Photonic Network Communications (2019) 37:212–223
https://doi.org/10.1007/s11107-018-0810-8

ORIGINAL PAPER

Research on the intrusion detection model based on improved cumulative summation and evidence theory for wireless sensor network

Fengjun Shang[1] · Dan Zhou[1] · Cheng Li[1] · Hanyun Ye[1] · Yuting Zhao[1]

Received: 7 September 2018 / Accepted: 30 October 2018 / Published online: 19 November 2018

Abstract

In this paper, a new hybrid intrusion detection model which combines the distributed and centralized strategies is proposed in this paper as follows. Firstly, considering the network anomalies, situation cannot be captured in real time on the base station; by introducing the CUSUM (cumulative summation) GLR (generalized likelihood ratio), an anomaly detection model which runs on the node is given. It can conduct real-time network monitoring. Based on the "link quality" and "majority rule," a new algorithm to detect the "Sinkhole attack" in the base station is proposed, and a new model CUSUM_MV to detect intrusion is given. Secondly, the evidence theory is introduced to detect intrusion in wireless sensor network. We give the redundant information process mechanism in the relay node, an evidence-based intrusion detection model deployed on the base station and the intrusion detection model CUSUM_HDST. The hybrid model can detect not only Sinkhole and DoS attacks, but also other specific vulnerabilities. A simulation experiment on Castalia simulator is carried out, and results show that the proposed method has better performance than the traditional Sinkhole attacks detection method.

Keywords Wireless sensor networks · Intrusion detection · CUSUM GLR · Castalia · D–S evidence theory

1 Introduction

Wireless sensor network (WSN) is a new information acquisition and processing technology with broad applications [1, 2]. The wireless sensor network is a major technology that drives the development of precision agriculture. WSNs increase the efficiency of sustainable development. Increases in agricultural efficiency will stem from networking sensors that elucidate important spatiotemporal patterns and integrate their data streams to not only display or record information, but also actuate human and autonomous responses. This involves monitoring soil, crop and climate conditions in a field, generalizing the result and providing a decision support system (DSS) for actions such as real-time variation of fertilizer or pesticide application.

WSN must rely not only on intrusion prevention technology, but also on intrusion detection system (IDS). Although there have been many research results about intrusion detection, many technical problems have not been solved, because of the particularity of the WSN itself.

Karlof and Wagner pointed out that many WSN routing protocols are not considered in the current routing protocols [3], and the need for the security of all routing protocols to identify the specific target. They demonstrate how to successfully introduce WSN, the ad hoc and end-to-end network and put forward some safety measures related to other routing attacks, such as Sybil [4, 5], wormhole, selective forwarding and spoofed routing information. In addition to the two new attacks against sensor networks, namely Sinkhole attacks and HELLO message flooding attack, they provide detailed analysis of the threat of Sinkhole attacks [6] on sensor networks in all attack types.

Ngai et al. proposed a method of detecting Sinkhole attack [7] using the base station to judge the consistency of the data in a region by detecting abnormal data. In order to locate the malicious node, the base station sends request messages to the node. Then, the base station uses the received information to generate a network topology. However, the base station is often not very good in the area of data tampering or selective

✉ Fengjun Shang
shangfj@cqupt.edu.cn

[1] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

forwarding because of fluctuations in regional data and the environment, so this method will have false positives.

Krontiris et al. first described the conditions for IDS of WSN [8] and the number of normal node being greater than the malicious node. Furthermore, it proposed the basic architecture of distributed IDS and the method of detecting black hole attacks and selective forwarding attacks. Based on the above research, the technology of detecting Sinkhole attack is proposed.

Shafiei et al. proposed two methods to detect the Sinkhole attacks [9]. One is based on the geographical statistical sampling method, and an energy consumption model is used to detect the possibility of Sinkhole attacks in every area of the network, and then, the distributed monitoring method is set up. Another is based on mitigation strategies to prevent traffic flow when hijacked by Sinkhole attacks. Finally, the two methods are verified by Castalia.

Rajasegarar proposed an anomaly detection technology based on the distributed clustering algorithm and k-nearest neighbor (KNN), which is based on the hyper-sphere [10]. By using the collected information in the node local clustering, similarity identification and the one-hop parent node performing the clustering task, it finds abnormal data sets and reduces the energy consumption caused by the node communication. Moreover, this can be used for Sinkhole attacks detection. However, according to the analysis, when the node to launch the Sinkhole attacks creates abnormal data flow hijacking, the possibility of finding the attacks in the network is very small [11].

In this paper, an adaptive network anomaly detection model based on CUSUM_MV is constructed, which is composed of two parts: the anomaly detection engine based on neighbor node monitoring and the centralized Sinkhole recognition engine. And then this paper analyzes the performance of intrusion detection algorithm. The proposed model is based on the CUSUM_HDST. The model can reduce the extra communication cost caused by intrusion detection to the network.

2 CUSUM_MV model based on cumulative summation

Assume that the network contains a sensor node and a base station. Sensor nodes and base stations can be transmitted through a variety of appropriate communication protocols; the sensor node will pass the message to the base stations through multi-hop. Each node can monitor the entire network in real time by monitoring the behavior of neighbor nodes and detect the abnormal of the network in this way. In each time window, each sensor node constructs a feature vector, which is used to record the neighbor node behaviors and the related network conditions observed in the time window. This feature vector is composed of a fixed number of attributes.

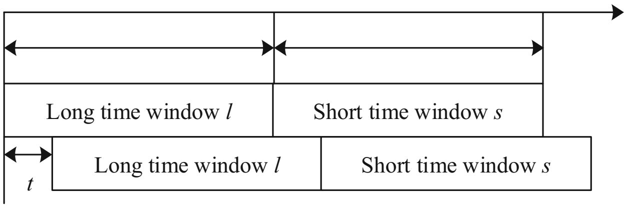


Fig. 1 Statistical time window forwarding

Due to the instability of the wireless channel in sensor networks, the different statistics contain noise, including signal conflict and environmental factors. So we need to perform a smooth processing of the statistics, such as with formula (2.1):

$$X_t^i = (1 - u_i)\,X_{t-1}^i + \mu X_{t'}^i, \quad 0 \le u_i \le 1 \tag{2.1}$$

where $X_{t'}^i$ is the observation value of the t' i node on that period in the network. X_t^i is the smoothed value. The value of the μ_i memory factor fluctuates with the value of different network conditions.

2.1 Anomaly detection of parent node based on CUSUM GLR

Cumulative summation algorithm has been widely used for network anomaly detection [12]. A sensor network is a data stream and network behavior is dynamic and stochastic, so the generalized likelihood ratio (GLR) and CUSUM are introduced to meet the needs of real-time monitoring of sensor networks. Under normal circumstances, the network behavior of adjacent nodes is basically similar. This paper selects the signal intensity RSSIij, C_{ij} and the link quality LQij to the base station to monitor the network behavior of the parent node. After the network enters the stable period, each node saves the RSSIij, the LQij of the parent node and the link quality C_{ij} to the intrusion detection module. Their expected value is:

$$E(RSSI_{N(i)}) = \varpi,\ E(C_{ii_p}) = \rho,\ E(\Delta_{lq}) = \vartheta,\ E_X = (\varpi, \rho, \vartheta)^{\mathrm{T}}$$

Based on the mean value estimation of sliding time window, the data of the long time window can detect the anomaly of the data in a short time. If there is no exception in the short time window, the two parameters (i.e., mean and variance) of the CUSUM GLR are estimated to move forward, and the value of forward moving is far shorter than that of the short time window. So the parameters of CUSUM GLR anomaly detection model can reflect the changes of the current network characteristics, as shown in Fig. 1.

