An Intrusion Detection Model for Wireless Sensor Network Based on Information Gain Ratio and Bagging Algorithm

Rui-Hong Dong, Hou-Hua Yan, and Qiu-Yu Zhang
(*Corresponding author:Qiu-Yu Zhang*)

School of Computer and Communication,Lanzhou University of Technology
No.287,Lan-Gong-Ping Road,Lanzhou 730050,China
(Email: zhangqylz@163.com)

(Received Aug. 5, 2018; Revised and Accepted Mar. 23, 2019; First Online June 11, 2019)

Abstract

Aiming at the problem that the dimension of the traffic data to be processed in the wireless sensor network (WSN) intrusion detection method is too high, which leads to the large amounts of computational complexity of the intrusion detection model and the weak detection performance of the intrusion behavior. Using the principle of ensemble learning algorithm, an intrusion detection model for WSN based on information gain ratio and Bagging algorithm was proposed. Firstly, the information gain ratio method is used to select the feature of sensor node traffic data in this model. Secondly, the Bagging algorithm is used to construct an ensemble classifier so as to train multiple C4.5 decision trees which are improved. The parameters of the ensemble classifier are optimized through 10 iterations, and the dynamic pruning process is introduced. Finally, the classification results of C4.5 decision tree are classified and detected by majority voting mechanism. The experimental results show that compared with the existing intrusion detection methods, the proposed model has higher detection accuracy for Blackhole, Grayhole, Flooding, Scheduling and other intrusion attacks. While ensuring the true positive rate of 99.4%, it can still maintain a low false positive rate and high detection performance for intrusions behavior.

Keywords: Bagging Algorithm; Ensemble Classifier; Intrusion Detection; Information Gain Ratio; Wireless Sensor Network (WSN)

1 Introduction

With the wide application of WSN in smart cities, smart grids, environmental monitoring, medical sensing, industrial and other fields [12], it also brings some security issues such as network attacks and intrusions. Due to the wireless transmission and unattended characteristics of WSN, the sensor node has limited energy, storage capacity and computing power, which makes it vulnerable to various malicious attacks, such as Wormhole, Sinkholes, Greyhole, and Flooding and so on. These typical attacks all cause the network traffic to deviate from the normal network traffic, which will bring great harm to the WSN in a short time. Therefore, as an important technical means of network security, WSN network intrusion detection technology has attracted wide attention from scholars [26].

At present, WSN intrusion detection is mainly divided into anomaly detection, misuse detection, specification-based detection and hybrid system detection [15]. The existing intrusion detection methods mainly include: support vector machine [7, 20], artificial neural network [2, 9], Naive Bayes [10], Bayesian Network [23], decision tree [22], random forest [14], artificial immunity [8], random weight neural network [25] and other methods. For example, in [7], a hybrid method of support vector machine and genetic algorithm was proposed. The genetic algorithm was used to select the feature subset from the original feature set, and SVM was used as the classifier for intrusion detection. The method obtained 97.3% detection rate. However, the detection efficiency of unknown attacks is not efficient.

In [20], an intrusion detection system based on SVM and principal component analysis (PCA) was proposed. For KDDcup99 data, PCA combined with SVM algorithm was used for intrusion detection. This method reduces data analysis time and improves intrusion detection performance, but it cannot identify the different types of attacks. In [9], a back propagation learning algorithm was proposed to optimize the back propagation neural network (BPNN) intrusion detection system. For the KDDcup99 data, it has higher detection rate and lower false detection rate, but the algorithm complexity is higher.

In [23], an intrusion detection method based on ensemble learning was proposed. By Using the KDDcup99 data, the Bayesian network and the random tree were first used as the base classifier for voting classification, and then identify if an attack has occurred. The algorithm as a whole has high detection efficiency, but the accuracy of U2R attacks was low.

In [14], a lightweight intrusion detection system based on decision tree was established, which improves the detection rate and reduces the complexity of the algorithm, but it does not detect unknown attacks. In [8], it compared the performance of supervised machine learning classifiers, proving that the detection performance of random forests is the best. In [25], an improved clonal selection algorithm was proposed. By selecting the best individual and cloning to detect the intrusion behavior, it was proved that the proposed artificial immune method is better than the artificial neural network.

In [6], a semi-supervised learning method based on fuzziness was proposed. The unlabeled sample was combined with the supervised learning algorithm to optimize the performance of the classifier. The random weight neural network was used as the base classifier to improve the classification ability. However, only two types of tasks can be detected, and multiple attacks cannot be detected.

In [21], considering the characteristics of wireless sensor networks, a detection model based on clustering mutual coordination was proposed. The intrusion detection rate was enhanced and the false detection rate was reduced. However, it is complicated to update the CA-AFSA-BP system during the detection process. And the detection rate of unknown attacks is not high.

In [24], a two-level feature selection method based on SVM was proposed. Fisher and information gain were used to filter noise and irrelevant features respectively in the filtering mode. By reducing the feature dimension, the modeling time and testing time of the system were reduced. However, when the number of training samples increases, the system overhead is large, and the classification detection performance is not high.

In [18], a cluster network intrusion detection system was proposed. Each node calculates the reputation value according to the behavior of observing neighbor nodes. The base station detects the malicious nodes by combining the reputation value and the misuse detection rules. However, because the reputation value calculation method has a great influence on the detection rate, which leads to the excessive dependence on the reputation value calculation method.

In [13], in order to solve the problem of dimension hazard in high-dimensional feature space, a SVM intrusion detection system based on self-encoding network was proposed, which is suitable for high-dimensional spatial information extraction tasks, and it can also reduce the intrusion detection model classification training time and test time. It satisfies the real-time requirements of intrusion detection, but the detection performance of R2L, U2R and other attack behaviors is not high.

In [4], a special WSN data set was developed, and the collected data set is called WSN-DS. It can help researchers better detect and classify WSN's four types of denial of service (DoS) attacks, including Blackhole, Grayhole, Flooding, and Scheduling. The data set is used to train the artificial neural network (ANN) to detect and classify different attacks. By analyzing the above research work, the existing WSN intrusion detection method generally has a large computational load, and the dimension of the traffic data to be processed is too high, which cannot effectively detect multiple attack types and the detection efficiency of unknown attacks is low.

In [17], a novel approach called SCDNN for sensor network intrusion detection was proposed, which combines spectral clustering (SC) and deep neural network (DNN) algorithms. es an effective tool of study. The algorithm has a strong ability of sparse attack classification and effectively improves the detection accuracy of the actual security system. However, the limitations of SCDNN are that its weight parameters and the threshold of each DNN layer need to be optimized, and the k and s parameters of the cluster are determined by experience, rather than by mathematical theory.

