Intrusion Detection Approach Based on Clustering and Statistical Model for Wireless Sensor Networks

Yinghua Zhou[1], Hui Shen[1]

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

Keywords: wireless sensor networks, intrusion detection, node clustering, statistical model.

Abstract. In recent years, Intrusion detection has been the focus of security research for Wireless Sensor Networks (WSN). Some approaches or mechanisms have been designed for WSN. But none of them has been widely applied. In this paper, a scheme based on sensor node clustering and statistical model for WSN is presented. First, the sensor nodes are divided into several clusters by using k-means algorithm, and then a kind of anomaly detection algorithm based on statistical model is applied to different clusters for anomaly detection. It is shown through experiments that the scheme can decrease the false alarm rate and increase the detection rate in comparison with existing intrusion detection approaches.

Introduction

Nowadays Wireless Sensor Networks (WSN) has become a focus of international research. WSN consists of a large number of tiny sensor nodes [1]. Each sensor nodes can perform sensing, data processing and communicating. WSN has been used in a variety of domains, such as military, environmental monitoring and traffic management. With the in-depth research, the problem of network security has become a key issue for WSN [2].

In general, sensor nodes are deployed in unattended environment and communicate with each other using wireless signals. The limited capacity of sensor node such as limitation processing capacity, memory and battery lifetime further increases the insecurity of WSN. Many kinds of attacks against WSN have been identified. For example, bogus routing, sensed data attack, selective forwarding attack, sinkhole attack, wormhole attack, black hole attack and hello flood attack.

At present, the research of security for WSN focuses on key management, security routing protocol, intrusion detection, authentication technology, etc. Intrusion detection technology, as a proactive defence method, is relatively mature in traditional network. But the intrusion detection system (IDS) for traditional networks cannot be used directly in WSN, since WSN has the following features: limited energy supply, limited communication capacity, limited computing capacity of sensor nodes, large number of sensor nodes and wide distribution. Therefore, the purpose of this paper is to propose an intrusion detection method for WSN.

Related research

Intrusion detection technique for WSN has two categories: misuse detection and anomaly detection [3]. Misuse detection detects intrusion behavior by comparing suspicious character with known attack signatures which are stored in the database. Since storage capacity of sensor nodes is limited and WSN data management system is immature, it is difficult to establish complete feature database. Anomaly detection should establish profiles of normal system state and user behaviors at first. Then the data are compared with the current activities. If these are obvious biases, it implies that abnormal behavior occurs. How to distinguish between abnormal behavior and normal behavior in WSN becomes a great challenge.

Some schemes based on intrusion detection technology have been proposed for WSN in recent years. Cao et al [4] presented a simple and efficient traffic prediction model for WSN. The deviation of real traffic and forecast traffic is used to identify anomalous nodes. The disadvantage of approach is that the intrusion detection rate becomes lower when attack strength is weaker.

Phuong et al [5] presented a scheme by using CUSUM to detect anomaly. By computing the CUSUM values of the number of incoming packets and the number of outgoing packets, then comparing with the thresholds, the anomaly can be detected. But this scheme has not any evaluation for WSN. Rajasegarar et al [6] presented a distributed anomaly detection approach based on one-class quarter sphere SVM for WSN. But this method spends higher computational complexity. Ho et al [7] presented a detection method based on sequential probability. The threshold is computed by time that sensor nodes do not communicate with the surrounding nodes. This scheme can detect dynamic attack node, but the detection of static attack node is inapplicable.

In this paper, a detection approach based on sensor nodes clustering and statistical model is proposed. The sensor nodes are divided into several clusters by using k-means algorithm. A kind of anomaly detection algorithm based on statistical model is proposed to be applied to different clusters for anomaly detection.

The design of intrusion detection approach

K-means clustering algorithm. The central idea of K-means clustering algorithm is that the data objects are divided into k different clusters by iteration to minimize the objective function and to make the formation of cluster as compact and independent as possible. The K-means algorithm is described in Fig.1.