For the variance of the computation, the Bessel standard deviation formula is used as follows.

$$\delta = \sqrt{\frac{1}{l-1}\sum_{i=1}^{l}(X_i - E(X))^2}$$

In order to reduce the cost of wireless sensor nodes, the unbiased range estimation is introduced to compute δ. First, long time window l, the maximum value $X_{\max}$ and minimum value $X_{\min}$ of the statistics are selected. So $R = X_{\max} - X_{\min}$. Dividing the data in time window l into three groups, the number of data is n in each group and the average value of each group is $\bar{R} = (R_1 + R_2 + R_3)/3$. Then, the calculation formula of the total standard deviation is estimated by using the principle of probability statistics, as shown in formula (2.2).

$$S = \frac{\overline{R}}{\sqrt{n}} \tag{2.2}$$

In the process of anomaly detection, the statistics can be expressed in the form of a vector: $X = (RSSI, C, LQ)$. It can also be expressed as a vector of a normal distribution.

In order to detect the anomaly dynamically, a long time window is required to estimate the process, as shown in formula (2.3).

$$E_X^{'} = \frac{1}{l}\sum_{i=k-l+1}^{k} X_i \tag{2.3}$$

If there is no exception, the current statistical value is updated, and the process reflects the adaptive mechanism of the anomaly detection. The process is shown in formula (2.4).

$$E(X_l^i) = \frac{1}{l}\sum_{j=l_0}^{m} X_l^i = \frac{1}{l}[(l-1)E(X_{l-1}^i) + X_l^i] \tag{2.4}$$

The mean vector E_x represents randomness before the exception occurs. After the exception occurs, the mean vector of the random statistic is $E_X^{'}$, and its log likelihood ratio is:

$$S_t = \ln\frac{p_{E_X^{'}}(X_t)}{p_{E_X}(X_t)} \tag{2.5}$$

Before the anomaly occurs, the value of S_t is negative. After the anomaly occurs, the value is positive. The value of S_n will continue to accumulate. When a given threshold is exceeded, an exception can be thrown. The decision rule can be given as shown in formula (2.6)

$$d = \begin{cases} H_0, & \text{if } \underset{n}{S} < \gamma \\ H_1, & \text{if } S_n \geq \gamma \end{cases} \tag{2.6}$$

The process of calculating in Fig. 5 is relatively inefficiency. It is basically consistent with the standard variance X of the statistical values. In order to reduce the number of calculation in the detection process, we can detect whether the standard deviation is too large in advance of any abnormality. Thus, the standard deviation can be expressed as δ. Furthermore, assuming X obeys the normal distribution, it can show $p_\theta(X) \sim N(E, \delta^2)$ shown in formula (2.7).

$$P_\theta(X) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{(X-E)^2}{2\delta^2}} \tag{2.7}$$

In the formula, when $\theta = \theta_0$, E is E_x. When $\theta = \theta_1$, E is E_X' t. Substituting into formula (2.5), it results in formula (2.8).

$$S_t = \frac{(E' - E)\cdot(2X - E' + E)}{2\,\delta^2} \tag{2.8}$$

And formula (2.9).

$$\begin{aligned} S_n &= \sum_{t=0}^{n} S_t \\ &= \frac{1}{2\,\delta^2}\sum_{T=0}^{n}(E' - E)\cdot(2X - E' + E) \end{aligned} \tag{2.9}$$

Assume that the vector $\gamma = \{\gamma_d, \gamma_{lq}, \gamma_c\}$ represents the anomaly detection alarm threshold for each statistic. For the detection threshold set, the bigger it is, the higher is the false-negative rate of the anomaly detection system. Furthermore, the longer the time to find the anomaly is, the longer the delay for alarm is. The smaller the threshold for anomaly detection system is, the higher the false alarm rate is. And then, sensor nodes have the greater burden. By analysis, a list of memory factor, long time l and short time S, and anomaly detection alert threshold $\gamma = \{\gamma_d, \gamma_{lq}, \gamma_c\}$ can be set up in the laboratory environment. And the selection of these thresholds can be accomplished by a lot of training before deploying nodes to monitor.

2.2 Anomaly information transfer

When a node finds an anomaly, it sends piggybacking packets to the base station. If it does not send packets for some time, this will generate an anomaly intrusion frame, called an IF packet, as shown in Fig. 2. In the anomaly region, the transmission path uses the historical parent node as the next hop node to avoid flooding which is mentioned by E. C. H. Ngai.

So, after the anomaly occurs, the anomaly information is still able to break through the anomaly regions. In order

Intrusion detection packet（IF packet）

6 Bytes	10Bytes	2Bytes	5Byters	5Bytes	2Bytes
PHY Header	MAC Header	LE Header	CTP Routing Frame	Intrusion detection frame	MAC Footer

Fig. 2 Modified CTP routing packet

0						8 15
P	C	D	LQ	C	RD	THL (Time Has Lived, Number of forwards)
ETX (Link quality From source node to base station)						
Origin (Source node number)						
SeqNo (Serial number of data packets)						CollectID (Receiving end of the protocol stack)
SuspectedID (Anomaly neighbor node collected from the source node)						
LinkEtx (The link quality from the source node to the parent node)						

Fig. 3 Modified CTP frame

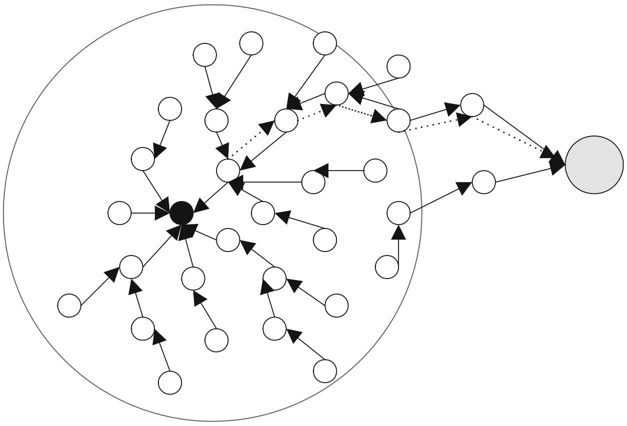


Fig. 4 IF routing diagram (dotted lines indicates the anomaly information transfer route, the arc shape in the region indicates the anomaly region, and the black circle represents an anomaly nodes)

to modify the data packets through multi-hop transmission to the base station, it may select the normal direction. The modified CTP data frame is shown in Fig. 3. The header from bit 2–7, totaling 6 bits, includes *D*, *LQ*, *C*, *RD*, where the RD domain occupies 2 bits (1 representing an anomaly and 0 indicating no abnormally) and where the RD domain is labeling the current node of the reverse, the maximum expressed 3°.

