Intrusion Detection Method Based on Clustering Information Entropy for Smart Home

Wentao Xiong

College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.

xwtao2015@126.com

Abstract

Smart home based on the rising internet of things (IoT) technologies is facing to lots of risks in information safety such as eavesdropping of users' privacy, interception and modification of measuring and controlling instructs. According to the characteristics and potential security problems of smart home network, an algorithm of artificial immune intrusion detection based on clustering information entropy is proposed for smart home. The cluster detectors are generated and the values of affinity are estimated through information entropy, the intrusion detection is realized by the improved method based on the integration of anomaly detection and misuse one. The results of test show that the proposed method is helpful to improve the accuracy of intrusion detection; the detection rate is higher than 98%, while the false detection rate is less than 2%. Moreover, this proposed method is able to reduce the complexity of time in intrusion detection and save the load overhead of network.

Keywords

The Internet of things, smart home, intrusion detection, immune algorithm, information entropy, cluster.

1. Introduction

As one typical application of networking technology, smart home is being rapidly developed. People enjoy a variety of convenience that smart home brings, but at the same time, faced with a lot of information security risks caused by communication open when the smart home system employed in the wireless sensor networks (Wireless Sensor Networks, WSN) technology. Such as, the node of counterfeiting, unauthorized access, loss of privacy, interception of system control instruction's tampering, forgery, replay or illegal injection, denial of service attacks and so on. These attacks will cause the system to not function properly in accordance with a predetermined function, and may reveal the user's privacy.

The research for smart home safety has just started, existing smart home security solution emphasis on passive defense, not enough to respond to security threats faced by the smart home system. Intrusion detection is the active defense means to guarantee the security of smart home, in as little as possible impact on the performance of the premise of intelligent home network, for intrusion detection, analysis, processing and early warning, thereby enhancing the ability of the system to deal with external threats.

2. Summary of Researches

In recent years, specifically for smart home safety study abroad is still limited, the only research report include: reference[1] proposed a GPRS-based smart home security monitoring strategy, smart home alarm system is limited to the hardware control plane were studied; reference[2] building intelligent home security system based on embedded Internet, not related to network and information security, mainly to study the safety methods and hardware design; reference[3] to construct a time series of encrypted smart home security control systems, research is focused on the safety of human-computer interaction, and does not involve the transfer of control of network information security; reference[4] to construct a safety smart home control system based on the SM4 encryption algorithm, mainly to discuss ways to realize the smart home system communication confidential. These studies are a passive defense mechanism, which can improve the security of smart home network to a certain extent, but for complex intrusion, which means a lack of active defense security research is not enough to build up a complete smart home security solutions.

Reference[5] provide an anomaly detection method in wireless smart home sensor network based on mobile agent, in that network, middleware is set up, the detection range is limited to the rich resources of the cluster head node memory, etc., detection range is small, limitations larger. Reference[6] Build detection system based on hierarchical specification intrusion in residential lan, according to the ZigBee specification defining the normal behavior of network protocol based on IEEE802.15.4, deviation from the defined behavior judged to be malicious acts , the disadvantage of this method don't design a method targeted intrusion on network characteristics in the smart home detection.

In this paper, combined the technical characteristics of smart home network and resource conditions, artificial immune intrusion detection algorithm based on information entropy of the cluster was proposed, seeking high detection rates and low false positive rate, in order to effectively enhance active safety smart home system.

3. Smart Home Intrusion Detection Model

3.1 Constitution of Smart Home Network

Based on the technology of Internet of things intelligent smart home will be conducted various home devices through the wireless sensor network, by the sensor nodes to completed the transfer of their equipment and control commands and information interaction. Smart home system involves relatively few number of sensor nodes (typically below 100), the node deployment method based on star topology; the node position is settled relatively .Intrusion detection function of smart home system is shown in figure 1.