In [16],a localization attack recognition method using a deep learning architecture was proposed, by learning the positional and topological feature based on SDA-based deep architecture, the classification accuracy can be significantly improved, but the time complexity and space complexity are relatively large.

In [3], a novel intrusion detection system based on neuro-fuzzy classifier in binary form for packet dropping attack in ad hoc networks was proposed. Simulation results show that efficiently detect the packet dropping attack with high true positive rate and low false positive rate.

Aiming at the shortcomings of the above research, this paper proposes a WSN intrusion detection model based on information gain ratio and Bagging algorithm. The model uses feature gain ratios for feature selection and reduces feature dimensions by removing extraneous features. The Bagging algorithm is used to construct an ensemble classifier to train the improved C4.5 decision tree, and the parameters of the C4.5 decision tree are optimized by multiple iterations to improve the classification accuracy of the classifier. A majority voting mechanism is used for the classification results to detect intrusion behavior. The experimental results show that the model can identify different types of attacks. Compared with the existing intrusion detection methods, the detection accuracy is improved, and many types of attacks can be detected.

The remaining part of this paper is organized as follows. Section 2 introduces related theory, including WSN network topology, feature selection and ensemble theory. Section 3 describes in detail the specific implementation process of the proposed WSN intrusion detection model in this paper. Section 4 gives the experimental results and performance analysis as compared with other related methods. Finally, we conclude our paper in Section 5.

2 Related Theory

2.1 WSN Network Topology

WSN mainly has three kinds of network topologies, which are divided into plane structure, cluster based structure and hierarchical structure [19], as shown in Figure 1. The WSN consists of three parts: Sensor nodes, cluster head nodes and base station. Sensor nodes are used to monitor the target area and collect data from the area. These nodes are arranged in respective clusters, and the sensed data is simply processed and transmitted to the cluster head node. The cluster head nodes collect and process the sensor node data in the cluster and transmit it to the base station. The cluster head nodes in the base station management scope can monitor the behavior of the cluster head nodes in real time, and the intrusion detection model can be deployed to the base station. When the base station receives the traffic data from the cluster head node, each piece of data is processed, and the intrusion detection model is used to determine whether an attack behavior has occurred in the WSN.

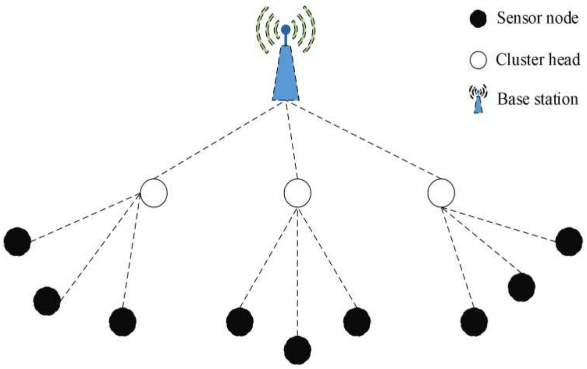

Figure 1: WSN network topology

2.2 Feature Selection

In the WSN, the traffic data dimension is high, some traffic characteristics are not related to the intrusion attack, and the node resources in the WSN are limited. Therefore, the WSN intrusion detection system introduces data preprocessing methods such as feature selection and data dimensionality reduction to remove irrelevant features and reduce the computational load of the intrusion detection method and enhance the intrusion detection efficiency.

The information gain ratio [1] is a feature selection method based on information theory, the specific definitions are as follows:

Definition 1. *Information entropy: The information entropy of a random variable is used to measure the degree of redundancy of the variable. Suppose that in a classification system, C indicates that the category is divided into c_1, c_2, ..., c_n, n represents the total number of classifications. Then the information entropy $H(C)$ of the classification system is defined as follows:*

$$H(C) = -\sum_{i=1}^{n} P(c_i)logP(c_i), \tag{1}$$

where $P(c_i)$ is the probability of the category $c_i(1 \leq i \leq n)$ at different values.

Definition 2. *Conditional entropy: Conditional entropy can evaluate the uncertainty of the value of a feature, suppose there are X pieces of data in the data set, and each piece of data has s features, which are expressed as A=$\{f_1, f_2, \ldots, f_s\}$. When the overall distribution of feature set A is fixed, the conditional entropy $H(C/A)$ is defined as follows:*

$$H(C) = -\sum_{f \in A}\sum_{c \in C} p(f,c)logp(c/f), \tag{2}$$

where $H(C/A)$ represents the uncertainty of the category C under the condition that the feature set A is different in value, and $P(c/f)$ represents the conditional probability that the category c takes the value under the condition of the feature $A = f$.

Definition 3. *Information gain: The information gain reflects the importance of the feature. The greater the information gain, the more important the features are. Then the information gain IG brought by the feature set A to the system is defined as follows:*

$$IG(A) = H(C) - H(C/A). \tag{3}$$

The information gain tends to select attributes with more branches, which may lead to over-fitting. In order to change the shortcomings of information gain, the information gain ratio is used to judge the partitioning attribute.

Definition 4. *Information gain ratio:*

$$G\_R(A) = IG(A)/H(A), \tag{4}$$

where $G\_R(A)$ is the information gain ratio of feature set A. $H(A)$ is the information entropy when feature A is a random variable according to the Equation (1).

The pseudo code of the information gain ratio feature selection algorithm is defined as Algorithm 1. where $num(S)$ represents the number of features in the selected feature set S, and $max(G\_R(f_i))$ represents the maximum information gain ratio in the feature set $A = f_1, f_2, \ldots, f_s$, The first k selected features are added to the set S, and finally the feature set S is obtained. The algorithm description is detailed in the appendix.

Algorithm 1 Information gain ratio feature selection algorithm

```
Input: Training data_set and feature selection quantity $k$
Output: Selected feature set $S$
Initialize feature sets $S = \varnothing$ /* Initialize feature set $S$ to an empty set */
Initialize all feature sets $A = f_1, f_2, \ldots, f_s$ /* $s$ is the number of attribute features */
Calculate the information gain ratio of each feature in feature set $A$ from Equation (4)
while num($S$) $< k$ do
    Select $max(G\_R(f_i))$, add the attribute $f_i$ to the feature set $S$.
end while
The selected feature set $S$ is obtained, and the number of selected features is $k$.
```

2.3 Ensemble Theory

2.3.1 Bagging Algorithm

The ensemble classifier is a kind of supervised learning method. As a kind of ensemble classifier, Bagging can avoid the over-fitting of the classifier and can improve the detection efficiency of unknown attacks. The ensemble learning classifier includes m base classifiers, which are trained by Bootstrap sampling method. After m times sampling, the results of m base classifiers are obtained. Finally, the classification results of the ensemble classifier are integrated according to the majority voting principle. Figure 2 shows the specific flow of the Bagging algorithm.