At the end of the CTP data frame, the area occupied by the 16-bit description of the suspected node intrusion detection is called SuspectedID. The 8-bit LinkEtx saves the single-hop link quality estimation. If the node detects a reverse link, it will send the data frame through the reverse of the RD domain to piggyback on the parent node. The transfer path is shown in Fig. 4.

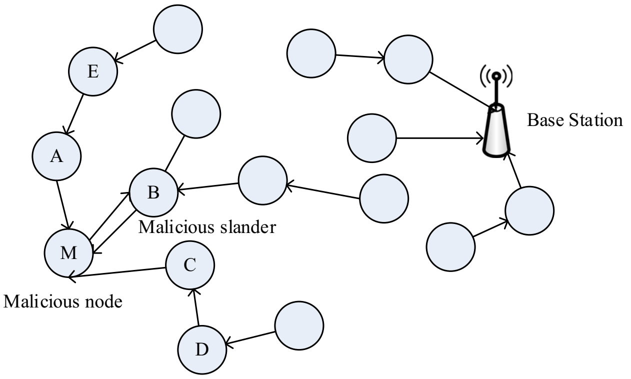


Fig. 5 Topology map obtained from base station

2.3 Sinkhole attacks detection algorithm

When the base station collects enough CTP data frames, *D*, *LQ*, *C*, *RD* and *SuspectedID* domain can be extracted from the intrusion detection system. They are used to construct the network behavior graph and route pattern of a certain area. If the node's anomaly detection works well, then the constructed graph can be regarded as a tree which is based on a Sinkhole, and thus identified as a Sinkhole attack. In Fig. 5, node M has launched a Sinkhole attack; the base station collects the anomaly information to constitute a tree, with node E and node D, for the malicious nodes. There are pointers to node M—A, B, C. There are also pointers to node B—D, E, M. The method of E. C. H. Ngai is analyzed, and the simplified intrusion detection process—putting forward the sink node link quality instead of hop number—is detected. The sinkhole attacks detection algorithm is based on majority voting, abbreviated as MV.

The detection technology principle of Sinkhole attacks is: If the node is a malicious node, because of the traffic aggregation effect of Sinkhole attack node, the node can be found in the suspicious region. Because the anomaly detection can generate false alerts, it is necessary to introduce a mechanism which is based on the network. If the node *a* is a father of the node b, the link quality from b to the base station is equal to the link quality from a to base station multi_Etx and the link quality from b to a link_ETX [13]. It is shown as formula (2.10).

$$multi\_ETX_a + link\_ETX_{(a,b)} = multi\_ETX_b \qquad (2.10)$$

Similarly, the process of identifying an attack node is transformed into a process of searching for the root node based on the link quality. If the root is a node in the anomaly region exceeds the detection index (setting the value of the detection intensity), then the node is considered to be the source of the Sinkhole attack. When malicious nodes testified the normal nodes, they will forge single-hop link quality information.

In the detection, if this is found not to conform to formula (2.10), the testifying is considered illegal to testify and ignore automatic. Figure 5 is an example. The detection algorithm iterates through the suspicious nodes. (In Fig. 5, they are M and B.). If the number of nodes is greater than the number exceeding ρ, an attack has occurred.

Definition 2.1 Node a is a legitimate testifying, that is, all the child nodes send an anomaly information frame of single-hop *linkETX* and multi-hop *multiETX* to satisfy formula (2.10).

Definition 2.2 Hypothesis mal_1 being the suspicious node, then mal_1^n is the number of suspect node mal_1. Namely, the node mal_1 is the total number of legitimate testifying all nodes and all its child nodes and mal_1 is one of the suspicious areas of multiple root nodes.

Because of the instability of the wireless channel [14], this will lead to the network anomaly link map. So it may be that in the process of finding an attack node, the link of RSSI is the lowest. In order to extract the useful link graph from the disrupted anomaly map, the abnormal link of the RSSI is removed from the graph with the exception of the dense region. If the total number of suspicious nodes is m, and the number N all nodes in the suspicious region is sent up, if it is established:

$$\max_{j\in m}(mal_j^n/N) > \rho \tag{2.11}$$

It is considered an attack, in order to avoid false alarm rate, if the proposed value ρ is set to be greater than 0.5.

3 CUSUM_HDST model based on D–S evidence theory

CUSUM_MV does not solve the other type of attacks, and it has a relatively large communication overhead in the network. In order to further improve the performance of the intrusion detection algorithm, the evidence theory (Dempster–Shafer) is introduced, which is also called the D–S evidence theory. Although the Bayesian network has been widely used for the classification of anomalies, the application of Bayesian networks requires the formation of a probability set and an abnormal distribution in advance. In contrast, evidence theory supports a reliability method implicitly embedded in system knowledge. And it does not need to clearly calculate the probability that it can be expressed with uncertainty in the cognitive domain and with no intellectual invention. Moreover, in the presence of uncertainty it does not require knowledge of the nature of decision making, so evidence theory is more suitable for carrying out the classification and detection of abnormalities.

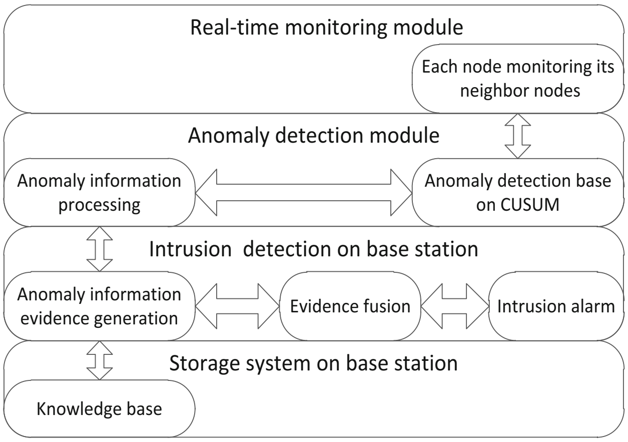

Real-time monitoring module
Each node monitoring its neighbor nodes
Anomaly detection module
Anomaly information processing
Anomaly detection base on CUSUM
Intrusion detection on base station
Anomaly information evidence generation
Evidence fusion
Intrusion alarm
Storage system on base station
Knowledge base

Fig. 6 Intrusion detection framework based on evidence theory

The efficiency of using statistical analysis and evidential reasoning to carry out the network anomaly diagnosis is studied in the paper [15]. In this paper, we first use the dual-loop auto-regression to model the increase in network monitoring variable values to detect the network anomaly accurately. In order to find out the causes of an abnormal occurrence, the evidence theory is used to combine all kinds of evidence. The results are verified by real data. The results show that the proposed method has higher classification efficiency. Accumulated evidence verifies that the evidence is in the right category, and it is not necessary to consult the class's estimates. A trust evaluation model is proposed [16] which is suitable for wireless network, and a trusted routing protocol is constructed based on AODV. Evidence theory is attractive because it is able to deal with uncertainty or incomplete knowledge (that is, the lack of a comprehensive probability model of knowledge). In the network environment, a host of reasons can lead to a variety of abnormalities, so evidence theory, which is based on an incomplete probability model, is more suitable for intrusion detection and network anomaly detection [17–19, 21–24].

