Research of Intrusion Detection Method Based on Ant Colony Clustering

Yue Qiang[1, a *], Hu Zhongyu[2, b], Shen Shikai[1, c], Zhang Dawei[3, d]

[1] School of Information Technology, Kunming University, Kunming 650214, China

[2] School of Auto-control and Mechanical Engineering, Kunming University, Kunming 650214, China

[3] Modern Educational Technology Center, Kunming University, Kunming 650214, China

[a]wallay@126.com, [b]poundblue@126.com, [c]kmssk2000@sina.com, [d]10061405@qq.com

Keywords: intrusion detection; data mining; clustering analysis; ant colony clustering

Abstract. Network intrusion detection has been intensively investigated in recent years. In this paper, we propose an adaptive method based on ant colony clustering to discover unknown attacks. The focus of the method is the clustering process of an ant colony movement. The structure of the intrusion detection system based on ant colony clustering is designed. We use the KDD99 data set to design and evaluate our algorithm. The experimental results show the capability of our method successfully to detect network intrusions compared with the K-Means clustering algorithm. The method can not only improve the detection rate but also reduce false positive rate significantly， and can automatically detect various kinds of attacks.

1. Introduction

With the extensive growth of the Internet, the increasing availability of tools and tricks attack networks, so network security become important to information on the network and system. As a hot point and forward application area, Intrusion Detection Systems (IDSs) are designed to defend computer systems from various cyber attacks and computer viruses, and become critical components of network administration.

Data mining has the wide availability of huge amounts of data for turning such data into useful knowledge and information. Data mining has attracted a great deal of attention in the information society, and has been widely used in many application areas. The methods and technologies of data mining are applied to intrusion detection for extracting knowledge. The advantages of using a data mining approach to IDSs is that it can be used to automatically build the detection models for IDSs, so that new attacks can be recognized automatically as well [1].

As one of the best important approaches of data mining, clustering analysis is a method that divides a dataset into groups of similar objects, thereby minimizing the similarities between different clusters and maximizing the similarities between objects in the same cluster. Clustering is widely applied in data mining, such as in document clustering and Web analysis. Classic clustering approaches are the following. (1) Partitioning methods, such as K-means, K-Medoids, and K-Prototypes; (2) hierarchical methods, such as BIRCH; (3) density-based methods such as DBSCAN ;(4) grid-based methods such as STRING; (5) model-based methods, such as neural networks and Self-Organizing Map (SOM).

In recent years, ant colony clustering, which is a type of clustering algorithm that imitates the behavior of ants, has attracted researchers' attention. Ant colony clustering can be divided into two classes. The first class imitates the ant's foraging behavior, which involves finding the shortest route between a food source and the nest. This intelligent behavior is achieved by means of pheromone trails and information exchange between ants. The key elements of these algorithms are the pheromone matrix updating rule and the heuristic function. The second class imitates ants' behavior of clustering their corpses and forming cemeteries.

The rest of this paper is organized as follows. Section 2 introduces the network intrusion detection. Section 3 proposes the intrusion detection method based on ant colony clustering. Section 4 presents and analyzes the experimental results. Section 5 concludes this paper.

2. Network intrusion detection

Intrusion detection is the detection of actions that attempt to compromise the integrity, confidentiality, or availability of a resource. It attempts to detect attacks by examining various data records observed through processes on the same network. Although many different IDSs have been developed, their detection methods generally are split into two categories: misuse-based detection and anomaly-base detection.

2.1. Misuse-based detection

Misuse-based detection takes advantage of signature. This detection method searches for patterns of program or user behavior that match intrusion scenarios, which are stored as signature. The manual-coded signatures are provided by experts based on their extensive knowledge of intrusion techniques. If a pattern match is found, the IDS signal an alarm. Because data from networks or systems are dynamic, the signatures need to be updated when new software versions occur. Therefore, the major disadvantage of misuse-based detection is that it only can identify attacks that match the signatures which exist. That is, it cannot detect new or unknown intrusions.

2.2. Anomaly-based detection

Anomaly-based detection constructs models of normal network behavior which called profiles. It uses the profiles to detect new patterns that significantly deviate from normal network behavior. Such patterns maybe represent new intrusions, and can be added to the set of signatures for misuse-based detection. The main advantage of anomaly-based detection is that it can detect unknown intrusions that have not yet been identified. But anomaly-based detection has a high percentage of false positives.

