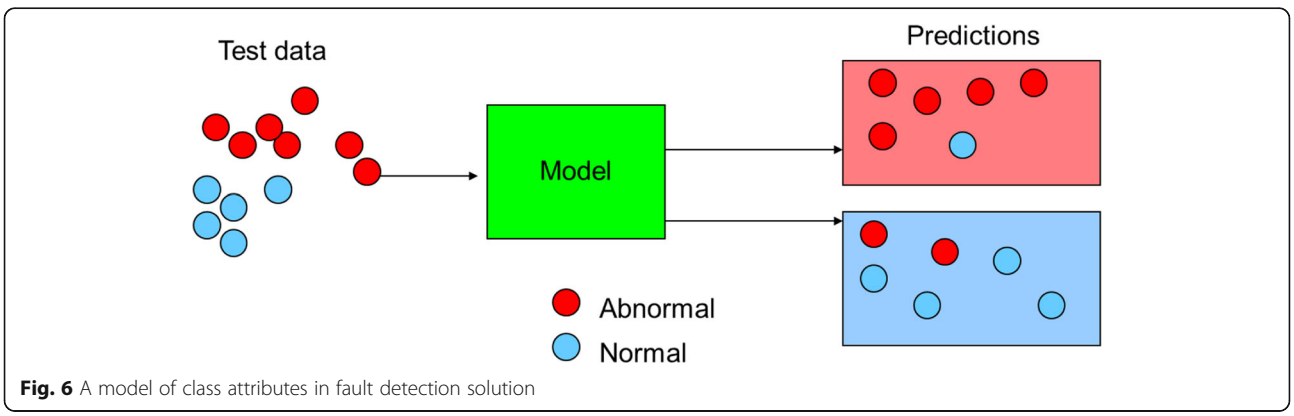


Fig. 6 A model of class attributes in fault detection solution

Table 1 Initial values for setting parameters of the experiment

Parameters noticed	Denoted symbols	Initial values
Initial node energy	E_j	0.5 J
Data aggregation energy	E_{DA}	5n J/bit/signal
Receiving and transmitting energy	E_{fs}	10 pJ/bit/m^2
Radio electronics energy	E_{elec}	50 nJ/*bit*
Number bit of a data message	l	1024 bit
Amplifier energy	E_{mp}	0.013 pJ/bit/m^4
Number of nodes in WSN	N	100/200/300/nodes
Space distribution	M	100/200/300 m
Generations	MaxIter	2000
Number of runs	runs	25

faulty nodes within the system are generated. For an N-node topology, there are some combinations of faulty nodes with varying from 1 to 5, respectively. There are about 200 data packets under each scenario generated by sensor nodes randomly.

4.3 Measurement results

The metrics of the detection accuracy (DA), the true positive rate (TPR), and the false positive rate (FPR) are used to evaluate the accuracy of the proposed technique as follows.

$$\mathrm{DA} = \frac{\text{number of faulty data identified}}{\text{total number of faulty data present}} \tag{9}$$

where DA is detection accuracy that is clarified as the ratio of the number of erroneous data identified to the total number of current incorrect data.

$$\mathrm{TPR} = \frac{\text{number of faulty data identified as faulty}}{\text{total number of faulty data}} \tag{10}$$

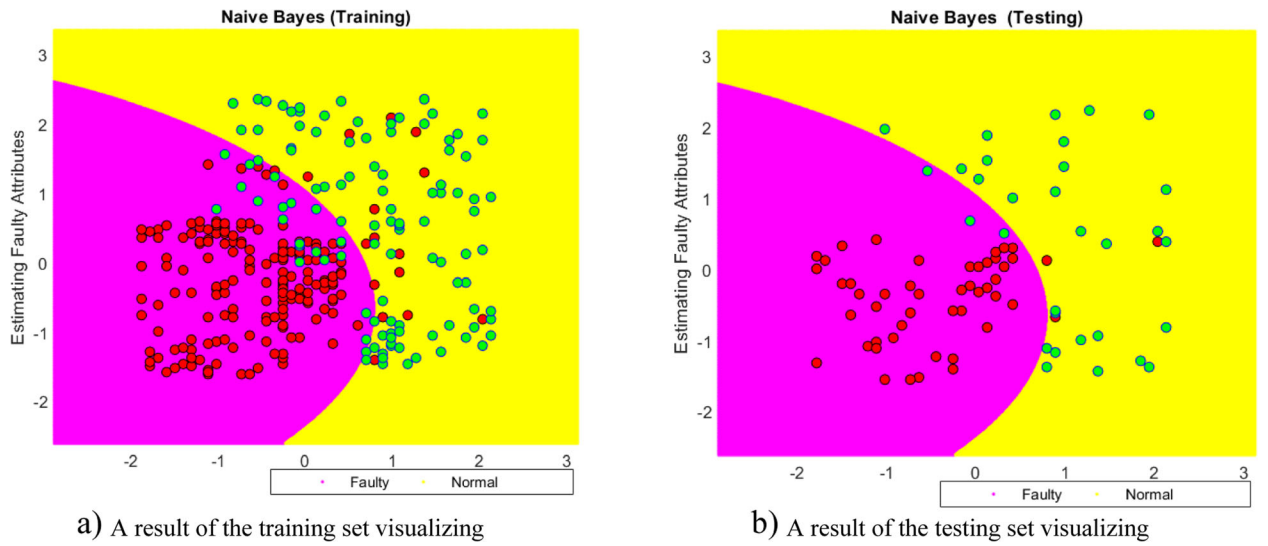
where TPR is the true positive rate that is classified as the ratio of some faulty data identified as faulty to the total number of incomplete data. It refers to the quantity that predicts the true positive as a positive.

$$\mathrm{FPR} = \frac{\text{number of non faulty data identified as faulty}}{\text{total number of faulty data}} \tag{11}$$

Here, FPR is the false positive rate that is classified as the ratio of some non-faulty data identified as faulty to the total number of incomplete data. Table 2 summarizes the results of the training accuracy of various fault types.