CUSUM_HDST algorithm is a distributed and centralized intrusion detection system shown in Fig. 6. The information including abnormal detection, misuse detection and hybrid CUSUM_MV is used to reduce information about abnormalities, which can reduce the burden on the network prior to transmitting the exception information to the sink node. The sink node is used to identify the attacks.

3.1 Feature selection

In order to detect the DoS attack, we add a statistic, which is used to count the traffic information of the neighbor nodes, that is, Str_{ij}, the mean sending packets and receiving packets.

Definition 3.1 S_{ij}, the number of times node i observes the number of packets sent in node j.

Definition 3.2 R_{ij}, the number of times node i observes node j receiving the data packet, that is, the node i listens to the node j sending the confirmation ACK packet, because nodes in each receiving a data packet send a confirmation ACK packet, to confirm to the other side that the packet has been received. Str_{ij} means that node i observes the traffic of nodes j information, as defined as formula (3.1).

$$Str_{ij} = \left|(R_{ij} - S_{ij})\right| / \, R_{ij} \tag{3.1}$$

Under normal circumstances, the number of packets sent and received should be in balance; that is, the ratio should fluctuate in the vicinity of 0. From the formula, it can be seen that if R_{ij} is far greater than S_{ij}, this ratio is close to 1. When a node initiates a DoS attack, this ratio will be more than 1. Because the sensor network takes data as the center, it is hard to avoid the network congestion caused by the sudden network behavior, which makes the packet loss probability not stable. Smooth processing of packet loss rate is shown in formula (3.2).

$$Str^{i}_{ijt} = (1 - u_i)\, Str^{i}_{ijt-1} + \mu\, Str^{i}_{ijt'}, \quad 0 \le u_i \le 1 \tag{3.2}$$

$Str^{i}_{ijt'}$ is the observation value of the nodes i in the time t' and Str^{i}_{ijt} is the smoothed value. The value of the memory factor depends on the specific network environment.

3.2 Relay node anomaly handling

In the process of anomaly information transfer, the malicious slander node may broadcast false information packets to launch an attack disrupting the anomaly detection process. This section introduces the concept of fuzzy set theory taking certain node anomaly information as a domain. Accordingly, the various neighbors sending anomaly indication information will be a fuzzy set. It will use legitimate nodes testifying to legitimate nodes as belonging to a certain sample. The approach examines the relationship of the individuals testifying, which is used for calculating the degree of conflict between various indicators.

The conflict degree is constructed into a $n \times n$ matrix, in which the evidence is to be considered as a forgery of malicious information and is not transmitted. Thus, the communication of malicious information and redundant information is reduced. In Fig. 7, the node m is a malicious node. Node 1 has neighbor nodes 2, 3, 4, which are responsible for observing the behavior of node 1. In the network, the solid node m initiates intrusion resulting in anomaly network behavior. While the network is anomalous, nodes 1, 2, 3 will think that the node 1 is a slight anomaly because the node m broadcast of the anomaly report will mislead the node 1.

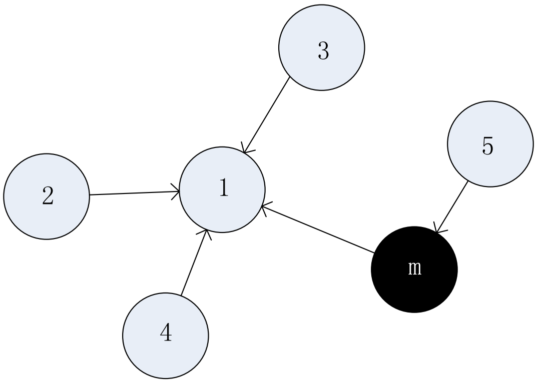


Fig. 7 Libel case

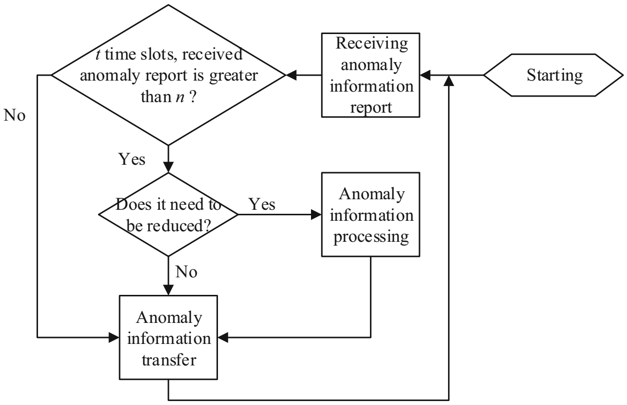

t time slots, received anomaly report is greater than *n* ?
Receiving anomaly information report
Starting
No
Yes
Does it need to be reduced?
Yes
Anomaly information processing
No
Anomaly information transfer

Fig. 8 Relay node work flow

When the relay node receives the information from the neighbor nodes, it needs to deal with the anomaly information and eliminate the conflicting information. In CUSUM_MV algorithm, according to the CUSUM GLR and the threshold, we get the anomaly information of neighbor nodes. This section uses this exception information for further processing. The CUSUM GLR model is used to detect the anomaly, and the anomaly in the short time window will be sent to the next hop node. In order to avoid being hijacked by a Sinkhole, the IF transfers uses the CUSUM_MV method. Its work flowchart is shown in Fig. 8.

The relay node i information received from the node j is a_{rssi}, a_{cn}, a_{lq} and a_{str}, respectively, and each of them is expressed as the anomaly degree of the nodes (signal intensity, convergence, link quality, traffic). As a result, the detection unit is based on a short time window, so the calculation method is shown in formula (3.3).

$$(|E_s - E_l|) / \, E_l \tag{3.3}$$

where E_l is expressed in the form of a long period (before the exception alarm) expectation. E_s expresses a short time window to detect the anomaly occurrence in the previous statistics. This ratio represents the extent of the exception, as an important basis for judging the occurrence of attacks in the network. Assuming the node f from different neighbor nodes receiving n vectors is the same node a, which can be expressed as a vector table of 3.4, then the node f will be responsible for obtaining n information from the anomaly, which may remove evidence of malicious information. Using fuzzy mathematics, node f can be assessed. We can calculate the degree of conflict between various anomaly reports.

$$f_n = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} = \begin{bmatrix} m_{11}^a & m_{12}^a & m_{13}^a & m_{14}^a \\ m_{21}^a & m_{22}^a & m_{23}^a & m_{24}^a \\ \vdots & \vdots & \vdots & \vdots \\ m_{n1}^a & m_{n2}^a & m_{n3}^a & m_{n4}^a \end{bmatrix} \tag{3.4}$$

The approach degree between the fuzzy sets of each neighbor node can be calculated using fuzzy mathematics. First of all, formula (3.4) is normalized as shown in formula (3.5).