3. Intrusion detection method based on ant colony clustering

The study of the behavior of real ants has greatly inspired and motivated the developments of ant colony optimization (ACO) and ant colony clustering (ACC) [2]. ACO was developed based on the nature of ants in seeking foods. Each ant randomly starts to find without any location information of food and communicates with each other by releasing a chemical substance called pheromone in the trail. If the path is short, the pheromone trail is reinforced quickly. Ant colony clustering can be divided into two categories. The first category imitates the ant's foraging behavior, which involves finding the shortest route between a food source and the nest. This intelligent behavior is achieved by means of pheromone trails and information exchange between ants. The second category imitates ants' behavior of clustering their corpses and forming cemeteries. Some ants can pick up dead bodies randomly distributed in the nest and group them into different sizes [3].

3.1. Ant colony perceptions

At the time t, the new location is randomly selected from the neighborhood with respect to a probability $P_{ij}(t)$. $P_{ij}(t)$ is the probability of an ant choosing a path from nodes i to j [4]. It is defined proportionally to an aggregation pheromone factor and a heuristic factor, i.e.,

$$P_{ij}(t) = \begin{cases} \dfrac{[\tau_j(t)]^{\alpha}[\eta_j]^{\beta}}{\sum_{l \in Neighbor(i)}[\tau_l(t)]^{\alpha}[\eta_l]^{\beta}}, & j \in Neighbor(i) \\ 0, & \\ & 0, \ otherwise \end{cases} \quad (1)$$

where $\tau_j(t)$ is the quantity of aggregation pheromone laying on the node j , η_j represents a local heuristic information between nodes i and j. Furthermore, α and β are the pheromone intensity and visibility weighting factors respectively. Finally, $j \in Neighbor(i)$ ensures that node j is a member of a set of nodes in the neighborhood of node i, which the ant has not yet visited [5].

Once each solution is formed at each iteration, aggregation pheromone trails of the nodes will be updated. More specifically, the algorithm updates the pheromone intensity for all nodes in light of the profit value of the solution; the formula of pheromone updating is as follows:

$$\tau_j(t+1) = 1 - \rho\tau_j(t) + \Delta\,\tau_j(t, t+1) \quad (2)$$

where $\tau_j(t+1)$ represents the pheromone trail of the ant j used in the iteration (t + 1), $0 < \rho \leqslant 1$ is a coefficient which represents pheromone evaporation, and the $\Delta\,\tau_j(t, t+1)$ represents the pheromone increment that ants deposited on the ant j when the iteration t ends [6].

3.2. Intrusion detection algorithm based on ant colony clustering

In this model, each object is taken as an ant. Suppose X is the dataset composed by all objects: $X = \{X = \{X_{i1}, X_{i2}, X_{i3},\ldots, X_{im}),\ i = 1, 2,\ldots,n\}$. The algorithm proceeds as follows:

(1) Initialization.

(2) Compute the weighted Euclidean distance of each data object X_i~X_j according to Eq. (3).

$$d_{ij} = \sqrt{\sum_{k=1}^{m} p_k\left(x_{ik} - x_{jk}\right)^2} \quad (3)$$

(3) Calculate the pheromone for each route according to Eq. (4).

$$\tau_{ij} = \begin{cases} 1, & d_{ij} \leq r \\ 0, & d_{ij} \geq r \end{cases} \quad (4)$$

where r is radius.

(4) Ants clustering. Compute each probability P_{ij} according to Eq. (5).

$$P_{ij} = \frac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{s\in S}\tau_{ij}^{\alpha}\eta_{ij}^{\beta}} \quad (5)$$

(5) If $P_{ij} \geq P_0$ then move X_i to an empty place in neighborhood of X_j.

(6) Compute each cluster center according to Eq. (6).

$$C_j = \frac{1}{J}\sum_{k=1}^{J} X_k \quad (6)$$

(7) Compute the similarity of data objects in each cluster according to Eq. (7).

$$D_j = \sum_{k=1}^{J}\sqrt{\sum_{i=1}^{m}(x_{ki} - x_{ji})^2} \quad (7)$$

(8) Summarize all D_J of each cluster according to Eq. (8).

$$\varepsilon = \sum_{J=1}^{k} D_J \quad (8)$$

(9) If $\varepsilon \leq \varepsilon_0$ then return else jump step 4 to iterate again.