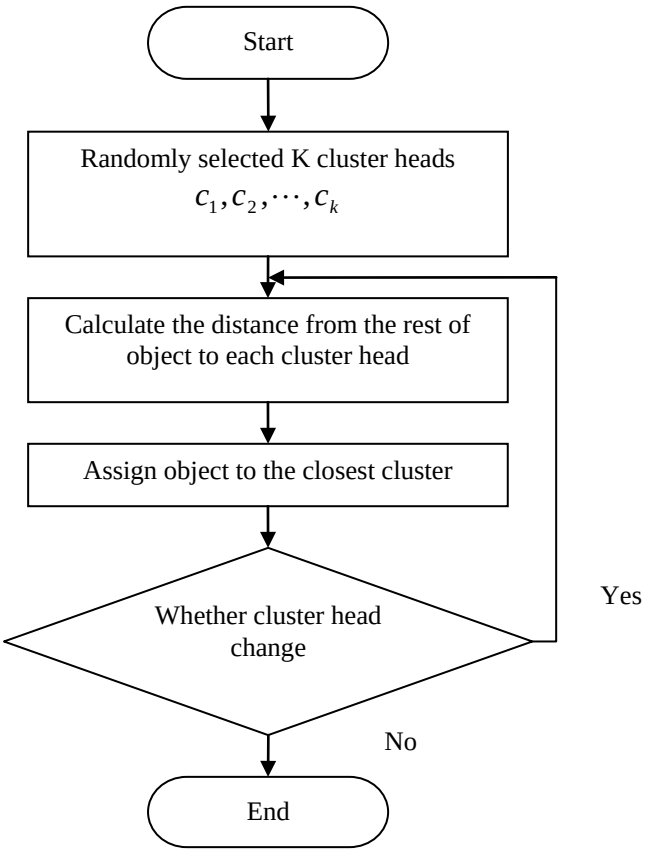

Fig.1: K-means clustering algorithm

The cluster number k can be determined based on regional division and LEACH clustering protocol:

$$k = \frac{\sqrt{N}}{\sqrt{2\pi}} \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}} \frac{L}{d^2} \tag{1}$$

where N is the number of nodes in WSN, ε_{fs} is free space attenuation amplification index of channel model, ε_{mp} is multipath fading amplification index of channel model, d is the distance from cluster head to base station in WSN, L is the side length of the square area in WSN.

Anomaly detection algorithm based on statistical model. An anomaly detection algorithm based on statistical model is applied to different clusters for anomaly detection. First, the characteristic parameters of WSN are identified. After IDS extract the characteristic parameters in the unit time, the threshold is set. The threshold is identified by statistical method. If a node behavior is in the range of threshold value, the node is normal; otherwise the node is abnormal.

Feature selection is one of the key issues in IDS. There is no linear relationship between the number of the features extracted and the performance of the detection. The high number of characteristic parameters does not imply the high intrusion detection rate. Some characteristic parameters are listed in [8], such as packet collision ratio, packet delivery waiting time, routing cost, power consumption rate, sensing data arrival rate, packet drop rate, packet reception rate. In the experiment of this paper, packet loss rate and packet reception rate of WSN are selected as the characteristic parameters of IDS. The threshold setting based on the number of packet loss and the number of packet reception in the unit time can help to detect a large number of common attacks.

The threshold setting. Many factors, for example, channel error, network congestion, electromagnetic interference, signal strength and all kinds of attacks can influence the characteristic parameters of WSN in the unit time. But each factor is unique and plays a role in the total. Thus, the characteristic parameters of unit time can express as a function based on a lot of random factors. This function can be expanded into the first order differential equation in the vicinity of certain center value, and is an algebraic sum of random variable for N factors. So we assume that the characteristic parameters of unit time can approximately obey the normal distribution.

Assume that d_n is characteristic parameters of the nth unit time and D_n is cumulative sum of characteristic parameters in n unit time. Cumulative sum of characteristic parameters in n+1 unit time is $D_{n+1}= D_n + d_{n+1}$.

So the relationship between D_n and d_n can be expressed as $D_n = \sum_{i=1}^{n} d_i$.

