LDFGB Algorithm for Anomaly Intrusion Detection

Shang-nan Yin, Zhi-guo Chen, and Sung-Ryul Kim

Division of Internet and Multimedia Engineering
Konkuk University, Seoul, Rep. of Korea
{yinshangnan,chenzhiguo520}@gmail.com, kimsr@konkuk.ac.kr
http://www.konkuk.ac.kr

Abstract. With the development of internet technology, more and more risks are appearing on the internet and the internet security has become an important issue. Intrusion detection technology is an important part of internet security. In intrusion detection, it is important to have a fast and effective method to find out known and unknown attacks. In this paper, we present a graph-based intrusion detection algorithm by outlier detection method which is based on local deviation factor (LDFGB). This algorithm has better detection rates than a previous clustering algorithm. Moreover, it is able to detect any shape of cluster and still keep high detection rate for detecting unknown or known attacks. LDFGB algorithm uses graph-based cluster algorithm (GB) to get an initial partition of dataset which depends on a parameter of cluster precision, then we use the outlier detection algorithm to further processing the results of graph-based cluster algorithm. This measure is effective to improve the detection rates and false positive rates.

Keywords: Graph-based clustering, outlier detection, Intrusion Detection.

1 Introduction

In modern society, we use internet at anytime and anywhere, so internet security becomes one of the hottest issues and we need to find an effective way to protect this network infrastructure. Intrusion detection system [1] is a useful method for detecting attacks. In 1987, Denning [2] introduced the first anomaly intrusion detection model, which is able to detect known and unknown attacks. After that research, many methods have been proposed for anomaly intrusion detection, such as machine learning [3], immunological [4] and data mining. Among these techniques, data mining has been widely used and it successfully solves the deficiencies of intrusion detection. The clustering algorithm is an important technology of data mining which can offset these deficiencies.

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (same attributes) to each other than to those in other groups (clusters). K-means algorithm [5]

Linawati et al. (Eds.): ICT-EurAsia 2014, LNCS 8407, pp. 396–404, 2014.

is a popular clustering algorithm which is used in anomaly intrusion detection. This algorithm classifies similar data set into same clusters, and classifies dissimilar data set into different clusters. However, the user must set the number of clusters k, meaning that the user has to have some knowledge about the data. Other methods also have some disadvantages, for example, combining simulated annealing and clustering algorithm [6] requires a lot of training data and thus consumes excessive resource. In recent year, many researchers have proposed to avoid excessive resource consumption. One of effective methods is graph-based clustering. For example, PBS algorithm [7] introduces a measurement method of data points similarity which is based on an approximate function. But, this algorithm cannot achieve an exciting detection rate. LDC algorithm [8] improves the detection rate, but it doesnt accurately and comprehensively analysis the distribution situation of data nodes.

In this paper, we present LDFGB algorithm for intrusion detection which is one of the graph clustering algorithms. It is based on the LDC algorithm, uses local deviation factor to differentiate the distribution situation of data nodes and to identify outliers. The experimental results show that the proposed method is efficiently for anomalybased intrusion detection.

The paper is organized as follows. In section 2, we introduce the graph-based clustering. In section 3, we describe the LDFGB algorithm in detail. In section 4, we describe our evaluation methods and experimental results. Finally, we conclude this paper.

2 Graph-Base Cluster Algorithm

Graph-based clustering algorithm is a method commonly used in automatic partitioning of a data set into several clusters. It proceeds by setting a parameter of clustering precision to control the result of clustering. Records in dataset are packaged as a node. These nodes 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 (Table 1).

$$d(i,j) = \sqrt{|x_i1 - x_j2|^2 + |x_i2 - x_j2|^2 + ... + |x_ip - x_jp|^2} \quad (1)$$

Table 1. Euclidean distance

	Id1	**Id2**	**Id3**	**.....**	**Idn**
Id1	0	0.81	0.11	.....	0.02
Id2	0.81	0	0.45	.....	0.71
Id3	0.11	0.45	0	.....	0.15
.....	.....	.....	.....	.....	.....
Idn	0.02	0.71	0.15	.....	0

According these values of distance, we construct a distance matrix I. And the threshold δ is computed from a parameter of cluster precision α.

$$\delta = dismin + dismax - dismin * Clusterprecision \tag{2}$$

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 as shown in Figure.1

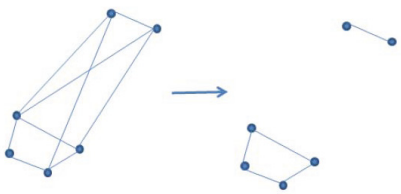

Fig. 1. GB Cluster

Finally, we transverse the whole graph, the nodes would be classified into the same cluster if there is an edge between them. Therefore, several sub-graphs are created. Each sub-graph represents a cluster. Finally, outliers are processed.