3.3. Intrusion detection structure based on ant colony clustering

The overall structure of the intrusion detection system based on ant colony clustering is shown in Fig. 1. The system mainly consists of two functional modules, and they are data processing engine and ant colony clustering module.

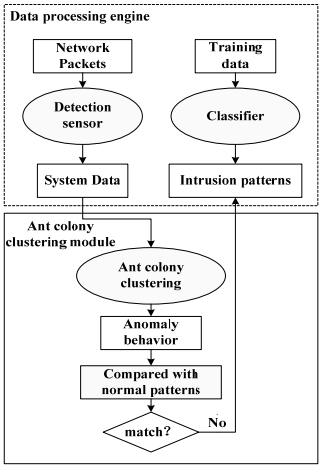

Fig. 1 structure of the intrusion detection system based on ant colony clustering

(1) Data processing engine: Firstly the system captures network packets from the transaction log on a key host according to certain strategy. Then detection sensor converts packets to data structure with a specific format, which is fit for mining. Training data are used to generate intrusion patterns according to classifier.

(2) Ant colony clustering module: System Data are converted to type object and divided into different clustering according to ant colony clustering algorithm. In intrusion detection based on ant colony clustering, each feature vector is looked as type object, as far as possible separation of normal behavior and anomaly behavior is achieved after clustering. Anomaly behavior is compared with normal patterns, if it doesn't match, and then the IDS append a new pattern into intrusion patterns storage.

4. Experimental results and analysis

4.1. Experimental design

The KDD99 data set is selected for training and testing of the IDS. As a version of the 1998 DARPA date set, KDD99 was first used in The Third International Knowledge Discovery and Data Mining Tools Competition, and now is considered as a standard benchmark for evaluation of IDSs based on data mining. In KDD99, the data records of attacks are divided into four main categories: (1) Probe, (2) DoS (Denial of Service), (3) U2R (User to Root), (4) R2L (Remote to Local). Each network connection data item in KDD99 contains 42 features, and the last feature indicates the item is a normal connection or attack.

We selected 5 data subsets from KDD99: D_1 (including 50 records), D_2 (including 100 records), D_3 (including 500 records), D_4 (including 1000 records), D_5 (including 10000 records). The distributions of all classes are shown in Table 1.

Table 1 Class distributions in network connection records.

Actual class	**D_1**	**D_2**	**D_3**	**D_4**	**D_5**
Normal	42	90	450	900	9000
Dos	2	2	15	20	300
U2R	2	2	15	20	300
R2L	2	3	10	30	200
Probe	2	3	10	30	200
Total	50	100	500	1000	10000

4.2. Performance evaluation

In order to evaluate the performance of the clustering algorithm, detection rate and false positive rate are used as the criteria.

$$\text{detection rate} = \frac{C_i}{C_m} \times 100\% \tag{9}$$

where C_i is the number of intrusion records detected exactly. C_m is the number of total intrusion records.

$$\text{false positive rate} = \frac{C_j}{C_n} \times 100\% \tag{10}$$

where C_j is the number of normal records detected as intrusion records by mistake. C_n is the number of total normal records.

4.3. Experimental results and analysis

We compare ant colony clustering (ACC) algorithm with K-Means algorithm using detection rate and false positive rate to evaluate the performance. The detection rates and false positive rates for the subsets D_1-D_5 are shown in Fig. 2 and Fig. 3.