The experimental results of the proposed scheme for the classification dataset of collecting temperature are compared with the cases of the methods, e.g., the support vector machine (SVM) [26], the decision tree (DT) [36], the hidden Markov model (HMM) [38], and the cloud computing scheme (CLOUD) [37] concerning the false positive rate of training set. In the measurement of the results of the proposed scheme in comparison with other methods, a mathematical tool known as the Hausdorff metric [25] is used to determine the distance difference between two datasets. Assumed, there are two non-empty of subsets A and B; the distance between them calculated as follows.

$$D_H(A,B) = \max\left\{ \sup_{x\in A} inf_{y\in B} d(x,y),\ \sup_{y\in B} \inf_{x\in A} d(x,y) \right\} \tag{12}$$



a) A result of the training set visualizing

b) A result of the testing set visualizing

Fig. 7 A visualization of the training and testing set results

Table 2 Summary of the effects of training accuracy of different fault types

	Data-loss fault			Hardware fault			Drift fault			Gain fault		
Node ID	TPR (%)	DA (%)	FPR (%)	TPR (%)	DA (%)	FPR (%)	TPR (%)	DA (%)	FPR (%)	TPR (%)	DA (%)	FPR (%)
001	86.5	89	26	86.5	99	32	86.5	87	32	86.5	92	28
002	83.5	88	32	83.5	98	37	83.5	84	31	83.5	90	32
030	85.2	90	24	85.2	89	34	85.2	85	34	85.2	92	34
034	86.3	93	31	86.3	93	34	86.3	86	31	86.3	93	31
050	84.5	87	30	84.5	97	36	84.5	85	30	84.5	87	30
056	86.1	86	22	86.1	96	32	86.1	86	32	86.1	91	32
070	85.2	89	34	81.8	89	34	85.1	85	36	83.5	89	26
078	86.3	88	31	84.9	89	35	97.3	97	32	85.2	88	32
099	87.0	91	30	84.1	88	35	86.1	89	34	86.3	85	34
AVG	**85.6**	**88.9**	**28.9**	**84.8**	**93.1**	**34.3**	**86.7**	**87.1**	**32.4**	**85.2**	**89.6**	**31.0**

where $D_H(A, B)$ is Hausdorff distance between two subsets A and B; sup and inf are the supremum and infimum respectively. Figures 8 and 9 depict the comparison of the proposed scheme for the classification of a collecting temperature dataset with the SVM, DT, HMM, and CLOUD methods on the faulty of hardware errors and data-loss fault respectively in the scenario of the adjusted option. It can be seen that our proposed method outperforms other competitors in that the average outcomes of DA are at 93.1% and 89.0% on the hardware faulty and data-loss fault respectively.

Figure 10 shows a comparison of the proposed method according to FPR with the SVM, DT, HMM, and CLOUD. The proposed method provides the best value of FPR that helps to make a significant and essential improvement of FPR as compared with others. The growth is starting from 59% compared with the SVM and reaches 59% compared with the DT, HMM, and CLOUD.

4.4 Adjusted options results

Due to the topology of the clustering WSN fashion, the system often falls into the hotspot problem. It means

