A Graph-based Clustering Algorithm for Anomaly Intrusion Detection

Zhou Mingqiang
College of Computer Science
Chongqing University
Chongqing, China
zmqmail@cqu.edu.cn

Huang Hui, Wang Qian
College of Computer Science
Chongqing University
Chongqing, China
{huanghui, wangqian}@cqu.edu.cn

***Abstract*—Many researchers have argued that data mining can improve the performance of intrusion detection system. So as one of important techniques of data mining, clustering is an important means for intrusion detection. Due to the disadvantages of traditional clustering methods for intrusion detection, this paper presents a graph-based intrusion detection algorithm by using outlier detection method that based on local deviation coefficient (LDCGB). Compared to other intrusion detection algorithm of clustering, this algorithm is unnecessary to initial cluster number. Meanwhile, it is robust in the outlier's affection and able to detect any shape of cluster rather that the circle one only. Moreover, it still has stable rate of detection on unknown or muted attacks. LDCGB uses graph-based cluster algorithm (GB) to get an initial partition of data set which is depended on parameter of cluster precision rather than initial cluster number. On the other hand, because of this intrusion detection model is based on mixed training dataset, so it must have high label accuracy to guarantee its performance. Therefore, in labeling phrase, the algorithm imposes outlier detection algorithm of local deviation coefficient to label the result of GB algorithm again. This measure is able to improve the labeling accuracy. The detection rate and false positive rate are obtained after the algorithm is tested by the KDDCup99 data set. The experimental result shows that the proposed algorithm could get a satisfactory performance.**

Keywords; intrusion detection;Graph-based clustering;cluster precision; outlier detection;labeling accuracy

I. INTRODUCTION

As the increase of the significance of computer networks in modern society, its security becomes one of the hottest issues to be solved. Therefore, it is extremely imperative to find an effective way to protect this valuable network infrastructure. There is imperative requirement to protect our computers from unauthorized or malicious actions. And intrusion detection system is a useful tool for detecting attacks. After Denning [1] introduced the first intrusion detection model to find these behaviors which are different from users', many approaches , that is to address the problems of IDS, have been proposed, such as machine learning [2], immunological [3] and data mining. Among these techniques, data mining has been widely used technology and successful in solving the deficiencies existed in intrusion detection and prevention systems by discovering users' behaviors from massive data. Wenke Lee presented an improved method of RIPPER which lead to the set of association rules and frequent episode patterns generated is easy to understand [4]. Besides, due to the bottleneck of frequent items of association rule-based Apriori algorithm, a Length-decreasing support could solve this problem [5]. In order to classify the high dimension data, GA algorithm is used to select a value subset of input features for decision tree classifiers [6].

Clustering is the method of grouping objects into meaningful subclasses so that members from the same cluster are quite similar and members from different clusters are quite different from each other. Therefore, clustering methods can be useful for classifying log data and detecting intrusions. Clustering intrusion detection is based on two assumptions. The first assumption is that the number of normal action is far greater than the number of intrusion action. The second assumption is that the intrusion action makes a difference with the normal action. Based on these two premises, many clustering techniques for anomaly intrusion detection have been proposed. However, these cluster-based IDS have many drawbacks: k-means is used for intrusion detection to detect unknown attacks and partition large data space effectively [7]. But it has two shortcomings: number of cluster dependence and degeneracy. To gain an optimal k is a NP hard problem. Ali and Yu guan presented a heuristic k-means algorithm called Y-means [8]. Otherwise, Wei used improved FCM algorithms [9] to obtain an optimized k. A. Prabakar [10] combined with C4.5 decision tree to improve performance of k-means intrusion detection system. However, all kinds of these algorithms should initial the cluster number k. Besides, other methods also have their own shortcomings, such as Simulated annealing combined with clustering algorithm [11] requires so a lot of training data that consume excessive resource. In recent year, many researchers have presented many effective methods on similarity and classification model. One of the important one is graph clustering. For example, PBS algorithm [12] introduced a measurement method of data points similarity based on the approximate function to improve the accuracy of graph-based clustering.

In this paper, we present LDCGB algorithm for intrusion detection. The major contributions of current work are two folds. First, we deploy graph-based clustering algorithm (GB) into intrusion detection successfully. GB algorithm could

detect any shape of cluster and do not have to give cluster number because it just uses a parameter of cluster precision to replace initial cluster number. A relatively good clustering result would be received by altering this parameter. However, the partition precision of GB algorithm is hardly to satisfy the requirement of IDS. So it just provides an input for the next step. And then, we apply outlier detection algorithm based on local deviation coefficient to improve the accuracy of data labeling. We demonstrate the power of this algorithm by testing it over KDD1999 data set.

