www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume – 4 Issue - 12 December, 2015 Page No. 15281-15286

Intrusion Detection Alarms Filtering System Based On Ant Clustering Approach

Xiao-long XU[1] Zhong-he GAO[2] Li-juan HAN[1]

(1.Experiment Teaching Center 2.Institute of Software, Qufu Normal University, Rizhao Shandong 276826, China)

E-mail: xiaolongxu@foxmail.com

Address: Experiment Center of Qufu Normal University. No.80 of Yantai Road, Donggang District, Rizhao, China. Zip code: 276826

Abstract: With the increasing of network attacks, network information security has become an issue of global concern. The problem with the mainstream intrusion detection system is the huge number of alarm information, it has high false positive rate. This paper presents a data mining technology to reduce false positive rate and improve the accuracy of detection. The technique is unsupervised clustering method based on hybrid ANT algorithm, it can discover clusters of intruders' behavior without prior knowledge. we use K-means algorithm to improve the convergence speed of the ANT clustering. Experimental results show that our proposed approach has higher detection rate and lower false alarm rate.

Key words: intrusion detection; alarms filtering; ant clustering; false alarms

1 Introduction

With the explosive growth of various applications on the Internet, information security is getting more and more attention. People need adequate protection to deal with cyber attacks. Intrusion detection system(IDS) plays an important role in network security, it monitors all kinds of behavior within the target area, collects and checks audit data to confirm the intrusion behavior. It will issue an alert when a suspicious or malicious attempt is found, to make administrators response quickly. Intrusion detection system can be divided into host-based intrusion detection system(HIDS) and network-based intrusion detection system(NIDS). HIDS can protect a host or system, and NIDS is capable of protecting all the hosts and systems within a network[1]. NIDS uses audit data, audit trail is a record of the usage of the system, it records the user's behavior in a file. NIDS generates an alert to notify network administrators that their network is under attack[2]. However, NIDS is likely to produce a large number of alarms, false alarms will make the network administrator exhausted. So we need more intelligent intrusion analysis technology to deal with the problem of false alarms, improve the accuracy of detection.

In this paper, a system based on ant colony clustering algorithm is proposed, the ant colony algorithm is applied to intrusion detection alert filtering. An unsupervised learning method

without the need for preparatory knowledge is provided, it enables network administrators to analyze intrusions, find false positives and false negatives. By combining the ant colony algorithm with the K-means clustering algorithm, it can detect more accurately and reduce the false positive rate. Ant colony algorithm can preprocess the data and produce a series of clusters centered on the behavior of intruders. K-means clustering algorithm is used to improve the final results of ant colony module, to get more evenly divided clusters.

Experimental results show that our method can find attack vectors hidden in normal cluster. That is, it can reduce the false negatives generated by NIDS, it can even exceed other data mining technologies such as SOM and GSOM.

2 Background of ant colony algorithm

2.1 Clustering based on ant colony algorithm

Ant colony algorithm provides a powerful natural heuristic method for solving clustering problem[3], there have been many clustering algorithms based on ant behavior. Deneubourg proposed clustering ranking based on ant colony algorithm. In his model, the ants move randomly in the workspace, select or release a data element. Ants have only local perception, it is able to sense whether the surrounding objects are similar to the objects it carries. Based on these information it can perform selecting or releasing activities.

The basic model of Deneubourg can be described as follows: Data items are randomly scattered into a two-dimensional grid, each ant movies randomly in the grid, selects or releases data items. The decision to select or release data items is random, but it is affected by the neighbor data items of the ants. If the ant is surrounded by neighboring similar data, then the probability of releasing a data item is increased. On the contrary, if an ant is surrounded by dissimilar data, or if there is no data around, the probability of selecting a data item will increase. In this way, we can achieve the clustering of the elements on the two-dimensional grid.

Lumer and Faieta used different evaluation methods of local density to modify the basic model of Deneubourg, the purpose is to make it more suitable for data clustering and applied to data mining. This algorithm is called the LF model or the standard ant colony clustering algorithm. In this algorithm, each ant-like agent cannot communicate with each other, they can only perceive the similarity of objects in neighboring regions.

Lumer and Faieta introduced the concept of short-term memory for each agent. Each ant only remembers a few places where it successfully released the data items. When it chooses a new data item, it will query memory to modify the direction where the ants will be forward. So an ant tends to move to the position where it finally releases the similar data items. Lumer and Faieta defined the probability of selection and release as:

$$P_p(O_i)=\left(\frac{k_1}{k_1+f(O_i)}\right)^2 \qquad (1)$$

$$P_d(O_i)=\begin{cases}2f(O_i) & IF \quad f(O_i)<k_2 z\\ 1 & IF \quad f(O_i)\geq k_2 s\end{cases} \qquad (2)$$

$$f(O_i)=\begin{cases}\frac{1}{S^2}\sum_{O_j\in R_S(r(O_i))}1-\frac{d(O_i,O_j)}{\alpha} & If\ f>0\\ 0 & Else\end{cases} \qquad (3)$$

Where

$f(O_i)$ represents the average similar degree of the data object O_i and its adjacent data object O_j.