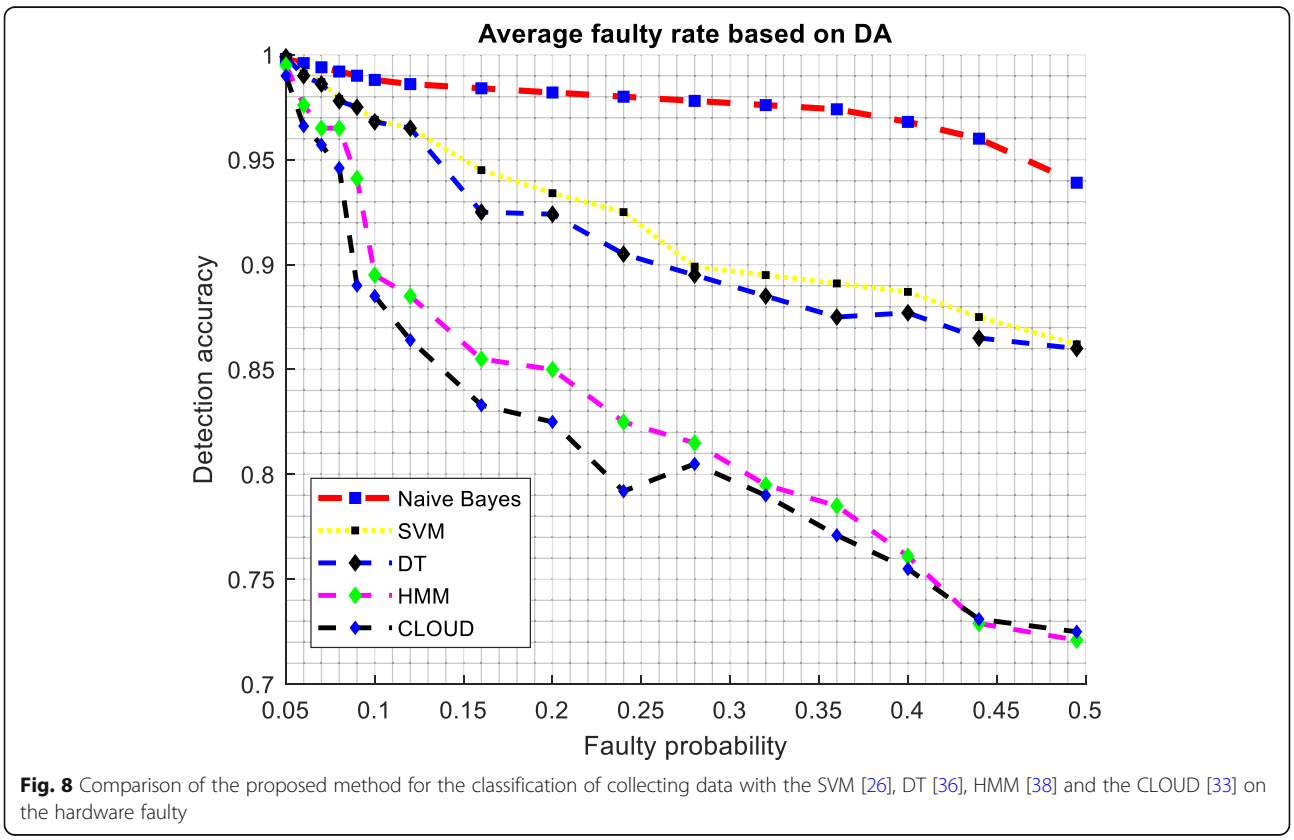


Fig. 8 Comparison of the proposed method for the classification of collecting data with the SVM [26], DT [36], HMM [38] and the CLOUD [33] on the hardware faulty

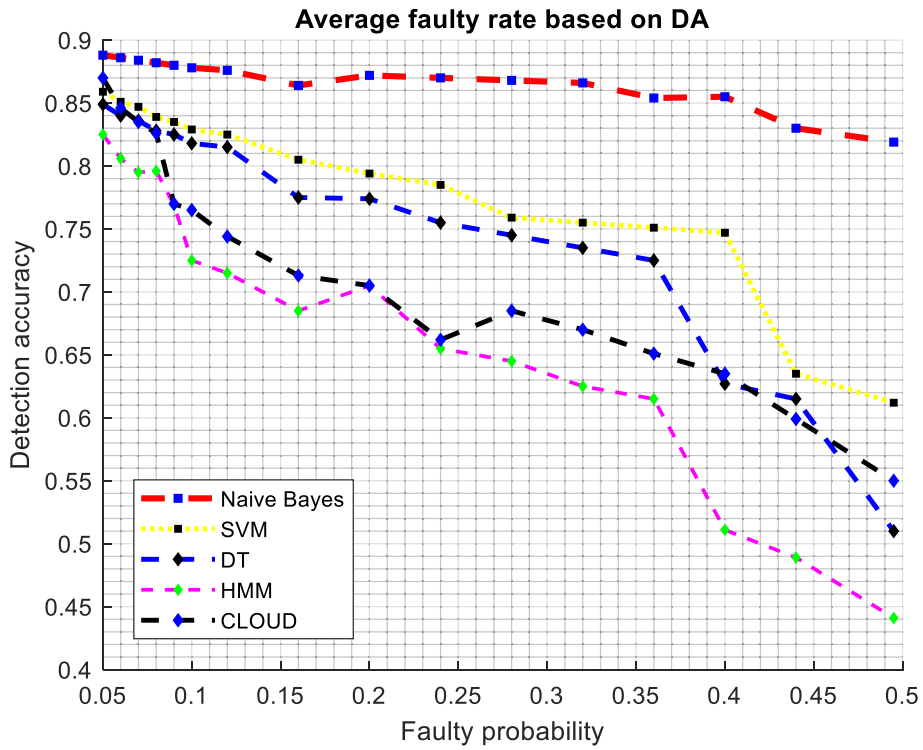


Fig. 9 Comparison of the proposed method for the classification of collecting data with the SVM [26], DT [36], HMM [38] and the CLOUD [33] on the data-loss fault