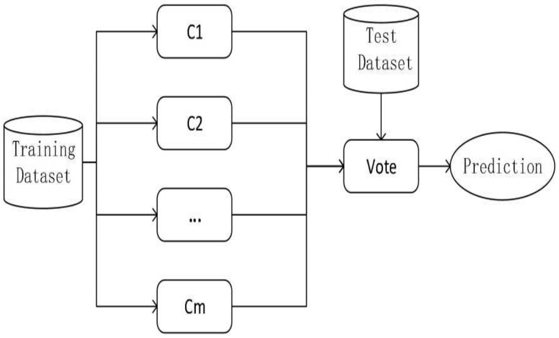

Figure 2: Bagging algorithm

2.3.2 Improved C4.5 Algorithm

The C4.5 algorithm is an algorithm to solve the problem of machine learning classification. The algorithm can find a mapping relationship between feature values and categories, and this mapping relationship can be used to classify unknown intrusion types. The C4.5 algorithm is a tree structure similar to a flow chart. A non-leaf node represents a test on an attribute. Each branch represents a test output, and each leaf node stores a class label. The advantage of this algorithm is that it does not require any domain knowledge, it is suitable for detective knowledge discovery, and it's highly efficient for detecting unknown attack types. For a leaf node, it covers q samples, there are e errors and the penalty factor is 0.5. Assuming that a decision tree has r leaf nodes, the prediction error of the decision tree is ER, which the formula is as follows Equation (5):

$$ER = (\sum_{i=1}^{r} e_i + 0.5 \times r) / \sum_{i-1}^{r} q_i \qquad (5)$$

where e_i is the number of samples misclassified in the i-th leaf node of the subtree, and q_i represents the number of samples in the i-th leaf node of the subtree.

The improved C4.5 algorithm pseudo code is defined as follows:

Algorithm 2 demonstrates the process of detecting anomalous intrusions in the WSN by the improved C4.5 classifier. First, if the node satisfies the stop split condition, all records belong to the same category, and it is set as a leaf node; Then the feature with the largest information gain rate is selected for splitting, and the first two steps are repeated until all data classification is completed. Finally, the generated tree needs to be dynamically pruned to reduce the prediction error. The algorithm description is detailed in the appendix.

3 The Proposed Model of WSN Intrusion Detection

WSN Intrusion detection model based on information gain ratio and Bagging algorithm, the shortened form is WI-IGRB, the information gain ratio is used for feature selection, and then the parameters of the ensemble classifier are optimized through 10 iterations, and the dynamic pruning process is introduced. The iteration 10 times is relatively suitable. The parameters of the Bagging algorithm are optimized during the iterative process, and the complexity of the algorithm cannot be too high that may lead to over-fitting of the model. The dynamic pruning process starts from the leaf node of the C4.5 decision tree, calculates the prediction error from the bottom to the node and the prediction error after pruning. If the prediction error after pruning is relatively small, the node is cut off. This process is repeated repeatedly until the prediction error is minimized. Finally, the majority voting system is used to count the type of the most predicted votes in the classifier and use it as the final result of the ensemble classifier. Figure 3 is a flow chart of the proposed WSN intrusion detection model.

As shown in Figure 3, the wireless sensor node collects environmental data and transmits the data to the cluster head node. The cluster head node processes the collected data and transmits it to the base station. The collected traffic data is selected from the base station as a training

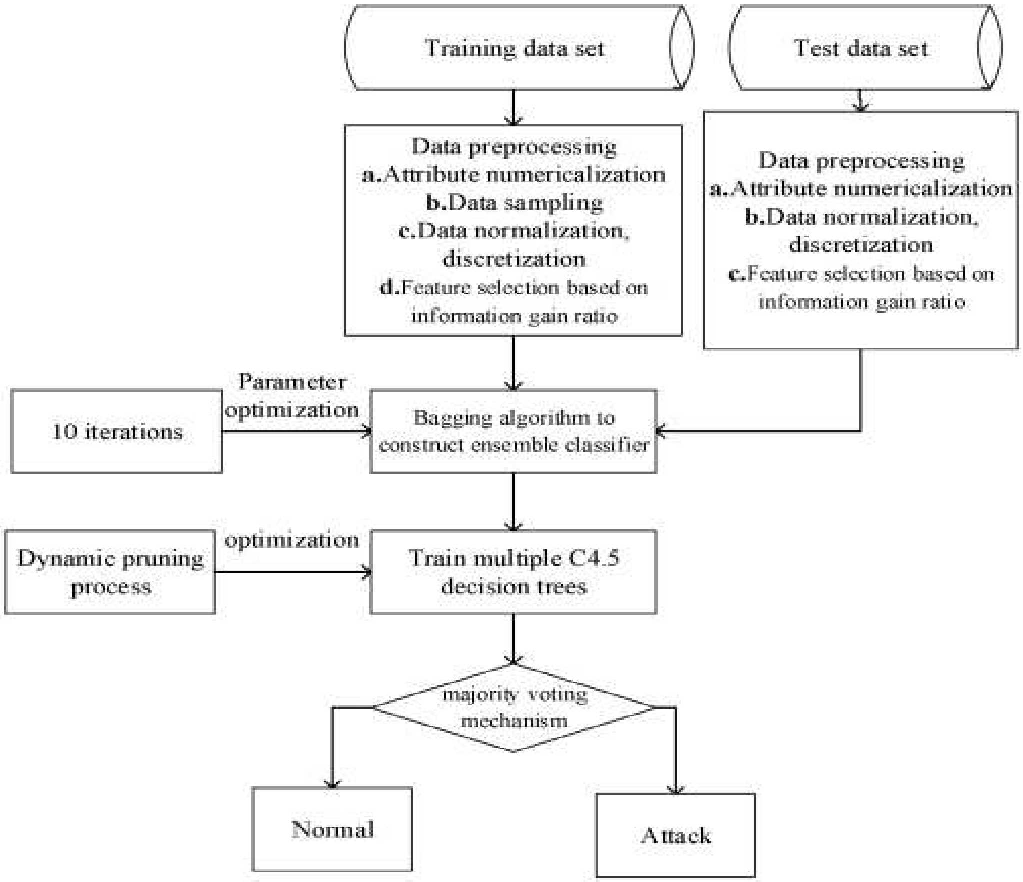
Training data set
Test data set
Data preprocessing
a.Attribute numericalization
b.Data sampling
c.Data normalization, discretization
d.Feature selection based on information gain ratio
Data preprocessing
a.Attribute numericalization
b.Data normalization, discretization
c.Feature selection based on information gain ratio
10 iterations
Parameter optimization
Bagging algorithm to construct ensemble classifier
Dynamic pruning process
optimization
Train multiple C4.5 decision trees
majority voting mechanism
Normal
Attack

Figure 3: Flow chart of WSN intrusion detection model

data set and a test data set respectively, and the proposed intrusion detection model is trained. It is mainly divided into the following two stages:

1) Model training phase: Preprocessing the training data set, including numeralization, proportional sampling of data, data normalization and discretization operations, and feature selection based on information gain ratio; using Bagging algorithm to construct ensemble classifier, multiple C4.5 decision trees are trained, and the dynamic pruning process is introduced to reduce the prediction error. Finally, the classification prediction is carried out by the majority voting mechanism.