The rest of paper is organized as follow. In next section, we introduce the graph-based clustering. In section 3, we describe the LDCGB algorithm in detail. In section 4, we describe our evaluation methods and experimental results. Finally, we summarize our contributions and point out our future work in this field.

II. GRAPH-BASE CLUSTER ALGORIGHM

First, Graph-based clustering algorithm [13] is a method commonly used in automatically partition for a data set in several clusters. It proceeds by setting a parameter of clustering precision to control the result of clustering. Records in dataset are packaged as a note. These notes are treated as vertex of a complete undirected graph, and the distance values between these notes as weight of the edge. The distance is calculated by Euclidean distance function. According these values of distance, we could construct a distance matrix ***I.*** And the threshold δ is computed by a parameter of cluster precision α.

$$\delta = dismin + (dismax - dismin) \times ClusterPrecision \qquad (1)$$

dismin and dismax represent the minimum and maximal value of matrix ***I*** respectively. So an edge is cut down from this graph if its value of weight greater than threshold δ in Fig.1.

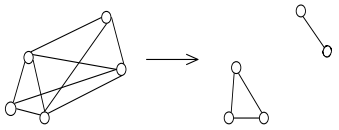

Figure 1. GB cluster

Finally, transverse the whole graph, the notes would be classified into the same cluster if there is edge between them. Therefore, several sub-graphs are created. Each sub-graph represents a cluster. Finally, outliers are processing. The steps of this algorithm is showed in Fig.2

```
1 For all i in RECORDSET DO{
2 Record i is packaged as a note
3 Put note i into GRAPH
4 Repeat {
5    calculate threshold δ by function (1)
6 cut down all the edges whose value is greater than
  the threshold δ
7 transverse GRAPH, label all the sub-graphs.
8 outlier processing
9          } until the outlier is processed completely
```

Figure 2. GB clustering algorithm

GB algorithm has been used for clustering for decades. However, it mainly has two shortcomings when it is applied for intrusion detection: the first one is that it distinguishes the normal and abnormal cluster just by a value of threshold. So the clustering accuracy is far from enough. Second, it doesn't offer a reasonable method to address outliers, but just simply throw it away. With this coarse granularity partition, it can't receive a satisfied detection rate. On the other hand, the ability to detect any shape of cluster is made it very suitable for the dataset with complex shape in real network.

III. GRAPH-BASED CLUSTER ALGORITHM FOR INTRUSION DETECTION

The In order to receive higher detection rate, we further propose an improved Graph-based clustering algorithm by using outlier detection method based on local deviation coefficient in label process. This method mainly focuses on how to classify the data on the boundary to be classified more accurately and then augment the difference between normal and abnormal clusters. Firstly, we define some related definitions:

Definition 1: (k distance of an object p)

For any positive integer k, the k-distance of object p, denoted as k-distance (p), is defined as the distance d (p, o) between p and an object o ∈D such that:

- For at least k objects o' ∈D \ {p}, it holds that d(p, o') ≤d(p, o)
- For at most k-1 objects o' ∈D \ {p}, it holds that d (p, o') < d (p, o).

Definition 2 :(k-distance neighborhood of an object)

Given the k-distance of p, the k-distance neighborhood of p contains every object whose distance from p is not greater than the k-distance.

N k-distance (p) = {q ∈D\ {p} | d (p, q) < k-distance (p)}.

These objects q are called the k-nearest neighbors of p.

Definition 3 :(local deviation rate of an object)

Given the k-distance of p, and p is a center of circle with radius k. All objects in this circle are k-distance neighborhood of p. p' is the centre of mass of this circle. So the local deviation rate is defined as:

$$LDC_{k(p)} = dis(p, p') \Big/ \left| N_{k-dis\tan ce(p)} \right| \qquad (2)$$

The dis(p, p') is the distance between object p and centre of mass of p'.

Definition 4: (local deviation coefficient of an object)

Given the k-distance neighborhood of p and LDR, the local deviation coefficient is defined as:

$$LDC_{k(p)} = \sum_{o \in N_{k-dis\tan ce(p)}} LDR_{k(o)} \Big/ \left| N_{k-dis\tan ce(p)} \right| \quad (3)$$

Intuitively, molecule is sum of the LDR of k distance neighborhood of p. The coefficient reflects the degree of dispersion of an object's neighborhood. Greater value of LDC means higher probability of one object being an outlier. On the other hand, a low LDC value indicates that the density of an object's neighborhood is high. So it's hardly to be an outlier. According to these definitions, the LDCGB algorithm is described as follow:

Step1: implement GB algorithm to cluster dataset and gain n clusters C1, C2…Cn, they are sorted in descending order according to the records they embraced.

Step2: initialize CN= {φ}, CS= {φ}, CA= {φ},

Step3: For i =1 to n

IF (C1.num+C2.num…Ci.num>λ1*M)

THEN CN={C1,C2…Ci-1},

IF (Cn+Cn-1..Cj+1 >λ2*M)

THEN CA = {Cj+1 … Cn}.