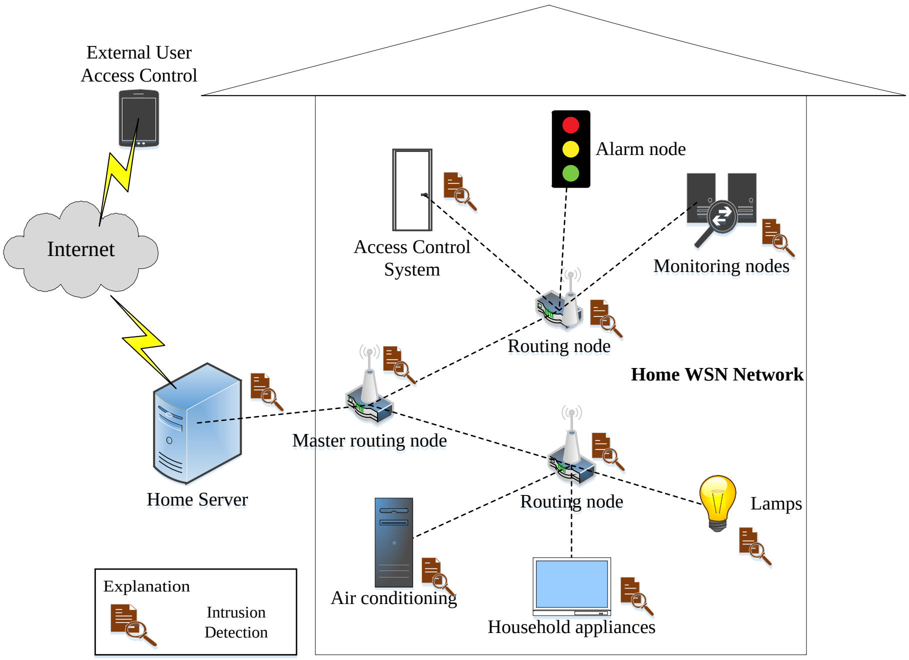

Fig.1 Smart home system with intrusion detection

In order to improve the security of smart home system based on the Internet of things, nodes and server in the network implementation of intrusion detection system. When the system is in the abnormal situation system, the implementation of intrusion detection system give alarms which is based on the test results. The main functions of the system components include:

(1) Monitoring node, it connected with various sensors, such as smoke detector, combustible gas detector, temperature and humidity sensors, light sensors, infrared sensors, cameras, etc.

(2) Control node, it connected to the household electrical appliances, and communications with routing nodes, such as CC2530.

(3) Routing nodes, it is mainly used for message routing, forwarding and information aggregation; Master routing nodes, collected information with routing nodes and interaction with home server.

(4) Family server, it can be used to filter access to network, is the information processing and storage center of smart home network, realize the human-computer interaction and intelligent control, etc.

(5) Alarm nodes, it is used to alarm the intrusion and system abnormal.

(6) The user remote intelligent terminals, such as smart phone, tablet, PC terminal equipment, can remote access.

3.2 Collaboration Detection Model

The traditional intrusion detection is divided into anomaly detection and misuse detection. Anomaly detection [7] is compared a predefined normal behavior data with the data to be detected, if match, judged to be normal; if not, judged to be invaded. Misuse detection [8] is the intrusion behavior date appeared by statistical, the data to be detected is matched with it, if match, the test data is intrusion behavior; if not; the test data is normal behavior.

Single detection method has many insufficient, such as anomaly detection has higher detection rate, but also has a high error detection rate; Misuse detection threshold is not easy to determine effectively, inappropriate threshold will increase the rate of false positives or non-response. In this paper, combining anomaly detection and misuse detection form collaborative intrusion detection model of smart home as shown in figure 2.