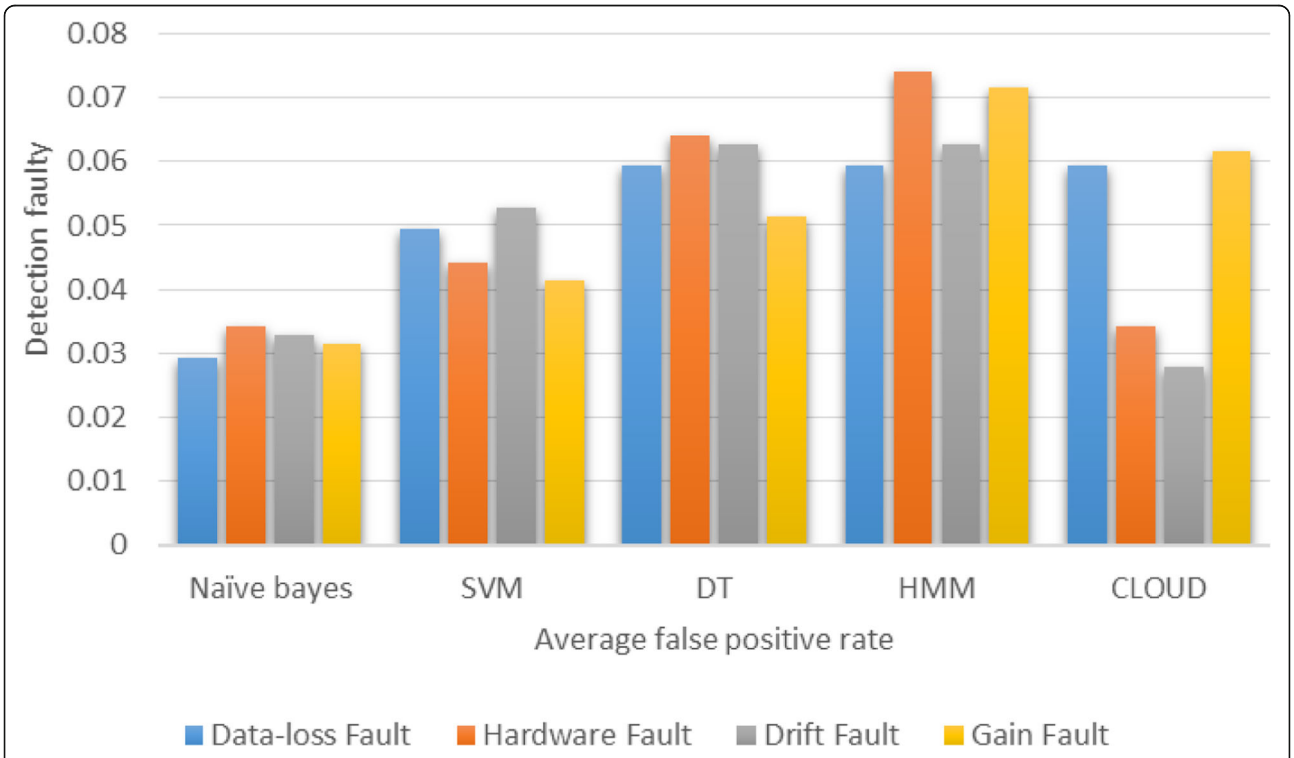


Fig. 10 Comparison of the proposed method for the classification of a collecting temperature dataset with the SVM, DT, HMM and CLOUD methods on concerning the FPR metric

that the clusters that are closer to the BS is getting hotter than the other groups that are far away [1, 10]. Because the group is the closer BS, the more profoundly is the traffic load that causes coverage issues of hotspot problems. It is necessary to adjust and amend this situation of faulty nodes according to the cluster WSN. The affection of a faulty node can affect its distributed neighbors of cluster size [1]. The calculations for the posterior fault probability of every node can affect the results of the classification. The fault node causes the similarity status to its neighbors in the cluster. The fault probability should be adjusted by exploiting the cluster's size. The new confidence of nodes is figured out as adjusting follows.

$$c_{i_j}^t = c_{i_j}^{t-1} \times R_c \tag{13}$$

where $c_{i_j}^t$ is the confidence of node j in the cluster i; t is the current time of generation; R_c is the ratio adjusting parameter that is determined as follows.

$$R_c = \left[1-\beta\frac{d_{\max}-D_j}{d_{\max}-d_{\min}}-\alpha\left(1-\frac{E_r}{E_{\max}}\right)\right] \times R_{\max} \tag{14}$$

where $d_{\max}$, $d_{\min}$, and D_j are the maximum and minimum distance of the CHs in the network to the BS and the distance from node CH_j to the BS; $R_{\max}$ is the maximum value of competition radius; β and α are the weighted factors with value in [0, 1]; E_r presents the residual energy of CH_j. The probability of the faulty nodes' confidence is calculated as follows:

$$P_{x_{ij}} = P\left(c_{ij}|x_{ij} \text{ is fault}\right) = \begin{cases} \frac{1}{2}, & \text{if all confidence } c < 0 \\ 1-P_{x_{ik}}, & \text{choose } x_k \text{ with maximum } c \\ P_{x_{ik}}, & \text{if } x_i \text{ is not same cluster} \end{cases} \tag{15}$$

where P_{x_i} and $P_{x_{ik}}$ are the probability of the faulty node x_{ij} and x_{ik} respectively.

Figure 11 shows a comparison of the proposed scheme for the classification of a collecting dataset of WSN with the SVM, DT, HMM, and CLOUD methods for different faulty, e.g., data loss, hardware, drift, and gain fault according to the DA measurements.

Figure 12 shows a visualization of the comparison between the adjusted and default options of the testing set results. Figure 13 depicts the comparison between the adjusted and default options of the proposed scheme for the testing set with different faulty, e.g., data loss, hardware, drift, and gain faults according to the DA measurements. Observing the results from Figs. 12 and 13, it is seen that the proposed scheme with adjusted options can provide more improved accuracy that reaches at 95%, 98%, 91%, and 90% in comparison with default cases for data loss, hardware, drift, and gains fault according to the DA measurements.