2) Model intrusion detection phase: Preprocessing the collected test data set, including digitizing some features, data normalization and discretization processing, and feature selection based on information gain ratio; using trained integrated detection model Classification; The majority of voting mechanisms are used to integrate classifications to determine whether intrusion has occurred.

Majority voting mechanisms are defined as Equation (6), where m is the number of samples collected by the Bootstrap sampling method, l is the traffic data to be classified, L is the result of the classification, and C^* is used to count the predicted votes in the m classifiers C_i. The most type and use it as the final result of the integrated classifier.

$$C^*(l) = max(\sum_i^m \alpha(C_i(l) = L). \qquad (6)$$

The proposed WSN intrusion detection model algorithm in this paper is defined as follows:

where m is the number of samples collected by the Bootstrap sampling method, and N is the number of iterations of the algorithm. In the model training phase, the Boostrap Sampling sampling method independently trains the decision tree C_i by randomly selecting m sample numbers. Finally, the prediction function is generated in parallel to get the ensemble classifier C^*. In the model intrusion detection phase, the trained ensemble classifier C^* is used to determine whether intrusion behavior occurs in the WSN. The algorithm description is detailed in the appendix.

Algorithm 2 Improved C4.5 algorithm

```
1: Input: Data Set B
2: Output: T-decision tree after dynamic pruning
3: [x, s] = size(B) /*  $x$ is the number of data set B, $s$ is the number of attribute features in data set B  */
4: T={}
5: if B belongs to the same category or other stopping criteria then
6:     break
7: end if
8: while feature set $S = f_1, f_2, \ldots, f_s$ do
9:     Calculate the branch information entropy and conditional entropy of each feature by Equations (1)-(2)
10:    Calculate the information gain rate $G\_R(f_j)$ of the feature $f_j$ according to the Equation (4)
11: end while
12: $f_best$=Select the maximum information gain rate $max(G\_R(f_j))$
13: Use $f_best$ as the decision node and join T
14: Remove $f_best$ from B to get subset B*
15: if $x > 0$ then
16:     Return to step 3
17: end if
18: while B* do
19:     T*=C4.5(B*)
20:     Attach T* to the corresponding branch of the tree
21: end while
22: while T is not NULL do
23:     Calculate the prediction error of the decision tree T according to Equation (5) and the prediction error of the pruning off T leaf node
24:     if Prediction error of pruning T-leaf nodes < Prediction error of unpruned T-leaf nodes then
25:         Pruning the T-leaf node
26:     end if
27:     Pruning upward
28: end while
29: Return dynamic pruned T decision tree.
```

Algorithm 3 WSN intrusion detection model algorithm

```
Input: Train dataset, test dataset
Output: Intrusion detection result
Model training phase:
Preprocessing the train dataset, the $k$ important features are selected by Algorithm 1
for $n = 1$ to $N$ do
    for $i = 1$ to $m$ do
        Sample Rifrom sample train dataset using Bootstrap sampling method
        The improved C4.5 decision tree $C_i$ in Algorithm 2 is trained by the sample $R_i$
    end for
end for
Using the Equation (6) to get the ensemble classifier $C^*$
Model intrusion detection phase:
Preprocessing the test dataset, the $k$ important features are selected by Algorithm 1
while test dataset do
    Using the ensemble classifier $C^*$ to determine whether an intrusion has occurred.
    Output intrusion detection results
end while
```

4 Experimental Results and Analysis

4.1 Experimental Data Set Selection

The experiment uses the WSN dataset WSN-DS [4], and the simulator NS-2 was used to simulate the wireless sensor network environment. Based on the LEACH routing protocol, each data has 23 features and simulates four attack types: Blackhole, Grayhole, Flooding, and Scheduling. A total of 374,661 traffic data were collected in the WSN-DS dataset, and 10% of the data were randomly selected as the experimental data set. 60% of the data were used as the training data set, and 40% of the data were used as the test data set. The experimental environment was performed on a 64-bit Windows 7 operating system with 8 GB of RAM and an Intel core i5-3230 CPU. Table 1 illustrates WSN simulation parameters. The data distribution is shown in Table 2.

Table 1: WSN simulation parameters

Parameter	Value
Number of cluters	100
Number of clusters	5
Network area	100m×100m
Base station location	(50,175)
Size of packet header	25 bytes
Size of data packet	500 bytes
Routing protocol	Leach
Simulation time	3600s

Table 2: Distribution of WSN-DS data sets

Data Set	Training set 60%	Testing set 40%
Blackhole	603	402
Grayhole	876	583
Flooding	199	132
Scheduling	398	266
Normal	20404	13603
Sum	22480	14986

The experiment also uses the NSL-KDD dataset, an improved version of the KDD'99 dataset, which removes a large amount of redundant data and maintains the original attack type ratio more suitable for evaluating the actual performance of the intrusion detection algorithm. Each traffic record contains 41-dimensional feature data of various continuous, discrete, and symbol types. The NSL-

KDD includes four attack categories (DoS, Probe, R2L, and U2R) [5]. The NSL-KDD includes a training dataset KDDTrain+_20Percent and a test dataset KDDTest-21. The training data set consists of 21 types of attacks, and 17 new attack types are added to the test set. First, the NSL-KDD data set needs to be preprocessed, and the feature protocol_type, service and attack class is digitized. Then, the data set is divided into five classes, normal, DoS, Probe, U2R, and R2L, mapped to values 1-5 respectively. Finally, normalize the values of the src_bytes and dst_bytes field columns to map the range to [0,1]. The specific data distribution of the NSL-KDD data set is shown in Table 3.