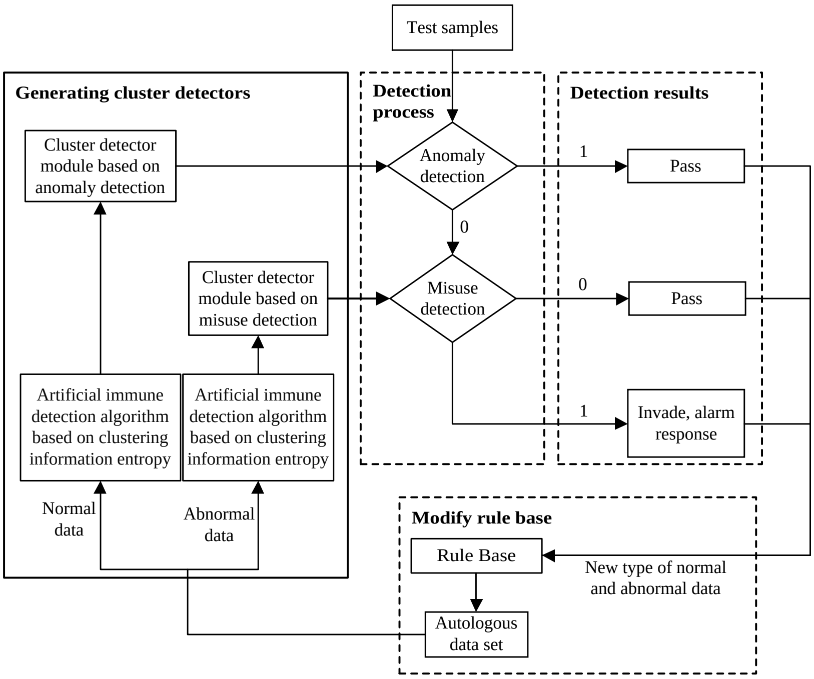
Test samples
Generating cluster detectors
Detection process
Detection results
Cluster detector module based on anomaly detection
Anomaly detection
1
Pass
0
Cluster detector module based on misuse detection
Misuse detection
0
Pass
Artificial immune detection algorithm based on clustering information entropy
Artificial immune detection algorithm based on clustering information entropy
1
Invade, alarm response
Normal data
Abnormal data
Modify rule base
Rule Base
New type of normal and abnormal data
Autologous data set

Fig.2 Cooperative intrusion detection model for smart home

The basic working process is: Under test data is being the first anomaly detection by cluster detector based on anomaly detection, if judged to be 1, it said that the data is normal; if judged to be 0, it said that the data is suspicious, and then to be misuse detection by the cluster detector based on misuse detection.

If the test output is 0, the said that question can be ruled out for the first time test result, and the date is judged to be normal data. Otherwise, it can confirm the abnormal data, give the alarm response.

Smart home intrusion detection model is the key to generate cluster detector based on anomaly detection and misuse detection, detector directly affect the intrusion detection system's detection rate and false detection rate. In order to improve the detection performance, the new tests results are used to the rule base modify and cluster detector generating algorithm for training. The generation of detector and the update method are as follows: the normal and abnormal data sets of data sets stored in the rule base will be set in advance before testing. As the initial autologous set data, through cluster information entropy of artificial immune algorithm is trained by cluster detector based on normal data generated anomaly detection and abnormal data generated misuse detection. Intrusion detection model for the new test of normal data and abnormal data feedback to rule base, implementation of the rule base updates, and it will detect the new data supplement for autologous set data, through further training update detector.

4. Artificial Immune Algorithm based on Clustering Information Entropy

The limited computing power and storage space of the sensor nodes in smart home network have restricted the application of complex intrusion detection algorithm. In this paper, based on the resource conditions of smart home system and the security threats, an artificial immune algorithm based on clustering information entropy is proposed. The algorithm is used to generate the detector by using the negative selection algorithm (NSA, Negative Selection Algorithms), to define affinity through information entropy theory, and take the idea of cluster to cluster data and detector to cluster classification to reduce the average number of detection and processing delay.

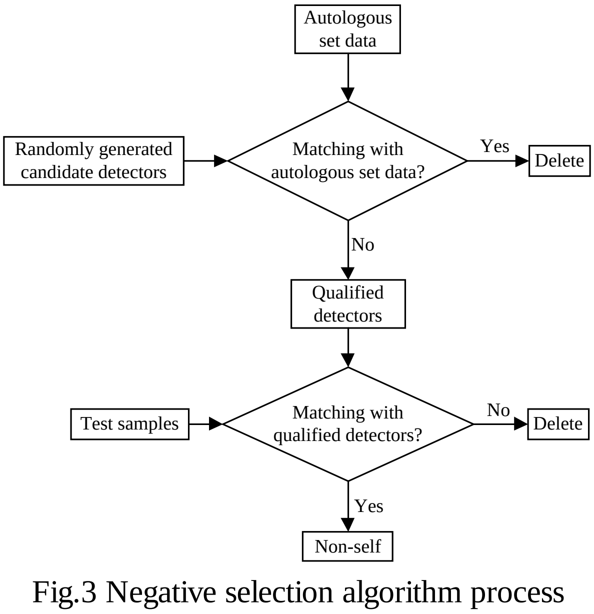
Autologous set data
Randomly generated candidate detectors
Matching with autologous set data?
Yes
Delete
No
Qualified detectors
Test samples
Matching with qualified detectors?
No
Delete
Yes
Non-self