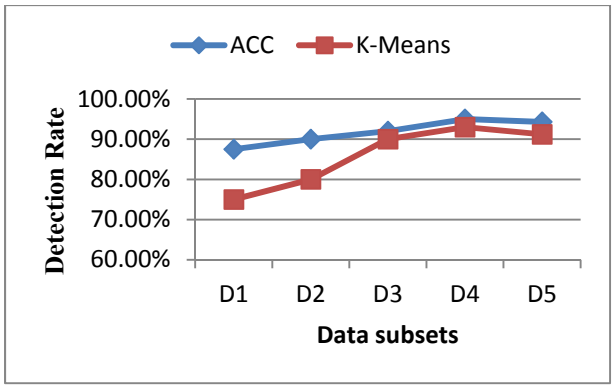

Fig. 2 Detection rate

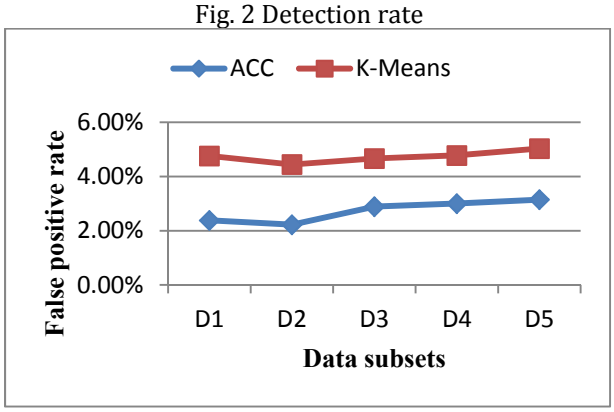

Fig. 3 False positive rate

The ACC algorithm outperforms the K-Means algorithm in our experiments with higher average detection rate and lower false positive rate. The reason of the K-Means algorithm having less detection rate is that it is not suitable for clusters of very different size. Especially when the distance between clusters is not very obvious, the clustering result is unsatisfactory. Moreover, the K-Means algorithm is sensitive to outlier data points because a small number of such data can substantially influence the mea value, and therefore it has higher false positive rate.

In order to further verify the superiority of the ACC algorithm, we selected about 50% of the 10% KDD99 data set as the testing data. The results are compared with the K-Means algorithm and shown in Table 2.

Table 2 Comparison of performance measures.

Actual class	Detection rate (%)		False positive rate (%)	
	ACC	K-Means	ACC	K-Means
Dos	93.1	90.4	3.3	5.2
U2R	92.2	90.8	2.7	7.1
R2L	90.7	88.2	4.1	7.8
Probe	95.5	91.3	3.6	4.7

It is noticed that the ACC algorithm is even better than the K-Means algorithm in terms of detection rates as well as false positive rates in various kinds of attacks. The ACC mode can be used to promote the performance of IDS in terms of efficiency and accuracy.

5. Conclusions and future work

In this paper, we have proposed an ACC algorithm with high efficiency and applied it to the intrusion detection problem. Compared with other clustering algorithms, the most advantage of ant colony clustering algorithm is it does not need any prior knowledge about clustering. It is very suitable for anomaly-based intrusion detection unsupervised clustering. The experimental results prove that it can find the unknown attacks, and has higher detection rate and lower false positive rate compared with the K-Means algorithm. As future work, we are considering combining our method with other methods in order to improving the efficiency and flexibility of IDS system.

Acknowledgement

In this paper, the research was sponsored by the Science Research Foundation Project of the Education Department of Yunnan Province (Project No. 2011Y237), Science Research Project of Kunming University (Project No. XJL15013) and Science Research Project of Kunming University (Project No. XJL14006).

References

[1] W. Lee, S.J. Stolfo, K.W. Mok, A data mining framework for building intrusion detection models, in: Proceedings of IEEE Symposium on Security and Privacy, 1999, pp. 120–132.

[2] Y.Liu, X.Yu, J.X. Huang, A. An, Combining integrated sampling with SVM ensembles for learning from imbalanced datasets, Information Processing &Management 47 (4) (2011) 617–631.

[3] S. Janakiraman, V. Vasudevan, ACO based distributed intrusion detection system, Journal of Digital Content Technology and its Applications 3 (1) (2009) 66–72.

[4] D. Martens, B. Baesens, T. Fawcett, Editorial survey: swarm intelligence for data mining, Machine Learning 82 (2011) 1–42.

[5] P.S. Shelokar, V.K. Jayaraman, B.D. Kulkarni, An ant colony approach for clustering, Analytica Chimica Acta 509 (2004) 187–195.

[6] D. Martens, B. Baesens, T. Fawcett, Editorial survey: swarm intelligence for data mining, Machine Learning 82 (2011) 1–42.