$$f_i = \left(m_{i1}^a \Big/ \sum_{j=1}^{n} m_{j1}^a, m_{i2}^a \Big/ \sum_{j=1}^{n} m_{j2}^a, m_{i3}^a \Big/ \sum_{j=1}^{n} m_{j3}^a, m_{i4}^a \Big/ \sum_{j=1}^{n} m_{j4}^a \right) \tag{3.5}$$

Obviously, formula (3.4) is normalized; this will not affect the degree of divorce between each vector. Then, the weighted Euclidean distance formula (3.6) is used to calculate the approach degree:

$$d(m_b^a, m_c^a) = \sqrt{\sum_{i=1}^{4} \zeta_i (m_{bi}^a, m_{ci}^a)} \tag{3.6}$$

In the formula, ζ_1, ζ_2, ζ_3 and ζ_4 express the signal intensity, the degree of convergence, the link quality and the weight of the flow, respectively. For example, when detecting a DoS attack, the assignment ζ_4 should be greater than the other three values. After calculating the distance, the matrix can be obtained shown in formula (3.7).

$$\begin{vmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{vmatrix}_{n \times n} \tag{3.7}$$

In formula (3.7), d_{ij} expresses the approach degree. Approach degree d_{ij} is 0 indicating no conflict, that is, the anomaly report is completely consistent. On the contrary, the larger d_{ij} is, the greater the conflict between the two reports is.

In order to filter out the impact of final evidence, the threshold value ρ_i is defined. It denotes the conflict degree of node i. ρ_{i_0} denotes the conflict proportion of node i_o. If it exceeds a certain threshold, the report blocking anomaly detection of the malicious slander report should be removed; the relay nodes will not transmit them. Conversely, less than ρ_i of the threshold should be classified as legitimate to testify the fuzzy set. The calculation method is shown in formula (3.8).

$$\rho_{i_0} = \frac{\sum_{i=i_0, j=1}^{n} d_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}} \tag{3.8}$$

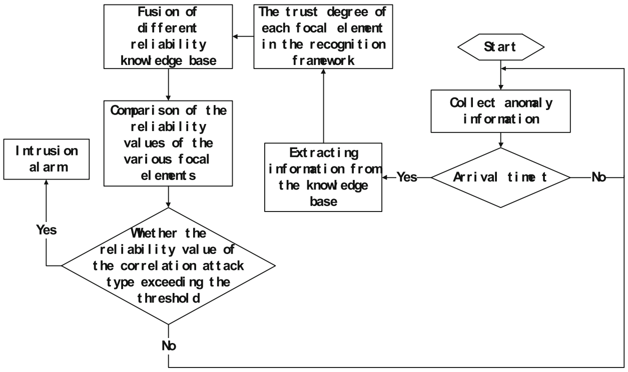

Start
Collect anomaly information
Arrival time t
Yes
No
Extracting information from the knowledge base
The trust degree of each focal element in the recognition framework
Fusion of different reliability knowledge base
Comparison of the reliability values of the various focal elements
Whether the reliability value of the correlation attack type exceeding the threshold
Yes
No
Intrusion alarm

Fig. 9 Base station detection process

3.3 Information fusion and attack judgment on the base station

Compared with Bayesian theory, the fusion of evidence theory does not require a priori knowledge of probability, and it is suitable for intrusion detection in dynamic changes. To introduce the concept of DST (D–S evidence theory), we first need to establish a sound prior knowledge about network failure or anomaly and list the hypothesis and the corresponding evidence. Through these hypotheses and evidence, we can determine the most likely causes of the current network anomalies. Evidence can be accumulated and calculated to determine the most likely categories of current abnormal information, so as to identify whether there is a network attack. Work flow on the base station is shown in Fig. 9.

On the basis of CUSUM_MV algorithm, the improved model is used to detect Sinkhole attacks, and it is also used to detect DoS attacks. For the detection of various attacks, we first need to define a recognition framework $F = \{F_1, F_2, \ldots, F_3\}$ and its relation to the knowledge base, where each element F_i represents a detection of intrusion attack category. In particular, F denotes a non-anomaly. Normal value is non-attack in this category.

In order to identify the attack in the current network, we first need to construct a knowledge set as the knowledge base for the attack detection. We will show that the knowledge set is represented as D_n (φ_n, f_n), where φ_n is the four-dimensional vector representing the degree of the different degrees of the individual statistics and f_n denotes the type of attack. Responding to the changing needs of detection, we can constantly expand the attack or exception types in the evolving knowledge base to improve the detection rate and expand the detection range.

The base station collects anomaly information sent from nodes. Each sub-node, in the t time period, can be expressed as a vector $\varphi(t) = \left[\varphi^1(t), \varphi^2(t), \ldots, \varphi^m(t)\right]$, which includes all the anomaly information in the network and the status of the current sensor network. M denotes the number of statistics. In order to construct the BPA, the Euclidean distance formula is introduced as shown in formula (3.9).

$$d(\hat{\psi}, \psi) = \sqrt{\sum_{j=1}^{m} \hat{\psi}^j - \hat{\psi}_i^j} \tag{3.9}$$

It is used to calculate the distance between two vectors. It can represent the degree of similarity between the vectors ψ and the current statistics $\hat{\psi}$. And the degree of trust is used to calculate the information generated in each of the exceptions and a knowledge set. With the increase in the distance of two vectors, we believe that the vector ψ and the network exception vector $\hat{\psi}$ are the same probability for the same class.

An efficient BPA should reply on $d(\hat{\psi}, \psi)$. It should reflect the relationship between the type of attack and the abnormal information vector, so that the BPA function [19] is introduced to generate the belief values, as shown in formula (3.10).

$$m_{\varsigma i_0}(A) = \begin{cases} p_i^{(l_{i_0})} & A = \{F_{l_{i_0}}\} l_i = l, \ldots, M \\ 1 - \ p_{i_0}^{(l_{i_0})} & A = F \\ 0 & A \in \{\{F_{l_{i, i \neq i_0}}\}, F\} \end{cases} \tag{3.10}$$

Formula $p_i^{(l_i)} = \alpha \, e^{-rd\,(\tilde{\phi}, \phi_i)^2}$, where $0 < \alpha < 1$ and $\gamma > 0$, expresses the distance between two vectors using anomaly information vector category judgments provided by the trust.

In order to avoid errors in evidence fusion, the rest trust degree averagely allocates a recognition framework for the rest of the class. In many cases, the same type of attack is expressed on multiple statistics, so a statistic and a class of attacks cannot be a good fit, while an exception vector for multiple attack types distribute the nonzero values. Formula $1 - p_{i0}^{(l_{i0})}$ expresses a degree of uncertainty, and the default is normal.

In order to detect the occurrence of intrusion attacks in the network, we need to focus on the knowledge of all the vectors in the type of attack included. Each evidence of ξ_i is not only one, but also generally the identification of the framework of multiple sets of processes shown in Fig. 9. Thus, we can use the results obtained by the fusion of evidence to identify the abnormal types. Therefore, in order to obtain the total BPA of each focal element, the evidence combination rule $m^{(N)}(A) = \oplus_{i=1}^{N} m_{\xi}(A)$ is used to obtain the BPA, where $A \in F$. For each type of attack set threshold, the detection model can reflect the changes of the current network characteristic, as shown in Fig. 1.