The remaining cluster is classified into CS {Ci…Cj}. End for.

Step4: compute LDC of every object p by the function (2) and (3), p ∈CS, sorted these values in descending order. The first k records are classified in CA, and the rest are classified in CN.

Step5: the data in CN are labeled as normal, while in CA, they are labeled as abnormal. After all data are labeled, the labeling process is over.

In this process, CN, CS and CA stand for the set of normal clusters, suspicious clusters and abnormal clusters respectively. CS is the set that need to be processed in next step. In step 3, M is the number of data set and λ1, λ2 (λ1+λ2=1) represent the percentage of normal and anomaly rate. They should meet the premise that the number of normal action is far greater than the number of intrusion action. So their values must satisfy λ1>>λ2. Otherwise, the isolated points were classified in abnormal clusters rather than discard them away. In detecting phase, a new record d, calculate its distance to each data, it will belong to the cluster the same to the data that has nearest distance with it. If the cluster is normal, d is normal. Otherwise, d is an attack.

IV. EXPERIMENT

A. Experiment data set

To evaluate the performance of LDCGB approach, a series of experiments were conducted KDDCup99 data set [14]. The dataset is a dedicated test dataset established for intrusion detection assessment by Massachusetts Institute of Technology. It contains 24 kinds of attacks that can be category into 4 types: Denial of Service, Remote to User, User to Root and Probing. In the dataset, a record has 7 classified attributes and 34 numeric attributes, and this belongs to the implementation of clustering in high-dimensional space.

B. Data preprocessing

The In order to improve the detection efficiency of the experiment, we remove the attributes that is useless for this experiment. After carefully analysis, we screen 20 properties as the objects of study, such as the lifetime of the TCP, window size and the length of the packet. On the other hand, preventing the problem that large numbers eliminate the effect of small ones, we do the following transformation:

- Calculating mean absolute deviation S_f

$$S_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|) \quad (4)$$

Here, $x_{1f}, x_{1f} ... x_{nf}$ are the n features of f, m_f is the average of f.

- Calculating the standardize feature property Y_{if} :

$$Y_{if} = (X_{if} - m_f) \Big/ S_f \quad (5)$$

C. Performance Analysis

In this experiment, we implemented and evaluated the proposed method in VS 6.0 on Windows PC with Duro-Core 2.0GHz and 2GB RAM. And then, we select 10000 samples for training data set. Otherwise, testing datasets are divided into group 1,2,3,4. Each group contains 2000, 4000, 5000 and 10000 records respectively. Besides, we select 2500 samples of intrusion which types are different from the training dataset. It is aim to evaluate ability of this algorithm on detecting unknown attacks. First, we alter the cluster precision of α. The result is showed in TBALE I:

TABLE I. CLUSTERING RESUT OF GB

Cluster precision (α)	Cluster number(n)
0.02	21
0.05	9
0.20	6
0.50	4

At Table I, we observe clearly that the change of cluster number with altering parameter of cluster precision. A relatively large α will lead to small clusters number. As a result, excessive data would be classified in one large class. And most of the abnormal behaviors can't be detected in this situation. On the other hand, with a small value of α, the partition will generate excessive clusters.

The next step, for the GB model, the best situation is that all data were divided into 9 subsets. So we fixed α = 0.05. To meet the one of premise of anomaly intrusion detection that normal action is far greater than the number of intrusion action, we try the parameter of λ1 and λ2 in (0.9,1.0) and (0.0,0.1) respectively. Finally, we change the values of parameter k and testing these constructed models by group 3. We find that, when k=9, λ1=0.95 and λ2=0.05, the performance of this

algorithm is best. The output of detection rate (DR) and false positive rate (FPR) showed in TABLE Ⅱ

TABLE II. THE PEFORMANCE OF LDCGB

k	DR	FPR
5	93.30%	2.24%
8	95.3%	2.08%
11	92.67%	2.14%
15	92.0%	2.1%

The experimental result shows that the parameter k is an important factor for the performance of this algorithm. The value of k should not be too large because a large k cause many isolated points to be classified into normal classes. On the other hand, a relative small value of k will make the algorithm uselessness because most of records have large value of LDC. So the abnormal records separated from suspicious cluster could not to be classified correctly. Both of these situations would decrease cluster precision. So when k is assigned 15, the detection rate is reduced.