Fig.3 Negative selection algorithm process

The algorithm mainly includes five aspects:

1) Negative Selection Algorithms

The negative selection algorithm [9] is one of the most important algorithms in artificial immune algorithm; it is the main method of generating the detector in the immune response, divide the object in the real environment (i.e., the detected data here) into the self and the non-self. The specific steps of the algorithm are as follows:

a) On a limited set of characters, defined the test sample and self-set data as b long string consisting of a set of data to be detected;

b) Randomly generating a set of candidate detectors D, delete the detectors which detected all the self, retain the detectors which detected all the non-self, to constitute the qualified detectors set D', each of detectors in D' are character string not matched with S;

c) A comparison between the detectors in D' and test data set S to monitor the changes. If the data character string is matched with the detector, the data in S is non-self, then judged as abnormal; if not match it is self and then judged as normal.

Step b) and c) to describe the process of negative selection, as shown in Figure 3.

Definition of self-set

Definition of self-set is generally using the observation method, in order to prevent the noise data input, and to observe the network data in a short time. And then define the pattern string of the auto set. The definition of an auto set is:

$$S = (s_1 \cdots s_i \cdots s_a) = \begin{vmatrix} s_{11} & \cdots & s_{1b} \\ \vdots & \ddots & \vdots \\ s_{a1} & \cdots & s_{ab} \end{vmatrix} \tag{1}$$

Where s_i is the first i autologous unit, s_{ij} ($i = 1,...,a$; $j = 1,...,b$) is the first j vector of the first i autologous unit. The definition of the sample data set is the same as this definition.

3) Generation of detector

In order to obtain the diversity of the detector, the generation of detectors is usually generated by random generation method. Each detector has its own life cycle; failure to develop a qualified detector in the life cycle will be replaced by a new detector which is randomly generated.

Randomly generated candidate detector set is:

$$D = (d_1 ... d_n) = \begin{vmatrix} d_{11} & ... & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{n1} & ... & d_{nm} \end{vmatrix} \tag{2}$$

The detector set is composed of a number of detectors, $d_i\,(i = 1,\cdots,n)$ is the first i detector, d_{ij} $(i = 1,...,n;\, j = 1,...,m)$ is the first j vector of the first i detector. A collection of all of the qualified detectors developed in the life cycle is composed of:

$$D' = (d'_1 \cdots d'_a) = \begin{vmatrix} d'_{11} & \cdots & d'_{1b} \\ \vdots & \ddots & \vdots \\ d'_{a1} & \cdots & d'_{ab} \end{vmatrix} \tag{3}$$

Here $1 \le a \le n; 1 \le b \le m$, $d'_i\,(i = 1,\cdots,a)$ is the first i qualified detector, d'_{ij} $(i = 1,...,a;\, j = 1,...,b)$ is the first j vector of the first i qualified detector. The generation of the qualified detector is achieved by matching rules, which are the same as the matching rules used in the detection process.

4) Matching rule

Matching rule, called as calculate affinity, is used to describe the similarity between antibody and antigen. It is the basic principle of detector data. We introduce the theory of information entropy [13] to calculate the affinity value between the elements of the self-set and the detector. The same as during the test phase, we use this theory to test the date matching. Information entropy is used to measure the amount of information, that is, according to the probability of the appearance of things to get the corresponding information entropy quantificational.