Table 1 Simulation environment parameter configuration

Parameter	Value
Simulation scene	200×300 m^2
Number of nodes	100
Rate of sending data in application layer	0.33 bit/s
Routing protocol of network layer	CTP
MAC layer protocol	CC2420
Physical layer protocol parameters	cc2440.txt
Malicious neighbor node list size	10
Malicious node routing beacon sending frequency	1
Malicious node radio layer sensitivity	− 90 dBm

4 Simulation and analysis

The simulation uses a Castalia simulator. The configuration of nodes needs to be done, and there are some simulation parameters, such as simulation scene, number of node and rate of wending data in. They are shown in Table 1. Attack nodes are deployed randomly in the network environment, and the Sinkhole and DoS attacks are launched after a period of time. The simulation program was run 10 times in various settings. Based on the analysis of the above research, for the detection of Sinkhole attacks, the base station periodically broadcasted information frames to the network. When the node receives the information frame, the local detection results were sent to the base station. To determine attack type, relay node detected the anomaly information. The base station generated evidence and fused evidence to the attack information.

The model was analyzed in terms of the detection rate, false alarm rate and the additional burden on the network. Due to the limitation of space, this paper analyzes the LQ of the nodes near the Sinkhole attack. As shown in Fig. 10, the horizontal coordinates of the graph represent the number of hops to the attack node, and the vertical coordinates are expressed as the average degree of the link quality changes of the nodes with different hops. The formula for calculating the degree of variation is (4.1).

$$\left[\left(\sum_{i=hop, j=0}^{j<n} \left| Etx_S^{i,j} - Etx_N^{i,j} \right| / \, Etx_N^{i,j}\right) / n\right] \times 100 \tag{4.1}$$

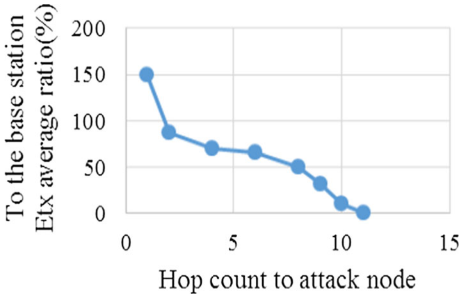


Fig. 10 Impact of attack on link quality

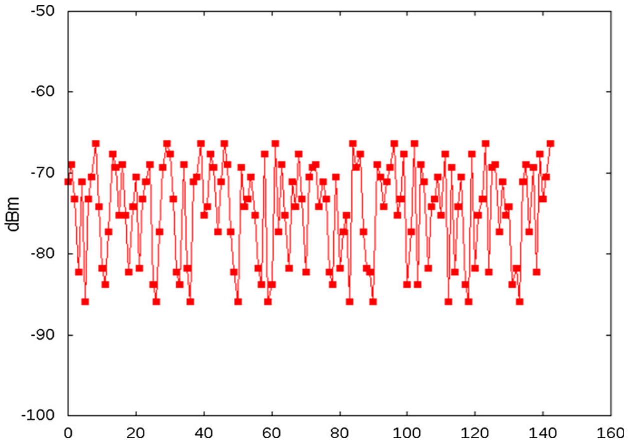


Fig. 11 RSSI distributions under normal circumstances

In formula (4.1), n is the number of attack nodes for hop count. Etx_S^j and Etx_N^j are the link quality of the base station before and after the node j changes. It can be clearly seen that the link quality of the node to attack node has a dramatic change and the change of the link quality is not obvious with the increase in the distance from the point of attack. After a period of time, the Sinkhole node had launched the attack, resulting in the change of the LQ in the base station. Attack nodes were launched by the attack, which was broadcasted by the route beacon frame. In the beacon frame, it pretended that link quality to the base station is high and the frequency of the transmitting beacon frames increased. The surrounding neighbor nodes will response to attack node, parent node selection, which ultimately led to the illusion of region near to the base station link quality rise. In addition, it can be known that the change of convergence degree is consistent with the change of LQ; that is, when a node to the base station has a high quality; it is bound to increase its attraction to the surrounding nodes. Figure 11 shows the RSSI of a node's parent node under normal circumstances. The horizontal coordinates are expressed in time sequence. The vertical coordinate is the corresponding time point. The RSSI value of the node is collected. It is clear that the RSSI of the node is always a normal distribution.

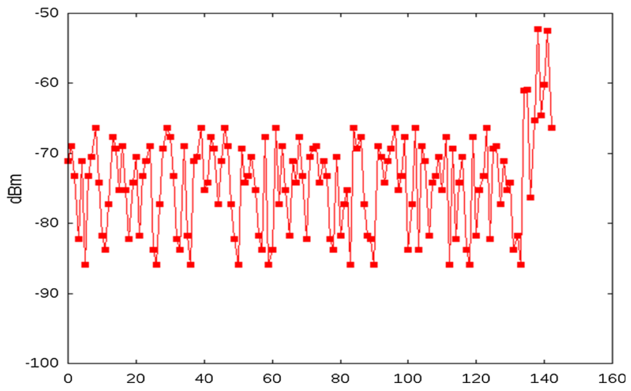


Fig. 12 Anomaly under attacking

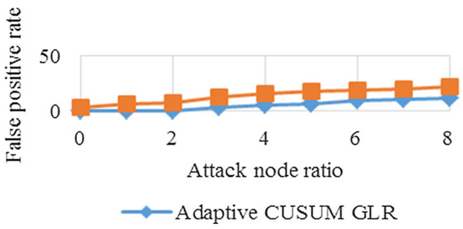


Fig. 13 False alarm rate changing with attack strength

As shown in Fig. 12, when a Sinkhole attack occurs, the attack node may be enhanced by the emission power, and the 138th point in the picture shows a sudden change in the power. In order to respond to a range of attacks, it is easy to use CUSUM GLR algorithm to detect the occurrence of anomaly time points and then quickly detect the Sinkhole attacks in the network.

Usually, the intrusion detection algorithm evaluation index was the detection rate and the false-positive rate for every simulation results. We need the false alarm rate and false-negative rate calculated. When a normal behavior is labeled as an anomaly, it is a false alarm. When an anomalous behavior is labeled as normal, it is false negative. The false-positive rate is the ratio of the number of false positives and the number of actual measurements. The false-negative rate is negative and the actual amount of the abnormal ratio.

The CUSUM_MV anomaly detection model introduces an adaptive mechanism, which is shown in Fig. 13 with respect to the false-positive rate of the anomaly detection with respect to the ordinary CUSUM GLR. False positives are the behavior of an abnormal alarm in the area of the attack. The relative rate of false positive refers to the number of false alarms as a proportion of the total number of nodes. The false-positive rate is a key factor to the detection rate, because if the false-positive rate is too high, the additional communication overhead is increased.