$d(O_i,O_j)$ represents a different degree of objects pair (O_i,O_j).

α is a factor that defines the scale of the dissimilarity.

DOI: 10.18535/Ijecs/v4i12.33

k_1 and k_2 are two constants, which are similar to k_1 and k_2 in the basic model.

2.2 AntClass algorithm

Monmarche combines the random search of ant colony algorithm with the heuristic search of K-means clustering algorithm to speed up the convergence rate of ant colony clustering algorithm. The proposed hybrid method is called AntClass and is built on the work of Lumer and Faieta. AntClass algorithm allows one ant to release multiple objects in the same unit, forming heaps of objects. Another important contribution of this algorithm is that it uses hierarchical clustering, it achieves this by allowing the ants to carry the whole heap of objects.

3 Alarm filtering system based on AntClass algorithm

In this system, AntClass algorithm is applied to solve the problem of alarm filtering. It is composed of three modules: alarm preprocessing, application of ant colony algorithm and application of K-means clustering algorithm. Figure 1 shows the general structure of the system. A variety of alarms generated by a NIDS for several computers in a period of time can be used as representative of the features of this session. To many connected machines this behavior is similar in different periods. This classification of similar behaviors in many typical behaviors can create continuous data clusters that may be a major potential attack.

3.1 Alarm preprocessing module

This module receives alerts from different NIDS installed in the network and extracts important information from these alerts. These alerts are retrieved from log files generated by NIDS, indicating that some external hosts are trying to connect to the internal hosts. The alert preprocessing module creates a number of data vectors from alerts that reflects intrusion behavior. In a certain time window, the number of type i alarms (IPintern, IPextern) is calculated. As a result,an aggregated alert data vector can be expressed as follows:

[DT, IPextern, IPintern] = #a1, #a2. . . #an

Where，

DT represents a span, IPextern represents the IP address of external host, IPintern represents the IP address of internal host. ai represents the number of alerts of type i.

In the next section we will use the results of the alert preprocessing module to create a classification of intrusion behaviors. The algorithm used is the combination of ant colony clustering algorithm and K-means clustering algorithm.

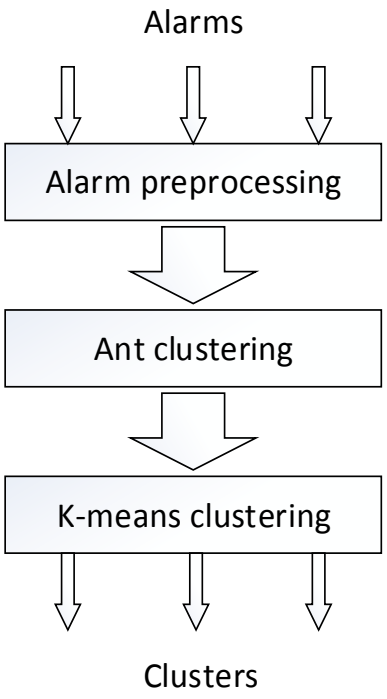

Figure 1 Structure of alarm filtering system

3.2 Application of ant colony algorithm

In order to create a preliminary distinguish of the intrusion behavior, this module receives data vectors generated by the alarm preprocessing module. We use ant colony algorithm to cluster data vectors. Ant colony algorithm is shown in Figure 2.

DOI: 10.18535/Ijecs/v4i12.33

At first, ants are randomly distributed in a two-dimensional plane, then each ant begins to move and choose or release an object(data vector)[4]. Each ant chooses a random direction, then each ant will continue to move to the original direction or to another new random direction. Each ant has a speed parameter to indicate how many steps will be moved along the selected direction. Each ant can also point out how many objects it can carry. Once the ant moves, it will choose or release an object according to the formula (1), (2). The stopping criterion of the algorithm is just the number of iterations.

1. Initialize the position of the ant
2. Repeat
3. For each anti Do
   (a) Move anti
   (b) If anti doesn't carry any object Then view adjacent cells and select an object
   (c) Else view adjacent cells and release the

Figure 2 Ant colony algorithm

3.3 Application of K-means clustering algorithm

The main advantage of the ant colony algorithm is it doesn't need any initial information about the future classification when providing the relevant data classification. However, there are two important problems in this algorithm. The first problem is that when the algorithm ends, some objects are not allocated to any heap, we call these objects as isolated objects. The second problem is that an object can take a long time to be transported to the correct cluster if it is assigned to the wrong heap. So we put forward the combination of ant colony clustering algorithm and K-means clustering algorithm to deal with these problems. We use K-means clustering algorithm to quickly remove the obvious error and provide an effective heuristic search method for the isolated object.