The average value of characteristic parameters in n+1 unit time is $\overline{D_{n+1}} = \frac{D_{n+1}}{n+1}$, while the standard deviation of sample is

$$S_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n+1}(d_i - \overline{D_{n+1}})^2} = \sqrt{\frac{1}{n}\{\sum_{i=1}^{n+1} d_i^2 - (n+1)\overline{D_{n+1}}^2\}} \quad (2)$$

Theorem 1: assume that $X_1, X_{2,} \cdots, X_n$ is a sample of the total $N(\mu, \sigma^2)$, $\overline{X}$ is the sample average and S^2 is the sample variance. So $\frac{(\overline{X} - \mu)\sqrt{n}}{S} \sim t(n-1)$

By Theorem 1, $t = \frac{(\overline{x} - \mu)\sqrt{n}}{s} \sim t(n-1)$ for a given α, check the t distribution table to get the critical value $t_{\frac{\alpha}{2}}(n-1)$, so $P[-t_{\frac{\alpha}{2}}(n-1) < t < t_{\frac{\alpha}{2}}(n-1)] = 1-\alpha$

The inequality $-t_{\frac{\alpha}{2}}(n-1) < \frac{(\overline{x} - \mu)\sqrt{n}}{s} < t_{\frac{\alpha}{2}}(n-1)$ is translated into

$$\overline{x} - t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}$$

So confidence interval of μ is

$$\{\overline{x} - t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}, \overline{x} + t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}\} \quad (3)$$

Thus, if we know the sample average $\overline{D_{n+1}}$ and cumulative sum of sample square $\sum_{i=1}^{n+1} d_i^2$, we can get observation in each unit time. In the actual situation, the standard deviation of total σ is unknown, so confidence interval of ensemble average is

$$\left\{\overline{D_{n+1}} - t_{\frac{\alpha}{2}}(n)\frac{S_1}{\sqrt{n+1}}, \overline{D_{n+1}} + t_{\frac{\alpha}{2}}(n)\frac{S_1}{\sqrt{n+1}}\right\} \quad (4)$$

If the values of the characteristic parameters satisfy the interval of formula (4) in the current unit time, the characteristic parameters is considered as normal. If not, anomaly is reported. If we want to judge whether or not the behavior of a sensor node is normal, we only need to observe certain capacity of d_{n+1}, rather than observe a large number of sample data.

The typical steps of scheme

The typical steps of the proposed scheme are as follows:

Step1: Calculate the parameters k based on the formula (1), WSN is divided into k clusters by using k-means algorithm.

Step2: Activate the IDS of cluster heads to monitor surrounding nodes in each cluster. The number of packet loss and the number of packet reception are recorded for each node. Assuming t is the unit time, and count data in every t time.

Step3: Assuming T is readiness time and m is the capacity of sample. We sample the number of packet loss and the number of packet reception in unit time after T time. Computing the confidence interval of total based on formula (4) for each cluster.

Step4: IDS continue to monitor nodes in the cluster. After t time, we count the number of packet loss and the number of packet reception for comparing with the confidence interval of step3. If the data satisfy the confidence interval, the node is considered as normal.

Simulation and analysis

The experiment use NS2 as simulation platform. A region of 100m*100m and a WSN consist of 100 nodes are simulated. All sensors are randomly deployed in the region. Setting unit time t to be 30s, the capacity of sample m to be 20 and readiness time T to be 10 minutes. The related parameters are set in Table 1.

Table 1: Simulation parameters of experimental environment

Parameter	Value
Network size(m2)	100*100
Number of nodes	100
Data packet size(Byte)	56
Rate(kbps)	19.2
Route protocol	DSR
MAC protocol	IEEE802.11

First, the parameters k is calculated and equals 8 in the experiment. So WSN is divided into 8 clusters by using k-means algorithm. Second, the confidence interval of every cluster is computed by the number of packet loss and the number of packet reception that has been recorded in readiness time. Third, black hole attack and selective forwarding attack for sensor nodes are simulated. Three kinds of approach are used for intrusion detection: the anomaly detection scheme based on Bayesian classification [9], the intrusion detection approach based on traffic prediction [10] and the method proposed in this paper. The detection effect of the three kinds of schemes is evaluated by calculating detection rate and false alarm rate. The experiment results are shown in Fig.2 and Fig.3.