In the experimental environment, the convergence degree CN, link quality LQ and the RSSI of the parent node were detected. The adaptive and predictive methods were used to

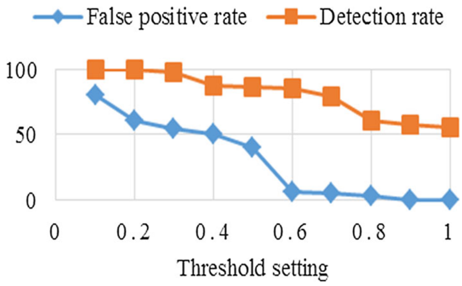


Fig. 14 Detection performance with the change of detection threshold

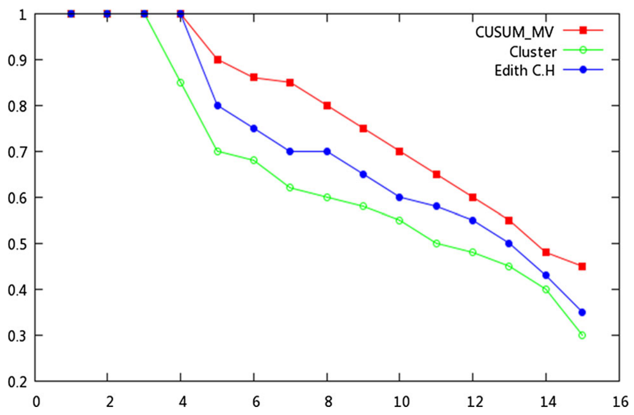


Fig. 15 Comparison of detection rate of intrusion detection model

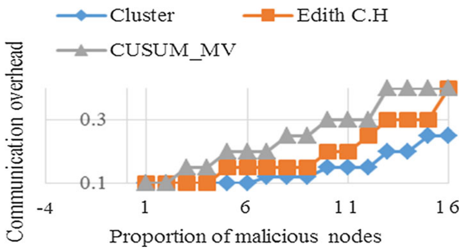


Fig. 16 Comparison of the communication burden of intrusion detection model

detect the three variables. The long time window was set to 60 s, the short time window was set to 10 s, and the time window movement was every 1 s. If one of the variables was an anomaly, the current node might have an exception based on judging whether the node is abnormal or not. Simulation results are shown in Fig. 13. From the graph, it can be seen that, in the same condition, the CSUSUM GLR anomaly detection model with the introduction of adaptive mechanism was significantly lower than that of low false-positive rate. Intuitively, CUSUM_MV attack detection model was only the most suitable ρ to achieve the highest detection effect.

In the experimental scene, with the random deployment of the 10 Sinkhole attacks, the false alarm rate and the detection rate of change with the value ρ, the trend is shown in Fig. 14. From the figure, it can be seen that the detection rate gradually increased. When the detection rate reached a certain level, the false-positive rate also increased, while the detection rate remained unchanged. Therefore, the intrusion detection module was deployed to the actual scene situations based on past data. In order to give the appropriate value, the ρ value needs training. From the figure, it can be seen that the value is 0.6. The false-positive rate and the detection rate achieved a better mutual balance with this value.

Comparing the algorithm proposed by E. C. H. Ngai, the hyper-sphere distributed clustering algorithm proposed by Sutharshan Rajasegarar and CUSUM_MV in the same network scenario, the relationship between the detection rate of each model and the change of malicious nodes is shown Fig. 15.

When the Sinkhole attacks, the attack node will be information, resulting in information being hijacked. The base station cannot receive effective information. In the cluster-based method, the so-called information transfer mechanism is not used; the detection rate is very low. The defects of the E. C. H. Ngai et al. proposed methods are mitigated. And link quality is a testimony standard, so as to improve the detection rate of the proposed a method of E. C. H.

With the increase in attack nodes, the detection rate of each detection method will fall. Because the anomalies are too much, there is message conflict. Moreover, the corresponding information cannot be transmitted to the base station; the base station is thus not accurate. With the increase in malicious nodes in the network, it is more difficult to detect the attacker, especially when the number of malicious nodes reaches a certain ratio. The network has been caught in a state of non-work. Obviously, with the increase in malicious nodes in the network, the resolution of the malicious nodes and normal nodes is more difficult, and the false-positive rate will also rise.

In terms of energy consumption, because the signal intensity of the set of nodes is certain, the energy consumption of the individual node is certain. If the node needs to maintain the network traffic information in the hybrid mode, the energy consumption is constant [20]. In this paper, we only consider the additional communication burden caused by the intrusion detection module, which is only a proportion of the communication packets in the network. According to E. C. H. Ngai's approach, the method is based on the base station. The proposed method is based on the node and transmitting the data to the base station. Thus, the communication relatively load is relatively low and the communication cost is relatively low. Figure 16 shows a comparison of the communication burden of intrusion detection models.

From Fig. 17, it can be seen that the detection rate is equivalent to the CUSUM_MV algorithm, and the stability of CUSUM_HDST detection model is better with the increase in attack nodes. In the experimental scene, the detection rate of the CUSUM_MV and the CUSUM_HDST is the same

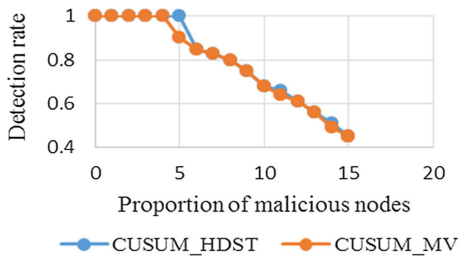


Fig. 17 Detection rate changes with the proportion of malicious nodes

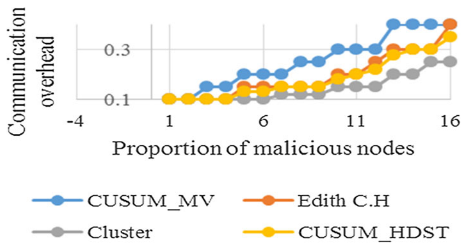


Fig. 18 Communication overhead with the change of the proportion of malicious nodes

when the proportion of malicious nodes less than 5. But until 5, the detection rate of the CUSUM_MV keeps 1 and the detection rate of the CUSUM_ HDST decline to 0.9. They both synchronously decline between 5 and 10, and they do not have changes until 15. It can be seen that the detection rate is equivalent to the CUSUM_MV algorithm, and the stability of CUSUM_HDST detection model is better with the increase in attack nodes.

From Fig. 18, it can be seen that the false-positive rate of CUSUM_HDST detection model is stable and has no dramatic change.

5 Conclusions

In this paper, we study the anomaly detection behavior of the nodes and the base station.