Ant colony algorithm provides the initial partition of the cluster, K-means clustering algorithm is used to calculate the center of each cluster and computes a new partition by allocating the nearby objects to the heap[5]. This cycle is repeated until the number of iterations is reached or the distribution has not changed in a single cycle. Figure 3 shows the K-means clustering algorithm.

1. Take data sets found by the ants as input in the form of k heaps(H1,H2,...,Hk).
2. Repeat
   (a) Calculate the center of each heap
   (b) Remove the objects of all heaps
   (c) For each object Oi
      Suppose the center of heap Hj is closest to Oi
      Assign Oi to Hj
   (d) Calculate the new heap H1 Hk' after

Figure 3 K-means clustering algorithm

4 Verification of algorithm

4.1 Implementation of algorithm

We implemented the proposed scheme using Java language, development environment is eclipse. And we apply this scheme to the Snort log file from NIDS. These files contain 32032 alert events recorded from 2014.11.20 to 2014.12.10 within 20 days. These alert events are generated by 4639 external hosts attempting to connect 289 internal hosts. The log generated by Snort contains 16 real attack scenarios and 1 non-attack scenarios. These 16 attack scenarios include: 4 kinds of POP3 brute force, 3 kinds of web crawler, 2 kinds of attacks against IIS, 2 kinds of vulnerability scanning, 1 attacks against Apache server, 3 kinds of FTP brute force, 1 SNMP attack. To evaluate our system, we divide the data set into 4 categories as shown in Table 1.

Table 1 Test and training data sets

Data set	Number of alerts	Number of data vectors	Number of windows
Test 1	167	50	11
Test 2	2816	1692	159
Test 3	11963	4001	331
Training	32032	12001	436

4.2 Experiments

The main indicators of the system we concerned are detection rate, false positive rate and false negative rate.

·Detection rate(DR): Attack behaviors detected by the system account for the proportion of total test samples.

·False positive rate(FPR): The percentage of normal behaviors mistaken for attacks.

·False negative rate(FNR): The percentage of attacks mistaken for normal behaviors.

Table 2 Result comparison

Data set	FPR	FNR	DR
Test 1	5.3%	6%	87%
Test 2	4.2%	4%	90%
Test 3	3%	3.3%	94%

Table 2 shows the FPR and FNR of data set test3 is relatively low, we can see that with the increase of the data set, the reliability of the system is improved. This also shows that adding new normal connections will make the system more similar to the actual situation of user behaviors on the network. By table2 we can see that the detection rate can be improved with the increase of the data set.

Table 3 Result comparison

Approach	FPR	FNR
AntClass	2.1%	2%
GHSOM	4.6%	4.1%
SOM	15%	14.5%

Table 3 shows the results of our proposed ant colony clustering algorithm on the test data, and compared with the other two systems GHSOM and SOM which use the same data set. The results show that our system has better performance, the scheme based on ant colony clustering algorithm has the lowest false alarm rate. The system uses ant colony clustering algorithm is better than SOM and GHSOM in FPR and FNR[6], The FPR is reduced to 2.1% and the FNR is reduced to 2%.

5 Conclusions

After theoretical analysis and practical test, the performance of intrusion detection system based on ant colony algorithm and K-means clustering algorithm has been greatly improved, such as false alarm rate is significantly reduced, the convergence rate is faster than the simple use of ant colony algorithm, and the performance in large-scale networks is better than that of small-scale networks. But the main problem we concerned about the system is alarm filtering, to improve detection accuracy by alarm filtering, and we has not paid much attention to the other performance indicators of IDS, these will be studied in the following work. In order to make the system more perfect, in the future it can be integrated with a prediction module to stop the intruder's behavior in advance.

References：

[1] Liao H J, Lin C H R, Lin Y C, et al. Intrusion detection system: A comprehensive review[J]. Journal of Network & Computer Applications, 2013, 36(1):16–24.

[2] Mao J K, Zhan F. Study on Intrusion Detection System Based on Data Mining[J]. Applied Mechanics & Materials, 2015, 713-715:2499-2502.

[3] Enxing Z, Ranran L. Research on and Implementation of Ant Colony Algorithm Convergence[J]. Electronic Science & Technology, 2013.

[4]Song L I, Jiang N. Ant algorithm model of human resource scheduling[J]. Journal of Liaoning Technical University, 2014.

[5] Zhuo-lei X. An intrusion detection model based on k-means algorithm[J]. Journal of Fuyang Teachers College, 2013.

[6] Salem M, Buehler U. An Enhanced GHSOM for IDS[C]. //IEEE International Conference on Systems, Man, & Cybernetics. IEEE, 2013:1138 - 1143.