5 Conclusion

In this paper, a new scheme of collecting data classification for aggregating data in cluster heads (CHs) in hierarchical WSN based on the naive Bayes

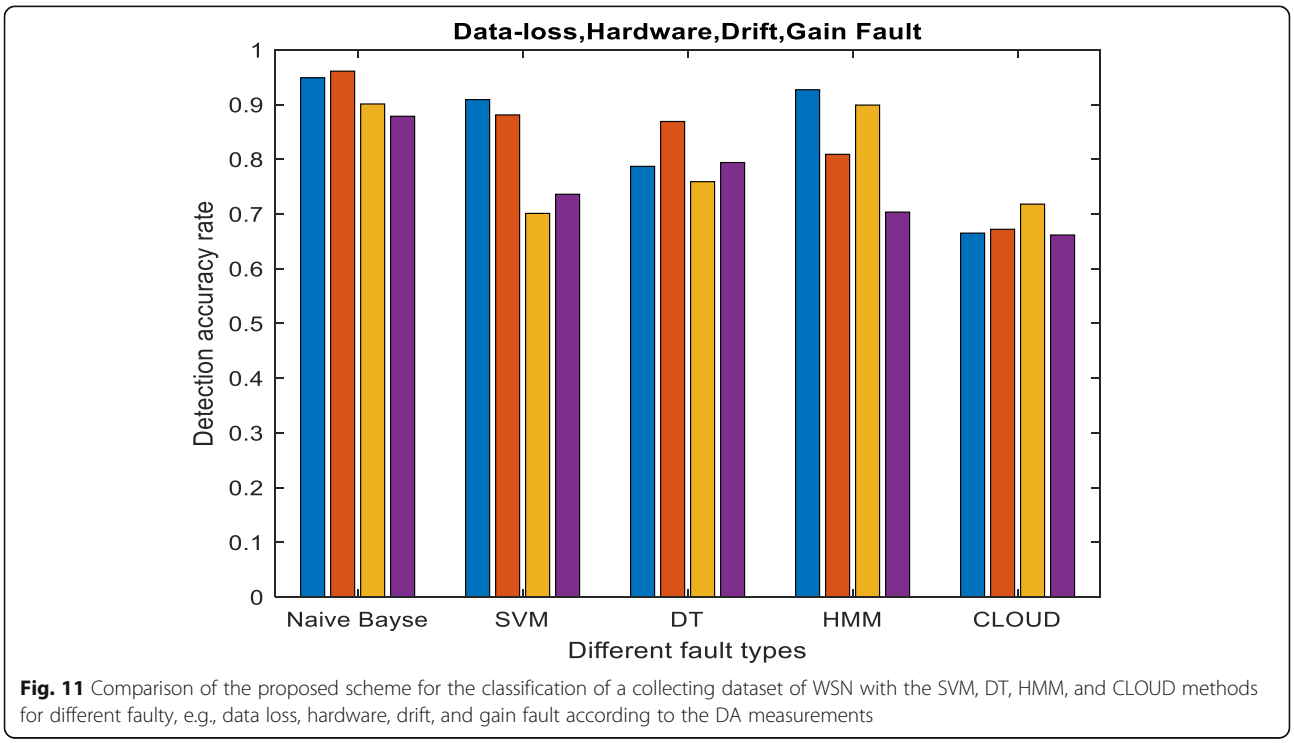


Fig. 11 Comparison of the proposed scheme for the classification of a collecting dataset of WSN with the SVM, DT, HMM, and CLOUD methods for different faulty, e.g., data loss, hardware, drift, and gain fault according to the DA measurements

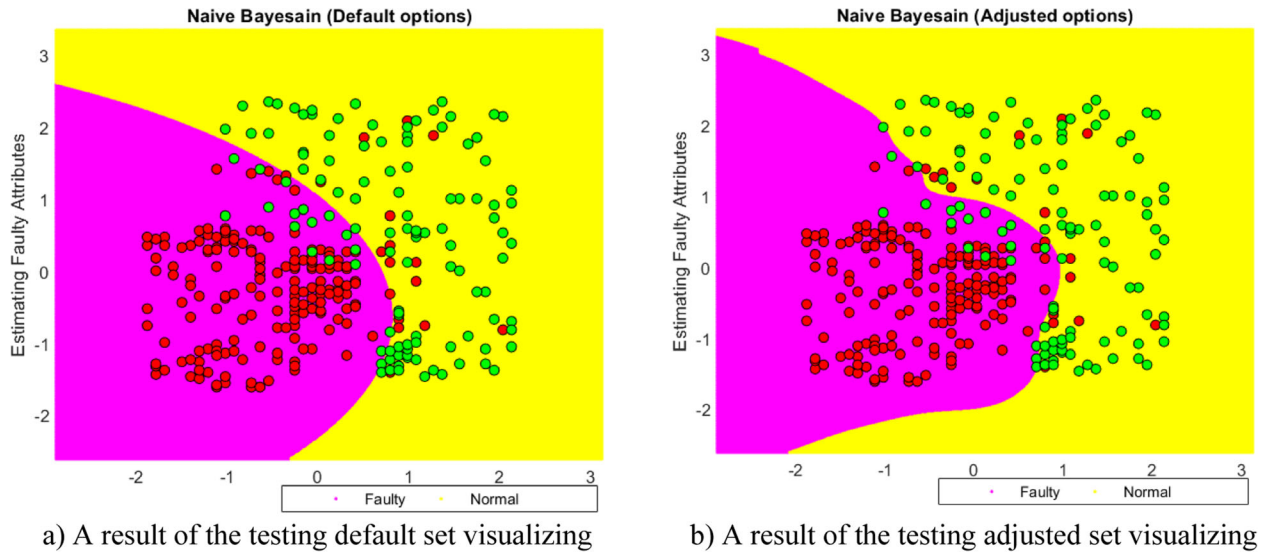


Fig. 12 A visualization of the comparison between the adjusted and default options of the testing set results

approach was presented. Due to the requirement of the precise data in several successful WSN applications, a decision function of classification should be deployed in CHs for faulty detection to aggregate the usual data for the next process. The collecting environmental information like temperature, humidity, and pollution levels is classified as collecting "fault" or "normal" data to aggregate and transfer them to the base station (BS).