The group G consists of N individuals and every individual contains M characters. In other words, group G is a set which has N length of number M of character string:

$$G = (S_1 S_2 \cdots S_i \cdots S_N) = \begin{vmatrix} S_{11} \cdots S_{1j} \cdots S_{1M} \\ S_{21} \cdots S_{2j} \cdots S_{2M} \\ \vdots \;\ddots\; \vdots \;\ddots\; \vdots \\ S_{N1} \cdots S_{Nj} \cdots S_{NM} \end{vmatrix} \tag{4}$$

Here $1 \le i \le N, 1 \le j \le M$. Defining any two individual's information entropy between p and q as:

$$H(p,q) = \frac{1}{M}\sum_{i=1}^{2}\sum_{j=1}^{M}\left(-p_{ij}\log_2 p_{ij}\right) \tag{5}$$

Where p_{ij} is the percentage of the value of j character of individual p or q in the group G .If individual p is different from individual q , $p_{1j} = q_{2j} = 0.5$; if not, $p_{1j} = 0$, $p_{1j} = 0$ .If every character of individual p and q are same , $H(p,q) = 1$;if not, $H(p,q) = 0$.

The affinity between individuals indicates the matching degree, and affinity is mainly used as matching computing between testing data and testing criterion in the intrusion detection. We can define affinity between individual p and q based on information entropy as:

$$\begin{aligned} A_{pq} &= \frac{1}{1+H(p,q)} \\ &= \frac{1}{1+\frac{1}{M}\sum_{i=1}^{2}\sum_{j=1}^{M}(-p_{ij}\log_2 p_{ij})} \end{aligned} \tag{6}$$

From the formula (5) and (6), when Individual p is more similar to q we can get greater value of affinity A_{pq}; conversely, we can get smaller value of A_{pq}, when individual p is less similar to q . It indicates individual p is different from q when $A_{pq} = 0.5$, what's more, individual p and q are completely same when $A_{pq} = 1$.

When we expand the degree of intimacy between two individuals to entire group G, we can get the degree of intimacy $A(G,N)$ of group G as follows:

$$\begin{aligned} A(G,N) &= \frac{1}{1+H(N)} \\ &= \frac{1}{1+\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}(-p_{ij}\log_2 p_{ij})} \end{aligned} \tag{7}$$

In the above formula, the greater $A(G,N)$, the greater degree of intimacy, but lower the diversity of group G . The same is true in reverse. When $A(G,N) = 1$, the all individuals of group G are fully identical.

Matching rule is used to calculate the affinity value between the tested data and the detector for intrusion detection. If the value is less than the setting threshold, then it does not match the detector, so we determine the tested data as the self. So we determine the normal behavior when it does not match the normal cluster, and determine abnormal behavior, when it does not match the abnormal cluster.

5) The introduction of the cluster thought

The cluster [10] thought can help to reduce the cost of calculating the affinity value between the tested data and the detectors. The cluster means providing the same service with the aid of a set of independent, high-speed network interconnected computers and using a single system management model. The cluster is equivalent to a whole with the interaction between the clients. The cluster can improve the availability and scalability of computer, to get a relatively high performance at a lower cost.

The paper according to their resemblance make normal or abnormal behavior data divide into different clusters, and the generated qualified detectors were clustered and classified. During the period, it generated multiple normal cluster detectors and abnormal cluster detectors, and each cluster contains a number of detectors. The entity as a cluster detector, the detected data are computed and gain affinity value with each cluster. If the value does not match with normal behavior cluster detector, then judged the detected data as normal behavior; if it does match with the detectors, show that the data is dubious, next, make the value compares with abnormal behavior cluster detector again, if it does not match, then judge for abnormal behavior; if it matches, then, the test results for the first time the question can be ruled out, the data can be judged as normal data. When it does finish detection of the N data, the cycle is end. The algorithm process of artificial immune intrusion detection based on clustering information entropy is followed by Figure 4.