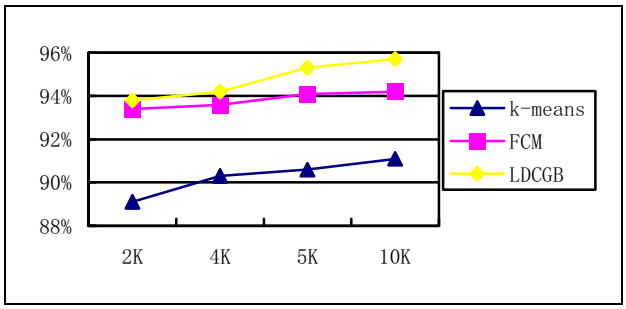

Figure 3. Mixed attacks of different algorithms

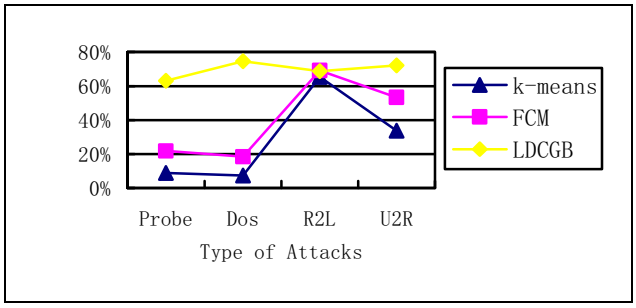

Figure 4. Unknown attacks of different algorithms

To further prove the superiority of LDCGB algorithm, we compare its performance with some well-known methods by group 1,2,3,4 of testing set. The result reveals that the detection rate of LDCGB is higher that these traditional cluster-based intrusion detection algorithm. The comparison of performance shows in Fig.3. Otherwise, the performance of detecting unknown attacks, as shown in the Fig.4, it's obvious that LDCGB have relatively stable and high detection rate on the unknown attacks. LDCGB could detect any shape of cluster and with a high labeling precision that is why it could detect more unknown attacks. But for the R2L and U2R attacks, the detection is comparatively low because there some types of U2R and R2L attacks are the same as the normal behaviors which make them impossible for the system to differentiate.

V. CONCLUSIONS

Intrusion detection system based on data mining increases the intelligence and reliability of network. Obviously, by means of clustering method, intrusion detection may be carried out. The LDCGB algorithm presented in this paper may overcome some disadvantages of the traditional cluster algorithm for intrusion detection and can obtain comparative satisfactory performance of intrusion detection. However, there are still have many deficiencies needed to be improved. Our further research will focus on how to reduce the complexity of this algorithm because the memory requirement for computation increase dramatically as growth of record's number. Another disadvantage should be improved is that the initial percentage of abnormal and normal records need manual control to find the suspicious clusters, it more or less influences performance of this algorithm.

REFERENCES

[1] D. Denning, An intrusion-detection model, In IEEE computer society Symposium on research security and privacy ,1986,pp,118-131

[2] Savage, Etherall D, Karlin A, et a1. Network Support for IP Trace-back. IEEE/ACM Transactions on Networking, 2001, 9(3)

[3] Dasgupta,D, Gonzalez, F. An Immunity-Based Technique to Characterize Intrusions in Computer Networks. IEEE Transations. evol. computer. 6(3), 1081–1088 (2002)

[4] Lee W· Scolfo S J. Data mining approaches for intrusion detection. In Pro.of the 7th USENIX Security Symposium .San Antonio TX,Jan 1998

[5] L. Li, D. Yang, and F. Shen, "A novel rule-based IntrusionDetection System using data mining", Proceedings of 2010 IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol.6, pp.169-172, July 2010.

[6] G. Stein, B. Chen, A. S. Wu, K. A. Hua, "Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection", in Proceedings of the 43rd ACM Southeast Conference, Kennesaw, GA, March 2005.

[7] M.Jianliang, S.Hai kun, B.Ling. The Application on Intrusion Detection Based on K-means Cluster Algorithm. International Forum onInformation Technology and Application.2009

[8] Yu Guan, Ali A. Ghorbani, and Nabil Belacel. Y-means: a clustering method for intrusion detection. In Canadian Conference on Electrical and Computer Engineering, pages 14, Montral, Qubec, Canada, May 2003.

[9] Wei Jiang, Min Yao, Jun Yan. Intrusion detection based on improved fuzzy c-means algorithm. Information science and engineering, 2008, ISIS

[10] Amuthan Prabakar Muniyandia, R. Rajeswarib, R. Rajaramc. Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm. International Conference on Communication Technology and System Design. Procedia Engineering 30 (2012):174-182

[11] Lin Ni, Hong-Ying Zheng. "An unsupervised intrusion detection method combind clustering with chaos simulated annealing". Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.

[12] W.Guo hui, L.Guo yuan. "Intrusion detection method based on graph clustering algorithm" . Journal of Computer Applications,July2011: 1888-1900

[13] Yang Liu. "GB-Cluster: a Graph-based Clustering Algorithm". Computer science,2002.

[14] KDD.KDDCup1999Data.http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999.