The Steps of GB Algorithm Are as Follows:

```
Input: Dataset (record set), Cluster Precision
Record I is packaged as a note
Put note I into Graph
Repeat {
   Calculate threshold (delta) by function
   Cut down all the edges whose value is greater than the
   threshold (delta)
   Transverse Graph, label all the sub-graphs.
   Outlier processing
} until the outlier is processed completely
```

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 doesnt offer a reasonable method to address outliers, but it just throws them away. With this coarse granularity partition, it cannot achieve a satisfactory detection rate. On the other hand, the ability to detect any shape of cluster has made it very suitable for the dataset with complex shape from real network.

3 LDFGB Algorithm for Intrusion Detection

In order to achieve higher detection rate, we further propose an improved graph-based clustering algorithm by using outlier detection method based on local deviation factor 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: (Outlier). These can be objects that are outlying relative to their local neighborhoods, particularly with respect to the densities of the neighborhoods. These outliers are regarded as local outliers. (Figure 2)

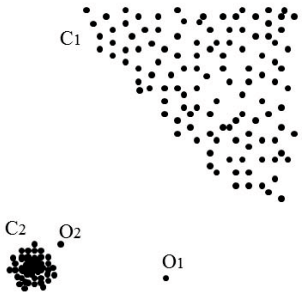

Fig. 2. O_1 and O_2 are outliers

Definition 2: (k distance of an object p). For any positive integer k, the k-distance of object p, denoted as k-distance (p), is defined to be 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 3: (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.

$$N_{k-distance}(p) = \{q \in D \setminus \{p\} \,|d(p,q) \leq k - distance(p)\}$$

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

Definition 4: (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 to be:

$$LDR_{k(p)} = \frac{dis\,(p, p')}{|N_{k-distance(p)}|} \tag{3}$$

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

Definition 5: (local deviation influence rate of an object). Given the k-distance neighborhood of p and LDR, the local deviation influence rate is defined to be:

$$LDIR_{k(p)} = \frac{\sum_{o \in N_{k-distance}}^{LDR_{k(o)}}}{|N_{k-distance(p)}|} \tag{4}$$

Definition 6: (local deviation factor of an object). Given LDR and LDIR, local deviation factor is defined to be:

$$LDF_{k(p)} = \frac{LDR_{k(p)}}{LDIR_{k(p)}} \tag{5}$$

The local deviation factor of object p reflects k-distance neighborhood of object p within the dispersion degree. High value of LDF means higher probability of one object being an outlier; a low LDF value indicates that the density of an objects neighborhood is high. So its hardly to be an outlier.