The system design of the proposed scheme consists of majority components such as the collecting and preprocessing data, identifying and normalizing attributes of data, and training and testing datasets. The noise is removed from the data that can be sufficient to obtain more accurate. The elements of selected attributes in datasets can enable detection accuracy. In the experimental section, the system was tested with the collecting data by naive Bayes classification. Compared with the

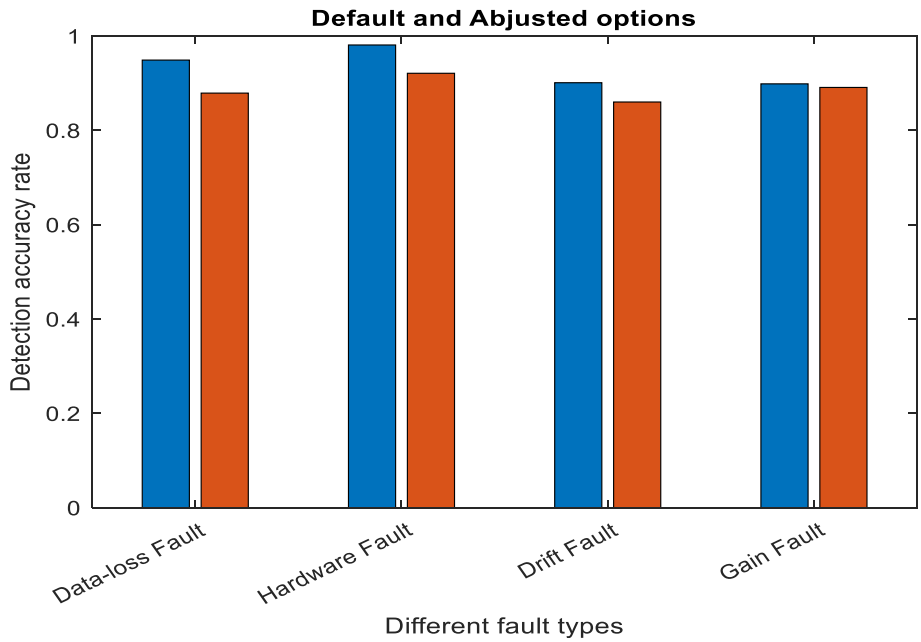


Fig. 13 Comparison between the adjusted and default options of the proposed scheme for the testing set with different faulty, e.g., data loss, hardware, drift, and gain faults according to the DA measurements

other methods in the literature, e.g., the support vector machine (SVM), decision tree (DT), hidden Markov model (HMM), and cloud computing scheme (CLOUD), it shows that the proposed method offers an effective way of forwarding the correct data for WSN applications. The system provides an accuracy of more than 97% throughout the data learning process. In data testing, the efficiency of the improved data fault detection offers more precise than any other competitive systems.

In future work, the proposed scheme may be further improved by adopting some efficient approaches [47–49] for optimal classification parameters; and it also may hybridize with the method of the neural network [50].

Abbreviations

BS: Base station; CH: Cluster head; CLOUD: Cloud computing scheme; DT: Decision tree; HMM: Hidden Markov model; NB: Naive Bayes; NM: Node members; SVM: Support vector machine; WSN: Wireless sensor network

Acknowledgements

This work was supported by the Natural Science Foundation of Fujian Province under Grant No. 2018J01638.

Authors' contributions

SC proposed the idea of the scheme of aggregation on WSN based on classification methods. TK carried out the simulations and drafted the paper. JP supervised the work and introduced the idea of applying the proposed scheme into dynamic deployment in WSN. TN gave some suggestions for the article and revised the manuscript. All authors read and approved the final manuscript.

Authors' information

Shu-Chuan Chu received the PhD degree from Flinders University of South Australia, Australia, in 2004. Currently, she is with the College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China. Her research interests include pattern recognition and data mining, computational intelligence, and cloud computing.

Thi-Kien Dao received her PhD degree in Electronics and Engineering from the National Kaohsiung University of Technology and Sciences, Taiwan, in 2019. She is currently a lecturer in the College of Information Science and Engineering, Fujian University of Technology. Her current research interests include computational intelligence, data mining, and sensor networks.

Jeng-Shyang Pan received the PhD degree in Electrical Engineering from the University of Edinburgh, UK, in 1996. Now, he is with the College of Computer science and Engineering, Shandong University of Science and Technology, Qingdao, China. His current research interests include soft computing, sensor networks, and signal processing.

Trong-The Nguyen received his PhD degree in Communication Engineering from the National Kaohsiung University of Applied Sciences, Taiwan, in 2016. He is currently a lecturer in the College of Information Science and Engineering, Fujian University of Technology. His current research interests include computational intelligence and sensor networks.