Table 3: Distribution of NSL-KDD data sets

Data Set	KDDTrain+_20Percent	KDDTest+
Normal	13449	9711
DoS	9234	7458
Probe	2289	2421
U2R	11	200
R2L	209	2754
Sum	25192	22544

4.2 Experimental Performance Index

In order to measure the performance of the wireless sensor network intrusion detection model, the true positive rate (TPR), false positive rate (FPR), accuracy (Acc), precision (P) indicators are used for measurement. TP indicates that the true value is a normal sample and is predicted as the number of normal samples. FN indicates that the true value is a normal sample and is predicted as the number of abnormal samples. FP indicates that the true value is an abnormal sample and is predicted as the number of normal samples. TN indicates that the true value is an abnormal sample and is predicted as the number of normal samples. Table 4 shows the definitions of TP, FP, TN and FN.

Table 4: Definition of TP, FP, TN and FN

True value	Predicted	
	Normal	Abnormal
Normal	TP	FN
Abnormal	FP	TN

$$\begin{aligned} TPR &= TP/(TP+FN) \\ FPR &= FP/(FP+TN) \\ P &= TP/(TP+FP) \\ Acc &= (TP+TN)/(TP+FN+FP+TN) \end{aligned}$$

where TPR indicates the probability that the true value is normal, the probability of the prediction is positive. FPR indicates the probability that the true value is abnormal, and the prediction is positive; P indicates the probability that the prediction is normal and the correct prediction is normal. Acc represents the accuracy of the prediction result, and the number of normal samples is predicted divided by the total number of samples.

4.3 Feature Selection Method

The existing intrusion detection method adopts data preprocessing methods such as feature selection and data dimensionality reduction to reduce the computational load of the intrusion detection method and enhance the detection efficiency. The main feature selection and dimension reduction methods are: correlation feature selection, linear discriminant analysis, mutual information, information gain, gain ratio, principal component analysis and other methods [11]. In this study, attribute reduction is used for WSN-DS data. First, features that have no or little impact on data types were eliminated, and then common feature selection methods and information gain ratio were selected for comparative analysis. The selected features and performance are shown in Table 5.

It can be seen from Table 3 that the experiment uses the Algorithm 1 information gain ratio to select features, set k=14, select 14-dimensional features from the 23-dimensional features, and select the flow feature set S={3,5,6,7,8,9,10,11,12,13,15,16,17,18}, and use Acc as an evaluation index. Using the information gain feature selection method, if the number of features is much larger than the number of categories, the information gain will become large, and the generalization ability will be reduced without using other more effective classification information. The information gain ratio introduces split information, and the feature splitting information with a large number of values becomes large, which can effectively control the problem of excessive information gain. Through the experiment, the principal component analysis method Acc reached 94.91%, and when using the information gain method, the Acc was 98.52%, and when the information gain ratio feature selection method was used, the Acc was 98.75%. The proposed WSN intrusion detection model has a better classification accuracy when choosing the information gain ratio as the feature selection method. Table 6 lists the selected traffic characteristics and specific description information in the WSN-DS. It includes the number of features in the WSN-DS, the name of features, and the description of corresponding features in the data set.

The information gain ratio method is used to select the important features of the WSN-DS data set traffic characteristics. Comparing the selected features and specific information listed in Table 6, the final selected traffic feature set is S={3,5,6,7,8,9,10,11,12,13,15,16,17,18}.

Table 7 lists the selected traffic characteristics in the NSL-KDD. It includes the number of features in the NSL-KDD, the name of features in the dataset.

The information gain ratio algorithm is used to select the important features in NSL-KDD. The last selected

Table 5: Comparison of feature selection methods

Feature selection method	Feature selection result	Acc (%)
Principal component analysis	1,2,3,4,5,6,7,8,9,10,11,12,13,14	94.91
Information gain	1,3,4,5,6,7,8,10,11,12,13,15,17,18	98.52
Information gain ratio	3,5,6,7,8,9,10,11,12,13,15,16,17,18	98.75

Table 6: WSN-DS data set selected traffic features

Feature number	Feature name	Description
1	Node ID	Node ID number
2	Time	Node runtime
3	IS CH	Used to mark whether the node is a cluster head
4	Who CH	Cluster head ID
5	Distance to CH	Distance between node and cluster head
6	Energy consumption	Energy consumed
7	ADV CH send	The number of the advertise CH's broadcast messages sent to the nodes
8	ADV CH receives	The number of advertise CH messages received from CHs
9	Join REQ send	The number of join request messages sent by the nodes to the CH
10	Join REQ receive	The number of join request messages received by the CH from the nodes
11	ADV SCH send	The number of join advertise TDMA schedule broadcast message sent to the nodes
12	ADV SCH receives	The number of scheduled messages received by the CH
13	Rank	Order of node TDMA scheduling
14	Data sent	The number of packets sent from the normal node to its CH
15	Data received	The number of packets received by the node from the CH
16	Data sent to BS	The number of packets sent to the BS
17	Distance CH to BS	Distance between CH and BS
18	Send Code	The cluster sending code
19	Attack Type	Type of the node

Table 7: NSL-KDD dataset features

No.	Feature	No.	Features
1	Duration	22	Is_guest_login
2	Protocol_type	23	Count
3	Service	24	Srv_count
4	Flag	25	Serror_rate
5	Src_bytes	26	Srv_serror_rate
6	Dst_bytes	27	Rerror_rate
7	Land	28	Srv_rerror_rate
8	Wrong_fragment	29	Same_srv_rate
9	Urgent	30	Diff_ser_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_rate
14	Root_shell	35	Dst_host_diff
15	Su_attempted	36	Dst_host_same_srv_port_ra
16	Num_root	37	Dst_host_serror_rate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_files	40	Dst_host_serror_rate
20	Num_outbound_cmds_files	41	Dst_host_srv_serror_rate
21	Is_host_login		

feature set is {9, 26, 25, 4, 12, 39, 30, 38, 6, 29, 5, 3, 37, 11, 22, 35, 34, 14}.

4.4 Performance Analysis

Table 8 shows the detection performance of the proposed WSN intrusion detection model for Normal type and attack types based on WSN-DS, such as Blackhole, Grayhole, Flooding, and Scheduling.

As can be seen from Table 8, the detection accuracy of the proposed WSN intrusion detection model for attacks in the WSN, such as Blackhole, Grayhole, Flooding, and Scheduling, is 99.04%, 97.96%, 99.02%, and 96.21%, respectively. The detection accuracy of the normal state is 98.85%. The weighted average results show that the model true positive rate is 99.4%, the false positive rate is 1.9%, the precision is 99.4%, and the classification accuracy rate is 98.75%. The experimental results show that the proposed WSN intrusion detection model has better performance in attack detection in WSN environment and can identify different attack types.

Table 9 shows the WSN intrusion detection model and PCA-SVM [20], Naive Bayes [10], Bayesian Nerwork [23], IG-C4.5 [22], Boosting-C5.0 [14], ANN [4] methods. The specific results of performance comparison were measured and compared using TPR, FPR, Acc, and P index.