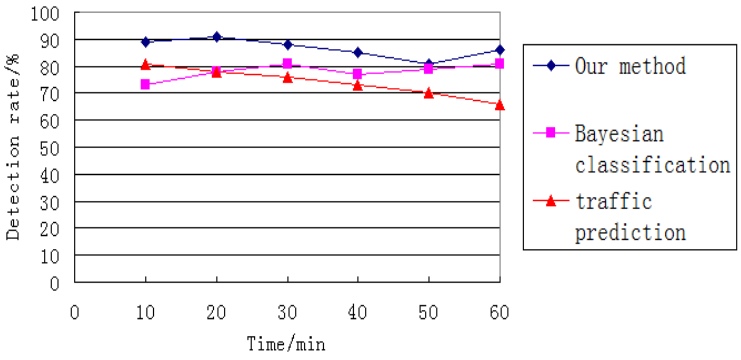

Fig.2: Detection rate

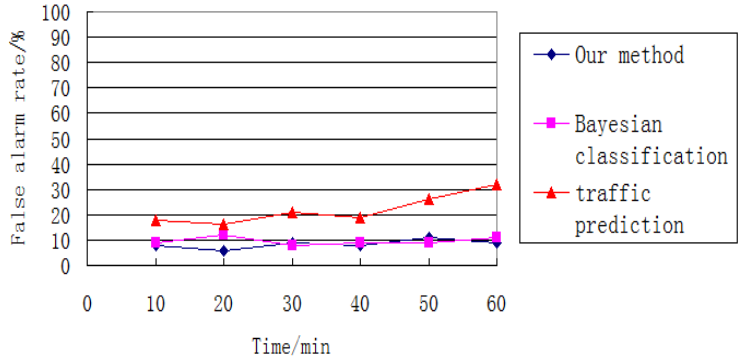

Fig.3: False alarm rate

The detection rate is the ratio between the correct number of attacks detected and the total number of attacks. The false alarm rate is the ratio between the number of a normal measurement identified as anomaly and the number of actual normal measurement. The results are known in Fig.2 and Fig.3. Our method can provide higher detection rate and lower false alarm rate than the other two schemes. K-means algorithm is a relatively good algorithm for the node clustering of WSN. The threshold setting based on statistical model is universal and spends much lower computational complexity. Therefore, the method proposed in this paper can satisfy the requirement of intrusion detection in WSN.

Conclusions

In this paper, a scheme based on sensor node clustering and statistical model is proposed. In this approach, the sensor nodes are divided into several clusters by using k-means algorithm, and then a kind of anomaly detection algorithm based on statistical model is applied to different clusters for anomaly detection. By using NS2 tool, two common attacks in WSN are simulated. The proposed method is compared with Bayesian classification and traffic prediction. The experiment results demonstrate that the proposed detection scheme provides higher detection rate and lower false alarm rate than the other two schemes.

Acknowledgements

The research work was supported by Chongqing University of Posts and Telecommunications. E-mail: sh1989315sh@163.com. The corresponding author: Hui Shen.

References

[1] Baig, Z.A. Pattern recognition for detecting distributed node exhaustion attacks in wireless sensor networks. *Computer Communications*, 34,pp.468-484,2011.

[2] Lee, J., Kapitanova, K. & Son, S.H. The price of security in wireless sensor networks. *Computer Networks*, 54(17), pp.2967-2978, 2010.

[3] Abduvaliyev, A., Lee, S. & Lee, Y.K. Energy efficient hybrid intrusion detection system for wireless sensor networks. *Electronics and Information Engineering*, pp.25-29, 2010.

[4] Cao, X.M., Han, Z.J. & Chen, G.H. Dos attack detection scheme for sensor networks based on traffic prediction. *Chinese Journal of Computer*, 30(10), pp.1798-1805, 2007.

[5] Phuong, T.V., Hung, L.X., Cho, S.J., Lee, Y.K. & Lee, S.Y. An anomaly detection algorithm for detecting attacks in wireless sensor networks. *Intelligence and Security Informatics*, pp.735-736, 2006.

[6] Rajasegarar, S., Leckie, C., Palaniswami, M. & Bezdek, J.C. Quarter sphere based distributed anomaly detection in wireless sensor networks. *Communications*, pp.3864-3869, 2007.

[7] Ho, J.W., Wright, M. & Das, S.K. Distributed detection of mobile malicious node attacks in wireless sensor networks. *Ad Hoc Networks*, 10(3), pp.512-523, 2012.

[8] Yu, Z. & Tsai, J.J.P. A framework of machine learning based intrusion detection for wireless sensor networks. *Sensor Networks*, pp.272-279, 2008.

[9] Xiao, Z., Liu, C. & Chen, C. An anomaly detection scheme based on machine learning for WSN. *Information Science and Engineering*, pp.3959-3962, 2009.

[10] Sun, B., Jin, X. & Wu, K. Integration of secure in-network aggregation and system monitoring for wireless sensor networks. *Communications*, pp.1466-1471, 2007.