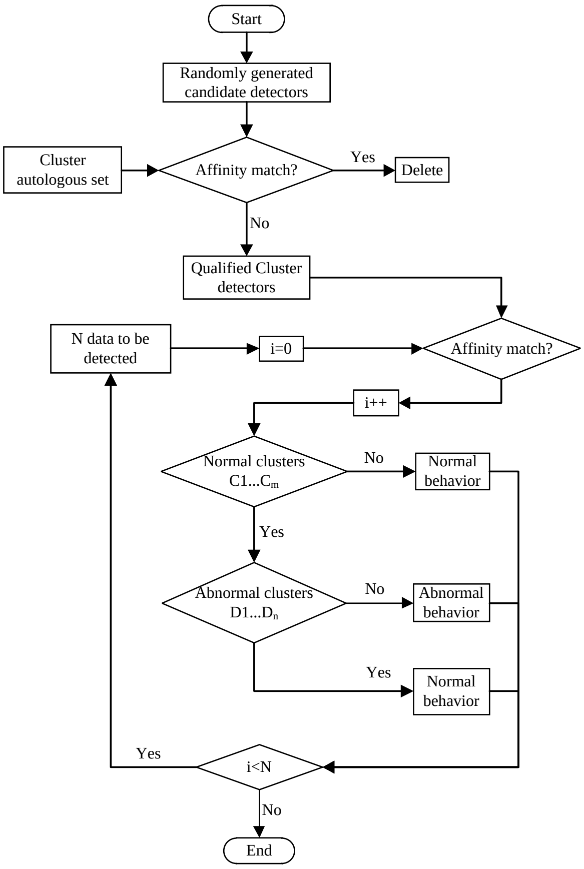

Fig.4 Process of artificial immune detection algorithm based on clustering information entropy

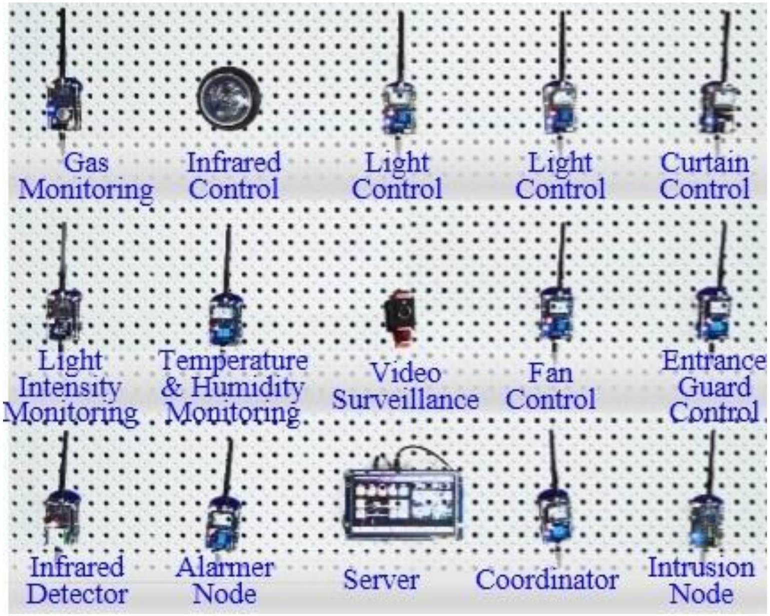

Fig.5 Test platform of intrusion detection for smart home

Processes of algorithm steps are as follows:

Step 1: Randomly generated candidate detector.

Step 2: Use Cluster thought to make autologous set and cluster classification, produces clusters autologous set.

Step 3: Set Random candidate detector and cluster autologous affinity match, delete the failed cluster detectors to obtain some qualified cluster detectors.

Step 4: Calculate the values of affinity between the cluster detectors and the tested data, If it does not match the normal cluster, it is determined that the data is normal behavior; If they match, the data is determined as suspicious, and then compares it with abnormal clusters, if not match, it is judged to be abnormal behavior, otherwise the data is normal behavior.

5. The Simulation and Performance

5.1 Simulation Test Platform and Experiment Explanation

Adopting Matlab simulation platform and VC6.0 structure smart home network intrusion detection model, to achieve the algorithm function and evaluate the performance, and as shown in figure 5 on smart home physical experimental platforms for validation.

Simulation data is from the Massachusetts Institute of Technology (MIT) Lincoln Laboratory KDDCUP99 [11] intrusion detection data set, the data set is divided into training data (with identification) and testing data (unmarked). Data set types are divided into five kinds that Normal, DOS, Probing, R2L and U2R, each containing 41 data features, identify the specific conditions shown in Table 1. KDDCUP99 original data set is large, this paper selected 10 percent of the training set (kddcup.data_10_percent.gz) were randomly selected 5000 abnormal data and 3000 normal data outliers as a training data, 10% of the test set (kddcup.newtestdata.unlabeled_10_percent.gz) were detected to evaluate the performance of this method.