As can be seen from Table 9, the TPR of the proposed method reaches 99.4%, which is higher than that of PCA-SVM, Naive Bayes, Bayesian Network, IG-C4.5, and ANN. Among them, the TPR of Naive Bayes method is 95.2%, which is the smallest compared with the above methods. However, the false positive rate FPR of this method is 1.9%, which is higher than Naive Bayes, Bayesian Network and ANN methods. When the detection rate of the WSN intrusion detection model is increased, the data that causes the true value of the attack behavior is incorrectly predicted as the probability of a normal sample increases, and then he false positive rate increases.

The false positive rate of IG-C4.5 method reaches 3.8%, which is the highest false positive rate compared with other methods. The Acc and P are respectively 98.8% and 99.4%, which are higher than PCA-SVM, Naive Bayes, Bayesian Network, IG-C4.5, and ANN methods. Among them, Acc is 0.25% higher than Boosting-C5.0. The reason is that the selection of Boosting training sets is related to the learning results of the previous rounds, which may lead to over-fitting and reduce classification accuracy. In summary, the proposed method performs better than other intrusion detection methods.

The NSL-KDD dataset is used to evaluate the performance of the proposed method and a 10-fold cross validation was performed. In a 10-fold cross validation, the data was divided into 10 replicates of equal size. In each iteration, each part of the data is used for verification and the remaining nine parts are used to train the model. Table 10 shows the performance of the proposed WSN intrusion detection model using the NSL-KDD data set and 10-fold cross validation. Detection performance for different attack types DoS, Probe, U2R, and R2L and Normal appearing in NSL-KDD.

It can be seen from Table 10 that the detection accuracy of the proposed intrusion detection model for NSL-KDD, such as DoS, probe, U2R and R2L, is 99.85%, 99.35%, 59.05%, and 94.0%, respectively. The detection accuracy of the Normal reaches 99.65%. The overall TPR of the model is 99.69%, the FPR is 0.31%, the P is 99.6%, and the Acc is 99.69%.

Table 11 shows the WSN intrusion detection model and PCA-SVM, Naive Bayes, Bayesian Nerwork, IG-C4.5, Boosting-C5.0, ANN methods. The specific results of performance comparison were measured and compared using TPR, FPR, Acc, and P index

As can be seen from Table 11, the performance of the proposed method was evaluated using the NSL-KDD data set and 10-fold cross validation was performed. The TPR of the method in this paper reaches 99.69%, which is higher than that of PCA-SVM, Naive Bayes, Bayesian Nerwork, IG-C4.5, and ANN. The FPR of this method is 0.31%, which is lower than other methods. The reason is that the Bagging ensemble algorithm effectively reduces the variance of the model.

The FPR of Naive Bayes method reaches 11.42%, which is the highest FPR compared with other methods. The Acc of this method is 99.69%, and the Acc is lower than that of Boosting-C5.0, compared with PCA-SVM and Naive Bayes. Bayesian Nerwork, IG-C4.5, and ANN methods are high. Comparing the model building time, we can see that the proposed method has a lower time. The reason is that the selection of Bagging algorithm training set is random, and each round of training sets is independent of each other, while Boosting the selection of each round of training sets is related to the learning results of the previous rounds.

The various predictive functions of Bagging can be generated in parallel, and the various predictive functions of Boosting can only be generated sequentially, such as neural networks, which are extremely time-consuming learning methods. Bagging can save a lot of time overhead through parallel training. At the same time, in the model establishment stage, the information gain ratio method is used to reduce the dimension of the traffic data, and the features with low importance are removed. The 18-dimensional features are selected from the 41-dimensional features. The calculation and time overhead in the detection process are effectively reduced, which is more suitable for the WSN intrusion detection environment.

The experiment also used the training dataset KDDTrain+_20Percent and the test dataset KDDTest-21 to evaluate the performance of the model. Table 12 shows the performance of the WSN intrusion detection model and PCA-SVM, Naive Bayes, Bayesian Nerwork, IG-C4.5, Boosting-C5.0, ANN methods.

Table 8: Performance of the WSN intrusion detection model based on WSN-DS

Performance	**Blackhole**	**Grayhole**	**Flooding**	**Scheduling**	**Normal**	**Weighted average results**
TPR (%)	98.2	96.1	98.2	92.4	99.8	99.4
FPR (%)	0.1	0.2	0.1	0.0	2.1	1.9
P (%)	96.5	96.3	90.2	97.6	99.8	99.4
Acc (%)	99.04	97.96	99.02	96.21	98.85	98.75

Table 9: Comparison of performance of different methods of WSN intrusion detection model

Methods	**TPR (%)**	**FPR (%)**	**P (%)**	**Acc (%)**
PCA-SVM	96.6	8.6	96.7	94.0
Naive Bayes	95.2	1.0	96.5	97.1
Bayesian Network	96.5	0.9	97.7	97.8
IG-C4.5	97.8	3.8	98.3	97.0
Boosting-C5.0	99.4	2.4	99.4	98.5
ANN	98.5	1.7	98.7	98.4
The proposed method	99.4	1.9	99.4	98.75

Table 10: Performance of intrusion detection methods using NSL-KDD

Performance	**Probe**	**DoS**	**U2R**	**R2L**	**Normal**	**Sum**
TPR (%)	98.8	99.9	18.2	88.0	99.8	99.69
FPR (%)	0.1	0.2	0.1	0.0	0.5	0.31
P(%)	99.4	99.7	98.9	96.8	99.6	99.6
Acc	99.35	99.85	59.05	94.00	99.65	99.69

Table 11: Comparison of performance of different methods using NSL-KDD 10-fold cross validation

Methods	**TPR (%)**	**FPR (%)**	**P (%)**	**Acc (%)**	**Model buildng time(s)**
PCA-SVM	93.02	6.97	94.26	93.02	-
Naive Bayes	88.58	11.42	88.67	88.58	-
Bayesian Network	96.69	3.72	96.68	96.48	-
IG-C4.5	96.6	5.25	96.53	95.7	-
Boosting-C5.0	-	0.38	-	99.96	6.38
ANN	99.24	0.83	99.18	99.2	198.67
The proposed method	99.69	0.31	99.60	99.69	5.78

Table 12: Comparison of performance of different methods using NSL-KDD

Methods	**TPR (%)**	**FPR (%)**	**P (%)**	**Acc (%)**	**Model buildng time(s)**
PCA-SVM	76.5	31.1	83.4	72.7	-
Naive Bayes	78.6	27.7	82.7	75.45	-
Bayesian Network	76.5	31.1	83.4	72.7	-
IG-C4.5	77.6	27.0	85.0	75.3	-
Boosting-C5.0	98.9	45.19	-	80.56	7.31
ANN	80.1	26.3	85.1	76.9	201.34
The proposed method	81.2	26.2	85.3	77.5	6.01

4.5 Algorithm Analysis

In order to further verify the performance of the proposed method, two performance indicators, time complexity and space complexity are analyzed in detail. The comparison results are shown in Table 13.