Funding

No funding.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China. [2]Fujian Provincial Key Lab of Big Data Mining and Applications, Fujian University of Technology, Fuzhou 350118, China. [3]Department of Information Technology, Haiphong University of Manage and Technology, Haiphong 180000, Vietnam.

Received: 16 November 2019 Accepted: 13 February 2020

References

1. T.-T. Nguyen, J.-S. Pan, T.-K. Dao, A compact bat algorithm for unequal clustering in wireless sensor networks. Appl. Sci. **9**(10) (2019) doi.org/10.3390/app9101973
2. H.Y. Kung, C.H. Chen, H.H. Ku, Designing intelligent disaster prediction models and systems for debris-flow disasters in Taiwan. Expert Syst. Appl. **39**(5), 5838–5856 (2012)
3. W.Z. Guo, W.P. Zhu, Z.Y. Yu, J.T. Wang, B. Guo, A survey of task allocation: contrastive perspectives from wireless sensor networks and mobile crowd sensing. IEEE Access **7**, 78406–78420 (2019)
4. J. Wang, Y. Gao, W. Liu, W. Wu, S.-J. Lim, An asynchronous clustering and mobile data gathering schema based on timer mechanism in wireless sensor networks. Comput. Mater. Contin. **58**(3), 711–725 (2019)
5. J.S. Pan, C.Y. Lee, A. Sghaier, M. Zeghid, J. Xie, Novel systolization of subquadratic space complexity multipliers based on toeplitz matrix-vector product approach. IEEE Trans. Very Large Scale Integr. Syst. **27**(7), 1614–1622 (2019)
6. C.F. García-hernández, P.H. Ibargüengoytia-gonzález, J. García-hernández, J.a. Pérez-díaz, Wireless sensor networks and applications : a survey. J. Comput. Sci. **7**(3), 264–273 (2007)
7. C.H. Chen, C.A. Lee, C.C. Lo, Vehicle localization and velocity estimation based on mobile phone sensing. IEEE Access **4**, 803–817 (2016)
8. N. Liu, J.-S. Pan, T.-T. Nguyen, A bi-population QUasi-Affine TRansformation Evolution algorithm for global optimization and its application to dynamic deployment in wireless sensor networks. EURASIP J. Wireless Com. Network **2019**, 175 (2019) doi.org/10.1186/s13638-019-1481-6
9. Y.Z. Chen, S.N. Weng, W.Z. Guo, N.X. Xiong, A game theory algorithm for intra-cluster data aggregation in a vehicular ad hoc network. Sensors **16**(2), 245 (2016)
10. J.-S. Pan, T.-T. Nguyen, T.-K. Dao, T.-S. Pan, S.-C. Chu, Clustering formation in wireless sensor networks: a survey. J. Netw. Intell. **2**(4), 287–309 (2017)
11. T.-T. Nguyen, J.-S. Pan, T.-K. Dao, A novel improved bat algorithm based on hybrid parallel and compact for balancing an energy consumption problem. Information **10**(6), 194 (2019)
12. T. Nguyen, J. Pan, T. Dao, An improved flower pollination algorithm for optimizing layouts of nodes in wireless sensor network. IEEE Access **7**, 75985–75998 (2019)
13. S.G.S.P. Yadav, A. Chitra, Wireless sensor networks - architectures, protocols, simulators and applications : a survey. Int. J. Electron. Comput. Sci. Eng. **1**(4), 1941–1953 (2012)
14. T.-T. Nguyen, T.-K. Dao, M.-F. Horng, C.-S. Shieh, An energy-based cluster head selection algorithm to support long-lifetime in wireless sensor networks. J. Netw. Intell. **1**(1), 23–37 (2016)
15. C.H. Chen, F.J. Hwang, H.-Y. Kung, Travel time prediction system based on data clustering for waste collection vehicles. IEICE Trans. Inf. Syst. **E102.D**(7), 1374–1383 (2019)
16. A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, J. Anderson, *Wireless sensor networks for habitat monitoring, WSNA '02: proceedings of the 1st ACM international workshop on wireless sensor networks and applications. Melbourne, Qld., Australia* (2002), pp. 88–97
17. W.Z. Guo, G.L. Chen, C.L. Yu, J.S. Su, Z.H. Liu, A two-stage clustering sleep scheduling algorithm with particle swarm optimization in wireless sensor networks. Adhoc Sens. Wirel. Networks **27**, 27–49 (2015)
18. H.J. Cheng, Z. Xie, Y.S. Shi, N.X. Xiong, Multi-step data prediction in wireless sensor networks based on one-dimensional CNN and bidirectional LSTM. IEEE Access **7**, 117883–117896 (2019)
19. S. Jia, L. Ma, D. Qin, Fault detection modelling and analysis in a wireless sensor network. J. Sensors (2018) doi.org/10.1155/2018/7935802
20. F. Yuan, Y. Zhan, Y. Wang, Data density correlation degree clustering method for data aggregation in WSN. IEEE Sensors J. **14**(4), 1089–1098 (2014)
21. J.-S. Pan, L. Kong, T.-W. Sung, P.-W. Tsai, V. Snasel, α-Fraction first strategy for hierarchical wireless sensor networks. J. Internet Technol. **19**(6), 1717–1726 (2018)
22. H.C. Shih, J.H. Ho, B.Y. Liao, J.S. Pan, Fault node recovery algorithm for a wireless sensor network. IEEE Sensors J. **13**(7), 2683–2689 (2013)