In accordance with the original data set proportion of various types of attack, select 960 records from test set randomly to obtain the groups of test data. Normal, DOS, Probing, R2L, U2R selected 300, 300, 300, 20, 40 records for performance testing correspondingly. This paper will use detection rate p_d, error rate p_f to evaluate the performance of the algorithm, and defines as follows:

$$p_d = \frac{N}{M} \times 100\% \tag{8}$$

$$p_f = \frac{E}{F} \times 100\% \tag{9}$$

Where N is the abnormal number of detected successfully, M is the total number of abnormal samples; E is the number of the individuals that the normal erroneously determined to be the abnormal, F is the total number of normal samples.

Table 1 ID type of KDDCUP99 dataset

Types	ID patterns	Meanings
Normal	normal	Normal data
DOS	Smurf, teardrop, pod, back, land, neptune	Denial of service attack
Probing	Ipsweep, nmap, portsweep, satan	Surveillance or other detection behavior
R2L	ftp_write,imap,guess_passwd, phf，multihop,spy,warezclient,warezmaster	Unauthorized access from remote machines

5.2 Simulation Results and Performance Analysis

Representative results [15] as shown in Table 2, which results of this study will compare with.

Table 2 Reference [15] test results

Types	Detection rate	False detection rate
Normal	N/A	4%
DOS	88%	N/A
Probing	82%	N/A

Figure 6 is the test results that paper select normal test data set to test network in normal behavior. Compare with the results of the reference [15], except when the number of samples is 50, the false detection rate is 4%; this paper method has improved significantly for false detection rate of Normal data.

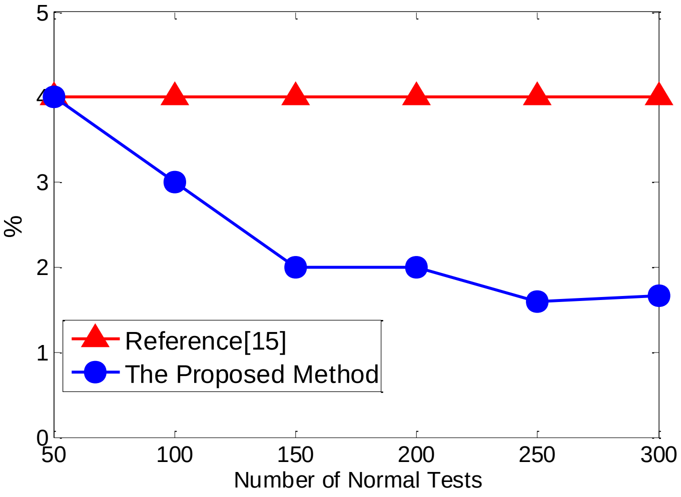

Fig.6 Error rate of normal data

Similarly, test the Dos、Probing datasets, and the results in the Table 3.

Table 3 Test results of the proposed method

Data types	Accounting	Detection rate	Error rate
Normal	295/300	N/A	1.67%
DOS	297/300	99%	N/A
Probing	295/300	98.33%	N/A

Compare table2 and table3, it shows that in this paper, the false detection rate of normal data and the detection rate of Dos and Probing are better than the reference[15].The detection rate is higher than 98% and false detection rate is less than 2%. Because of the introduced of the information entropy, the accuracy of detection is improved in this paper.

Figure 7 shows the comparison between the Artificial Immune Algorithm based on Clustering Information Entropy and traditional NSA and PSA in the average delay. The table shows the average detection time of method in this paper is better the other two methods, the larger number of tests, the more obvious advantages. Because of the introduced of the clustering, for the same number of samples to be tested, the number of required average detection is reducing; the average delay is reducing, so reduce the time complexity effectively.