As can be seen from Table 13, the number of data sets is X, and the number of flow characteristics is k, m is the number of samples collected by Bootstrap sampling method, and N is the number of algorithm iterations. The time complexity of the PCA-SVM method is $O(5kX)$ and the space complexity is $O(X^2)$. The time complexity of the Naive Bayes method is $O((k+k^2)X)$ and the space complexity is $O(2kX)$. The time complexity of the Bayesian Network method is $O((k+k^2)X)$ and the space complexity is $O(2kX)$. The time complexity of IG-C4.5 method is $O(X+log_2N)$, and the space complexity is $O(kX)$. The time complexity of the Boosting-C5.0 method is $O(X+Nmlog_2X))$, and the space complexity

Table 13: Comparison of time complexity and space complexity performance

Methods	Time complexity	Space complexity
PCA-SVM	$O(5kX)$	$O(X^2)$
Naive Bayes	$O((k+k^2)X)$	$O(2kX)$
Bayesian Network	$O((k+k^2)X)$	$O(2kX)$
IG-C4.5	$O(X+log_2N)$	$O(kX)$
Boosting-C5.0	$O(X+Nmlog_2X)$	$O(kX)$
ANN	$O(kNX)$	$O(kX)$
The proposed method	$O(X+Nmlog_2X)$	$O(mkX)$

is $O(kX)$. The time complexity of the ANN method is $O(kNX)$ and the space complexity is $O(kX)$.The time complexity of the proposed method is $O(X+Nmlog_2X)$, which is higher than that of the IG-C4.5 method. When the number of data sets is very large, $m < k << X$, then $O(X+Nmlog_2X)) < O(kNX)$, It can be seen that the algorithm time complexity of the proposed method is more time complex than that of the ANN method, and the space complexity is relatively constant.

5 Conclusions

A WSN intrusion detection model based on information gain ratio and Bagging algorithm is proposed. In the data preprocessing stage, the feature selection method based on information gain ratio is used to reduce the dimension of the collected WSN traffic data, which reduces the computational complexity of the intrusion detection method and effectively reduces the computation and time overhead in the detection process. In the integrated learning phase, the Bagging algorithm is used to construct the integrated classifier, and several improved C4.5 decision trees are trained. The dynamic pruning process is introduced to reduce the prediction error, and the parameters of the integrated classifier are optimized by 10 iterations. In the intrusion detection phase, the trained integrated classifier is used to classify the data, and the majority voting mechanism is used to judge whether the intrusion behavior occurs. The experimental results show that the detection accuracy of Blackhole, Grayhole, Flooding, Scheduling and Normal are 99.04%, 97.96%, 99.02%, 96.21% and 98.85%, respectively. Compared with other intrusion detection methods, the detection accuracy is improved. A variety of attack types can be effectively detected. Subsequent research focuses on extending other types of attacks in the WSN dataset, using other feature selection techniques combined with deep learning models for wireless sensor network intrusion detection.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61363078), the Research Project in Universities of Education Department of Gansu Province (2017B-16, 2018A-187). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] A. Abduvaliyev, A. S. K. Pathan, R. Roman J. Zhou, and W. C. Wong, "On the vital areas of intrusion detection systems in wireless sensor networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1223–1237, 2013.

[2] H. I. Ahmed, N. A. Elfeshawy, S. F. Elzoghdy, H. S. El-Sayed, and O. S. Faragallah, "A neural network-based learning algorithm for intrusion detection systems," *Wireless Personal Communications*, vol. 97, no. 2, pp. 3097–3112, 2017.

[3] V. N. T. AlkaChaudhary and A. Kumar, "A new intrusion detection system based on soft computing techniques using neuro-fuzzy classifier for packet dropping attack in manets," *International Journal of Network Security*, vol. 18, no. 3, pp. 514–522, 2016.

[4] I. Almomani, B. Al-Kasasbeh, and M. Al-Akhras, "Wsn-ds: A dataset for intrusion detection systems in wireless sensor networks," *Journal of Sensors*, vol. 2016, no. 2, pp. 1–16, 2016.

[5] Amrita and K. K. Ravulakollu, "A hybrid intrusion detection system: Integrating hybrid feature selection approach with heterogeneous ensemble of intelligent classifiers," *International Journal of Network Security*, vol. 20, no. 1, pp. 41–55, 2018.

[6] R. A. R. Ashfaq, X. Z. Wang, and J. Z. Huang, "Fuzziness based semi-supervised learning approach for intrusion detection system," *Information Sciences*, vol. 378, pp. 484–497, 2017.

[7] B. M. Aslahi-Shahri, R. Rahmani, and M. Chizari, "A hybrid method consisting of ga and svm for intrusion detection system," *Neural Computing and Applications*, vol. 27, no. 6, pp. 1–8, 2016.

[8] M. C. Belavagi and B. Muniyal, "Performance evaluation of supervised machine learning algorithms for intrusion detection," *Procedia Computer Science*, vol. 89, no. 2016, pp. 117–123, 2016.

[9] Z. Chiba, N. Abghour, and K. Moussaid, "A novel architecture combined with optimal parameters for back propagation neural networks applied to anomaly network intrusion detection," *Computers and Security*, vol. 75, no. 2018, pp. 36–58, 2018.

[10] D. H. Deshmukh, T. Ghorpade, , and P. Padiya, "Intrusion detection system by improved preprocessing methods and na?ve bayes classifier using nsl-kdd

99 dataset," in *International Conference on International Conference on Electronics and Communication Systems (ICECS'14)*, pp. 1–7, Feb. 2014.

[11] R. H. Dong, D. F. Wu, Q. Y. Zhang, and H. X. Duan, "Mutual information-based intrusion detection model for industrial internet," *International Journal of Network Security*, vol. 20, no. 1, pp. 131–140, 2018.

[12] Z. Feng, J. Fu, and D. Du, "A new approach of anomaly detection in wireless sensor networks using support vector data description," *International Journal of Distributed Sensor Networks*, vol. 13, no. 1, pp. 1–14, 2017.

[13] N. Gao, L. Gao, Y. He, Y. Y. He, and H. Wang, "Lightweight intrusion detection model based on self-encoding network feature dimension reduction," *Chinese Journal of Electronics*, vol. 45, no. 3, pp. 730–739, 2017.