23. C.I. Wu, H.Y. Kung, C.H. Chen, L.C. Kuo, An intelligent slope disaster prediction and monitoring system based on WSN and ANP. Expert Syst. Appl. **41**(10), 4554–4562 (2014)
24. A. Mahapatro, P.M. Khilar, Fault diagnosis in wireless sensor networks: a survey. IEEE Commun. Surv. Tutorials **15**(4), 2000–2026 (2013)
25. Z. Zhang, A. Mehmood, L. Shu, Z. Huo, Y. Zhang, M. Mukherjee, A survey on fault diagnosis in wireless sensor networks. IEEE Access. **6**, 11349–11364 (2018)
26. S. Zidi, T. Moulahi, B. Alaya, Fault detection in wireless sensor networks through SVM classifier. IEEE Sensors J. **18**(1), 340–347 (2018)
27. T. Muhammed, R.A. Shaikh, An analysis of fault detection strategies in wireless sensor networks. J. Netw. Comput. Appl. **78**, 267–287 (2017)
28. R. Sathiyavathi, B. Bharathi, *A review on fault detection in wireless sensor networks, International Conference on Communication and Signal Processing (ICCSP), 1487–1490* (2017)
29. R. Luckasson et al., *Mental retardation: Definition, classification, and systems of supports. American Association on Mental Retardation, 10th edn.*. American Association on Intellectual and Developmental Disabilities, Washington (2002)
30. M. Aly, Survey on multiclass classification methods. Neural Netw. **19**, 1–9 (2005)
31. H.B. Barlow, Unsupervised learning. Neural Comput. **1**(3), 295–311 (1989)
32. X.J. Zhu, *Semi-supervised learning literature survey, University of Wisconsin-Madison Department of Computer Sciences, 19* (2005)
33. P.-N. Tan, *Introduction to data mining* Pearson Addison-Wesley, Boston (2018)
34. B.C.P. Lau, E.W.M. Ma, T.W.S. Chow, Probabilistic fault detector for wireless sensor network. Expert Syst. Appl. **41**(8), 3703–3711 (2014)
35. T. Calders, S. Verwer, Three naive Bayes approaches for discrimination-free classification. Data Min. Knowl. Disc. **21**(2), 277–292 (2010)
36. A. Garofalo, C. Di Sarno, V. Formicola, *Enhancing intrusion detection in wireless sensor networks through decision trees, Dependable Computing. EWDC 2013. Lecture Notes in Computer Science*, vol 7869 (Springer, Berlin, Heidelberg, 2013), pp. 1–15
37. C. Yang, C. Liu, X. Zhang, S. Nepal, J. Chen, A time efficient approach for detecting errors in big sensor data on cloud. IEEE Trans. Parallel Distrib. Syst. **26**(2), 329–339 (2015)
38. P. Tang, T.W.S. Chow, Wireless sensor-networks conditions monitoring and fault diagnosis using neighborhood hidden conditional random field. IEEE Trans. Ind. Informatics **12**(3), 933–940 (2016)
39. C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data. J. Bioinforma. Comput. Biol. **3**(2), 185–205 (2005)
40. V. Aurich, J. Weule, *in Mustererkennung. Non-linear Gaussian filters performing edge preserving diffusion* (Springer, Berlin, Heidelberg, 1995), pp. 538–545
41. G.-R. Xue et al., in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil.* Scalable collaborative filtering using cluster-based smoothing (2005), pp. 114–121
42. A. Stoffelen, Toward the true near-surface wind speed: error modeling and calibration using triple collocation. J. Geophys. Res. Ocean **103**(C4), 7755–7766 (1998)
43. N. Kwak, C.-H. Choi, Input feature selection for classification problems. IEEE Trans. Neural Netw. **13**(1), 143–159 (2002)
44. J.-S. Pan, J.-W. Wang, Texture segmentation using separable and non-separable wavelet frames. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **82**(8), 1463–1474 (1999)
45. J.-W. Wang, C.-H. Chen, J.-S. Pan, Genetic feature selection for texture classification using 2-D non-separable wavelet bases. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **81**(8), 1635–1644 (1998)
46. P. Hu, J.-S. Pan, S.-C. Chu, Q.-W. Chai, T. Liu, Z.-C. Li, New hybrid algorithms for prediction of daily load of power network. Appl. Sci. **9**(21), 4514 (2019) doi.org/10.3390/app9214514
47. Z. Meng, J.S. Pan, K.K. Tseng, PaDE: An enhanced differential evolution algorithm with novel control parameter adaptation schemes for numerical optimization. Knowledge-Based Syst. **168**, 80–99 (2019)
48. T.-K. Dao, T.-S. Pan, T.-T. Nguyen, J.-S. Pan, Parallel bat algorithm for optimizing makespan in job shop scheduling problems. J. Intell. Manuf. **29**(2), 451–462 (2018)
49. H. Wang, S. Rahnamayan, H. Sun, M.G.H. Omran, Gaussian bare-bones differential evolution. IEEE Trans. Cybern. **43**(2), 634–647 (2013)
50. Y.F. Dai, W.Z. Guo, X. Chen, Z.W. Zhang, Relation classification via LSTMs based on sequence and tree structure. IEEE Access **6**, 64927–64937 (2018)

Publisher's Note