1. When the station cannot capture the network anomalies, the CUSUM GLR is introduced, and the anomaly detection model is given;
2. Sinkhole hijack traffic, and the mechanism of transmission of the anomaly information to the base station, is given in view of the Sinkhole attack nodes;
3. Based on the "link quality" and "majority rule," a new Sinkhole attack detection scheme is proposed, and a CUSUM_MV intrusion detection model based on node and base station communication is presented.
4. Based on the Castalia simulation experiments, the results showed that the CUSUM_MV intrusion detection model has a better performance than traditional methods in detecting Sinkhole attacks. The detection rate is improved, and the false-positive rate is reduced;
5. Based on weighted Euclidean distance, the redundant information removal mechanisms are established on the relay nodes. In order to reduce the communication overhead caused by intrusion detection, evidence theory is applied to the detection of wireless sensor networks. Based on node and base station, the CUSUM_HDST intrusion detection model is given;
6. Simulation experiments based on Castalia show that the CUSUM_HDST intrusion detection model can not only detect Sinkhole and DoS attacks, but also reduce the communication overhead caused by intrusion detection.

Acknowledgements The work has been supported by the National Natural Science Foundation of China (No. 61672004), the Chongqing Research Program of Basic Research and Frontier Technology under Grant No. cstc2016jcyjA0590, and the CERNET Innovation Project. The author would like to thank the Program for Innovation Team Building at Institutions of Higher Education in Chongqing (CXTDX201601021) and Chongqing Municipal Engineering Research Center of Institutions of Higher Education.

References

[1] Hodge, V.J., O'Keefe, S., Weeks, M., Moulds, A.: Wireless sensor network for condition monitoring in the railway industry: a survey. IEEE Trans. Intell. Transp. Syst. **16**(3), 1088–1105 (2015)

[2] Fouchal, H., Hunel, P., Ramassamy, C.: Towards efficient deployment of wireless sensor networks. Secur. Comm. Netw. **9**(17), 3927–3943 (2016)

[3] Karlof, D.W.: Secure routing in wireless sensor networks: attacks and countermeasures. Ad Hoc Netw. J. Special Issue Sens. Netw. Appl. Protoc. **8**(3), 293–315 (2003)

[4] Jan, M.A., Nanda, P., He, X., Liu, R.P.: A Sybil attack detection scheme for a forest wildfire monitoring application. Fut. Gener. Comput. Syst. **80**, 613–626 (2018)

[5] Bhise, A.M., Kamble, S.D.: Review on detection and mitigation of sybil attack in the network. Procedia Comput. Sci. **78**, 395–401 (2016)

[6] Yadav, H., Tak, M.S.: A surevy on detection of sinkhole attack in wireless sensor network. Int. J. Eng. Techn. Res. **V6**, (11) (2017)

[7] Ngai, E.C.H., Liu, J.C., Lyu, M.R.: An efficient intruder detection algorithm against Sinkhole attacks in wireless sensor networks. Comput. Commun. **12**(30), 2353–2364 (2007)

[8] Krontiris, I., Benenson, Z., Giannetsos, T., Dimitriou, T., et al.: Cooperative intrusion detection in wireless sensor networks. In: Roedig, U., Screenan, C.J. (Eds.) EWSN, pp. 263–278 (2009)

[9] Shafiei, H., Khonsari, A., Derakhshi, H., et al.: Detection and mitigation of sinkhole attacks in wireless sensor networks. J. Comput. Syst. Sci. **12**(1), 12–22 (2013)

[10] Rajasegarar, S., Leckie, C., Palaniswami, M.: Hyperspherical cluster based distributed anomaly detection in wireless sensor networks. J. Parallel Distrib. Comput. **74**(1), 1833–1847 (2014)

[11] Fessant, F.L., Papadimitriou, A., Viana, A.C., et al.: A Sinkhole resilient protocol for wireless sensor networks: performance and security analysis. Comput. Commun. **12**(35), 234–248 (2012)

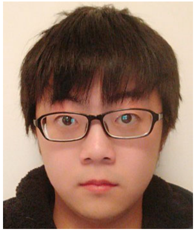
[12] Zhao, H.: The simulation experiment and research on an improved cumulative sum anomaly detection method. Appl. Mech. Mater. **743**(38), 219–225 (2015)

[13] Ozcelik, M.M., Irmak, E., Ozdemir, S.: A hybrid trust based intrusion detection system for wireless sensor networks. In: International Symposium on Networks, Computers and Communications. IEEE, pp. 1–6 (2017)
[14] Sun, Y., Zhang, Y.: New developments of characteristic analysis in wireless sensor networks. IETE J. Res. **2**, 221–227 (2016)
[15] Zang, T., Yun, X., Zhang, Y., Men, C., Cui, X.: Botnets' similarity analysis based on communication features and D–S evidence theory. J. Commun. **32**(4), 66–76 (2011)
[16] Yang, K., Ma, J., Yang, C.: Trusted routing based on D–S evidence theory in wireless mesh network. J. Commun. **32**(5), 89–103 (2011)
[17] Zhao, X., Liu, Y., Sun, J.: New network anomaly detection using transfer learning and D–S theory. Appl. Res. Comput. **33**(4), 1137–1140 (2016)
[18] Chen, Y., Liu, Y.: Application of extended D–S evidence theory in intrusion detection. Comput. Eng. Sci. **36**(1), 83–87 (2014)
[19] Chang, Y., Liu, F.: Wireless sensor intrusion detection system based on the theory of evidence. In: IEEE International Conference on Communication Software and Networks, pp. 2811–2814. IEEE (2016)
[20] Super User: Wireless Sensor Network Simulator User Manual. NICTA, Australia (2013)
[21] Song, X., Wang, C., Gao, J., Xi, H.: DLRDG: distributed linear regression-based hierarchical data gathering framework in wireless sensor network. Neural Comput. Appl. **23**(7–8), 1999–2013 (2013)
[22] Bacciu, D.: Unsupervised feature selection for sensor time-series in pervasive computing applications. Neural Comput. Appl. **27**(5), 1077–1091 (2016)
[23] Wang, G., Huang, C.: Energy-efficient beaconless real-time routing protocol for wireless sensor networks. Comput. Syst. Sci. Eng. **26**(3) (2011)
[24] Zhang, D.G., Zhou, S., Chen, J.: New Dv-distance method based on path for wireless sensor network. Intell. Autom. Soft Comput. **23**(2), 219–225 (2017)

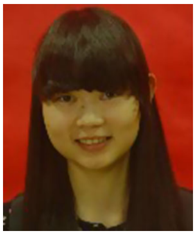
Fengjun Shang male, finished his Ph.D. degrees in Instrument Science and Technology at the College of Opto-electronic Engineering, Chongqing University, China, in 2005. Since then, he works at the Institute of Computer Network Engineer in Chongqing University of Posts and Telecommunications, China. He was a visiting scholar in University of Wollongong, Australia, from November 2007 to November 2008. His research interests include sensor network, future Internet, IOT, network optimization and cloud computing.

Dan Zhou is a candidate master's degree from Chongqing University of Posts and Telecommunications, and her main research direction is the network probes.

Cheng Li is a master's degree from Chongqing University of Posts and Telecommunications, and his main research direction is the network security for wireless sensor networks.



Hanyun Ye is a candidate bachelor's degree from Chongqing University of Posts and Telecommunications, and his main research direction is the network measurement.



Yuting Zhao is a candidate bachelor's degree from Chongqing University of Posts and Telecommunications, and her main research direction is the cloud computing.