[14] A. Garofalo, C. D. Sarno, and V. Formicola, "Enhancing intrusion detection in wireless sensor networks through decision trees," *Lecture Notes in Computer Science*, vol. 7869, no. 2, pp. 1–15, 2013.

[15] A. Ghosal and S. Halder, "A survey on energy efficient intrusion detection in wireless sensor networks," *Journal of Ambient Intelligence and Smart Environments*, vol. 9, no. 2, pp. 239–261, 2017.

[16] W. Hua, Y. Wen, and D. Zhao, "Identifying localization attacks in wireless sensor networks using deep learning," *Journal of Intelligent and Fuzzy Systems*, vol. 35, no. 2, pp. 1339–1351, 2018.

[17] T. Ma, F. Wang, J. J Cheng, Y. Yu, and X. Chen, "A hybrid spectral clustering and deep neural network ensemble algorithm for intrusion detection in sensor networks," *Sensor*, vol. 16, no. 10, pp. 1701, 2016.

[18] M. M. Ozcelik, E. Irmak, and S.Ozdemir, "A hybrid trust based intrusion detection system for wireless sensor networks," in *International Symposium on Networks, Computers and Communications*, pp. 1–6, Oct. 2017.

[19] N. Rachburee and W. Punlumjeak, "A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining," in *7th International, Conference on International Conference on Information Technology and Electrical Engineering*, pp. 420–424, Oct. 2016.

[20] M. C. Raja and M. M. A. Rabbani, "Combined analysis of support vector machine and principle component analysis for IDS," in *International Conference on Communication and Electronics Systems (ICCES'16)*, pp. 1–5, Oct. 2016.

[21] X. Sun, B. Yan, X. Zhang, and C. Rong, "An integrated intrusion detection model of cluster-based wireless sensor network," *PloS one*, vol. 10, no. 10, pp. e0139513, 2015.

[22] W. Wang, Y. He, J. Liu, and S. Gombault, "Constructing important features from massive network traffic for lightweight intrusion detection," *IET Information Security*, vol. 9, no. 6, pp. 374–379, 2015.

[23] Y. Wang, Y. Shen, and G. Zhang, "Research on intrusion detection model using ensemble learning methods," in *IEEE International Conference on Software Engineering and Service Science*, pp. 422–425, Nov. 2017.

[24] X. Wu, X. Peng, and Y. Yang, "Two-level feature selection method based on svm in intrusion detection," *Transactions of Communications*, vol. 34, no. 4, pp. 19–26, 2015.

[25] C. Yin, L. Ma, and L. Feng, "Towards accurate intrusion detection based on improved clonal selection algorithm," *Multimedia Tools and Applications*, vol. 76, no. 19, pp. 1–14, 2017.

[26] B. Zarpelao, Miani R S, and C. T. Kawakani, "A survey of intrusion detection in internet of things," *Journal of Network and Computer Applications*, vol. 84, no. 2017, pp. 25–37, 2017.

Appendix

A 1

Algorithm 1 Information gain ratio feature selection algorithm.

Proof. Information gain ratio is a filter method of feature selection. The Information gain ratio is defined in detail in Section 2.2 of this paper. The algorithm logic steps are described in detail below.

Step1: Initialize the feature set $S = \emptyset$, which will be used to save the selected feature.

Step2: From the original feature set $A = f_1, f_2, \ldots, f_s$, the information gain ratio of each feature is calculated by the formula (4), expressed as $G\_R(f_i)$, $i \in [1, s]$.

Step3: The first k features are sequentially added to the feature set S, and the selected feature set S is output.

From the logic point of view, Information gain ratio feature selection algorithm is correct. After experimental verification, the algorithm finally selects k important features. □

B 2

Algorithm 2 Improved C4.5 algorithm.

Proof. Algorithm 2 demonstrates the process of detecting anomalous intrusions in the WSN by the improved C4.5 classifier. The algorithm logic steps are described in detail below.

Step1: If the node satisfies the stop split condition, all records belong to the same category or the maximum information gain rate is less than the threshold, indicating that the B data set does not need to be classified and break out of the program.

Step2: According to the information entropy formula (1), find the information entropy $H(S)$ of each feature in $S = \{f_1, f_2, \ldots, f_s\}$, calculate the conditional entropy of each feature in feature set S according to the conditional entropy formula (2) and obtain the information gain ratio $G\_R(f_j)$ of each feature in feature set S according to formula (4). The feature fbes with the largest information gain rate($max(G\_R(f_j))$) is selected as the decision node and added to the decision tree T. Repeat the first two steps until all data classifications are complete.

Step3: The generated tree needs to be dynamically pruned to reduce the prediction error. Firstly, delete the subtree rooted at this node, Then, make it a leaf node, the most common classification of training data assigned to the node. Finaly, when the pruned tree is not worse than the original tree for verifying the performance of the set, the node is actually deleted.

From the above, the improved C4.5 algorithm logic is correct, after experimenting, this algorithm can classify each piece of traffic data to generate a decision tree with less prediction error. □

C 3

Algorithm 3 WSN intrusion detection model algorithm.

Proof. Where m is the number of samples collected by the Bootstrap sampling method, and N is the number of iterations of the algorithm. The algorithm logic steps are described in detail below.

Step1: In the model training phase, the Boostrap Sampling sampling method independently trains the decision tree C_i by randomly selecting m sample numbers.

Step2: Using the majority voting mechanism to get the ensemble classifier C^* . Majority voting mechanisms are defined as Equation (6).

Step3: In the model intrusion detection phase, preprocessing the test dataset, as in the training dataset preprocessing. the k important features are selected by Algorithm 1.

Step4: The trained ensemble classifier C^* is used to determine whether intrusion behavior occurs in the WSN.

In summary, the algorithm logic is correct, and it has been proved by experiments. There are detailed experimental results in 4 experimental results and analysis. □

Biography

Rui-Hong Dong. Researcher, worked at school of computer and communication in Lanzhou university of technology. His research interests include network and information security, information hiding and steganalysis analysis, computer network.

Hou-Hua Yan. received the BS degrees in Computer Science and Technology from Taiyuan Institute of Technology, Taiyuan, China, in 2015. Currently, he is studying for his masters degree at Lanzhou University of Technology. His research interests include network and information security, wireless sensor network security, intrusion detection.

Qiu-Yu Zhang. Researcher/Ph.D. supervisor, graduated from Gansu University of Technology in 1986, and then worked at school of computer and communication in Lanzhou University of Technology. He is vice dean of Gansu manufacturing information engineering research center, a CCF senior member, a member of IEEE and ACM. His research interests include network and information security, information hiding and steganalysis, multimedia communication technology.