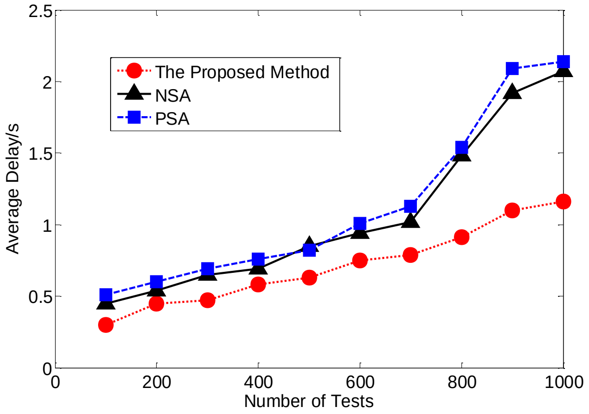

Fig.7 Delay comparison of algorithm

6. Conclusion

The smart home based on internet of things has broad prospects. It brings people home convenience and innovation experience while it is faced with disclosure of user privacy, and control instruction interception, tampering, forgery, replay, and other information security risks. We proposed artificial immune intrusion detection algorithm based on clustering information entropy for the composition and characteristics of the smart home. In this paper, in order to achieve the attacks detection of unauthorized access to system resources, command and control illegal probe or privacy theft, denial of service attacks, build a detection model combine anomaly detection and misuse detection, through affinity values described by information entropy and cluster detectors. Simulation and physical test results show the proposed method improves the accuracy of intrusion detection compared to existing solutions, its detection rate is higher than 98%, the false detection rate is less than 2% and this method reduces the time complexity.

Acknowledgements

Natural Science Foundation.

References

[1] Cheng Shuai, Zhong Xianxin, Liu Jixue, et.al. Based on the GPRS intelligent household security monitoring [J]. Computer measurement and control, 2011, 19(2):326-328.

[2] Yang X, Zhang Y, Zhao R. Study and design of home intelligent system based on embedded internet[C] //International Conference on Embedded Software and Systems Symposia. IEEE, 2008: 344-349.

[3] Deng Binwei, Li Chao: Time sequence to encrypt the design and implementation of smart home control system security[J].Journal of electronics world, 2012, 19(9): 33-35.

[4] Hu Xiangdong, Han Kaiming, Xu Hongru: Design and implementation of security-focused intelligent household Internet of things [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2014, 26(2):171-176.

[5] Usman M, Muthukkumarasamy V, Wu X W, et al. Wireless smart home sensor networks: mobile agent based anomaly detection[C]//Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic& Trusted Computing (UIC/ATC),2012 9th International Conference on. IEEE, 2012: 322-329.

[6] Jokar P, Nicanfar H, Leung V: Specification -based intrusion detection for home area networks in smart grids[C]//Smart Grid Communications, 2011 IEEE International Conference on. IEEE, 2011: 208-213.

[7] Xi X, Xia S, Tian X, et al. Anomaly detection of user behavior based on DTMC with states of variable-length sequences [J]. The Journal of China Universities of Posts and Telecommunications, 2011, 18(6): 106-115.

[8] Wang X, Cao J, Liu X, et al. Feature detection of triangular meshes via neighbor supporting[J]. Journal of Zhejiang University Science C, 2012, 13(6): 440-451.

[9] Jin Zhang-zan, Liao Ming-hong, Xiao Gang: Survey of negative selection algorithms[J]. Journal on Communications, 2013, 34(1): 159-170.

[10] Zhang Xiaofang, Hu Zhengguo, Zheng Jichuan, Tang Yan: Research and application of high availability cluster[J]. Computer Engineering, 2003, 29(4): 26-27.

[11] Information on HETTICH S, BAY S D. KDD cup 1999 data [EB/OL]. [2014-09-23]. http://kdd.ics.uci.edu/databases/kdd-cup99/kddcup99.html.

[12] Vasilomanolakis E, Karuppayah S, Mühlhäuser M, et al. Taxonomy and survey of collaborative intrusion detection[J]. ACM Computing Surveys (CSUR), 2015, 47(4): 55.

[13] Fu Z Y: Information theory-basic theory and application [M].Beijing: Electronic Industry Press, 2011:22-35.

[14] Butun I, Morgera S D, Sankar R: A survey of intrusion detection systems in wireless sensor networks[J]. Communications Surveys & Tutorials, IEEE, 2014, 16(1): 266-282.

[15] Richard L, Joshua W H: The 1999 DARPA off-line intrusion detection evaluation [J]. Computer Networks, 2000, 34(4): 579-595.

[16] Zhang L, Bai Z, Lu Y, et al. Integrated intrusion detection model based on artificial immune [J]. The Journal of China Universities of Posts and Telecommunications, 2014, 21(2): 83-90.