The Steps of LDF Algorithm Are as Follows:

```
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 ={}, CS ={}, CA ={},
Step3: For i =1 to n
          IF (C1.num+C2.num...Ci.num> (lambda)2*M),
          THEN CN={C1,C2...Ci-1},
          IF (Cn+Cn-1...Cj+1 >(lambda)2*M)
          THEN CA = {Cj+1...Cn}.
       The remaining cluster is classified into CS {Ci...Cj}.
       End for.
Step4: compute LDF of every object p by the function (3) (4)
       and (5), 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 $\lambda 1$, $\lambda 2$ ($\lambda 1+\lambda 2=1$) 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 $\lambda 1 \gg \lambda 2$. 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.

4 Experiments and Results

To evaluate the performance of LDFGB approach, a series of experiments was conducted on a 2-dimensional artificial dataset. (Figure 3)

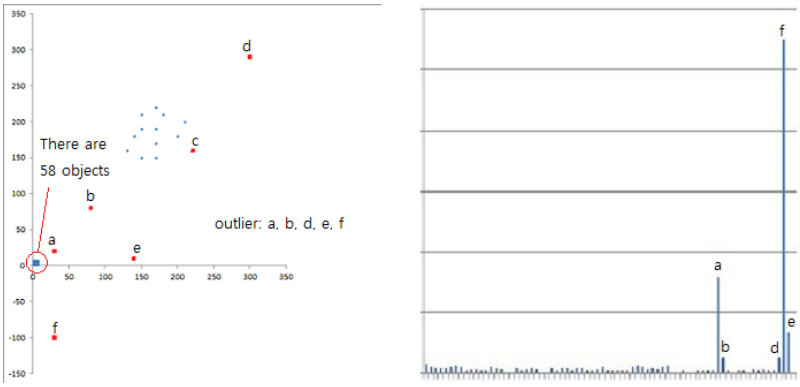

Fig. 3. 2-dimensional dataset and LDF values

Figure 4 shows that the data nodes distributed in two kinds of situations. Similar results are obtained by the LDC algorithm, so the LDC algorithm cant differentiate these kinds of situations. If we use the proposed method (LDFGB Algorithm) by adjusting K values that we can differentiate two situations. So the performance of LDFGB is the better than LDC algorithm.

KDDCup99 dataset is a dedicated test dataset established for intrusion detection assessment by Massachusetts Institute of Technology. It contains 24 kinds of attacks categorized 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. The In order to improve the detection efficiency of the experiment, we remove the attributes that is useless for this experiment. After careful 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.

We randomly select 10000 samples for training data set. Besides, we randomly select 2500 samples of intrusion which types are different from the training dataset. It is aimed to evaluate ability of this algorithm on detecting unknown attacks. First, we alter the cluster precision of α. The result is shown in Table 2:

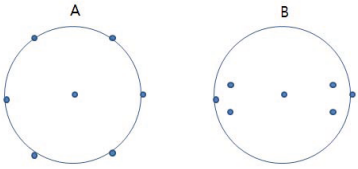

Fig. 4. Two kinds of distribution

Table 2. Clustering result of GB

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

On Table 2, we observe clearly that the change of cluster number with altering parameter of cluster precision. A relatively large α will lead to small number of clusters. As a result, excessive data would be classified in one large class. And most of the abnormal behaviors cant 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 $\lambda 1$ and $\lambda 2$ 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, $\lambda 1$=0.95 and $\lambda 2$=0.05, the performance of this algorithm is the best. The output of detection rate and false positive rate showed in Table 3:

The LDC algorithm output of detection rate and false positive rate showed in Table4:

We discover the LDFGB algorithm achieves better detection rate than the LDC algorithm. Table 3 and Table 4 show that when there are the same K values, the performance of LDFGB algorithm is superior to the LDC algorithm. The experimental result shows that the parameter K is an important factor for

Table 3. The performance of LDFGB

K	Detection rate	False positive rate
5	94.30%	2.26%
8	96.14%	2.09%
12	97.25%	2.03%
15	96.52%	2.13%

Table 4. The performance of LDC

K	Detection rate	False positive rate
5	93.30%	2.24%
8	95.31%	2.08%
12	92.67%	2.14%
15	92.00%	2.11%

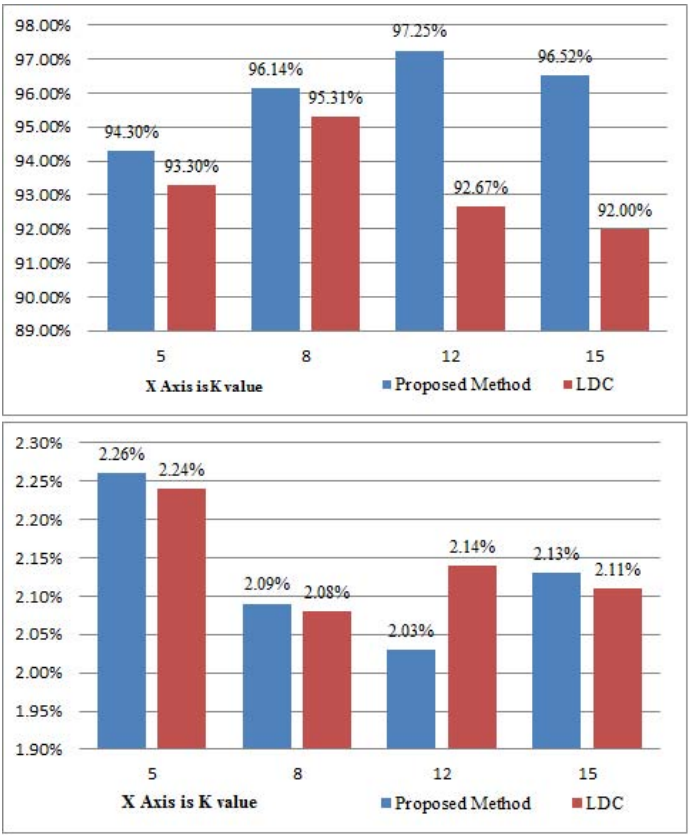

Fig. 5. The preferment of our method and LDC method

the performance of this algorithm. We should not set a value for K that is too large, because a large K value will cause many isolated points to be classified from normal classes. On the other hand, a relatively small value of K will lead to the most of records have a large LDF value. Therefore, we could not separate the abnormal records from suspicious cluster. Both of these situations would decrease cluster precision. Figure 5 shows that the proposed method always outperforms the LDC method.

5 Conclusions

Intrusion detection system based on data mining increases the safety and reliability of network. Obviously, by means of clustering method, intrusion detection may be carried out. The LDFGB 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 many deficiencies that need to be improved. Our further research will focus on how to reduce the complexity of this algorithm because the memory requirement for computation increases dramatically as the number of records grow. Another disadvantage which should be fixed 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.

Acknowledgments. This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2011-0029924).

References

1. http://www.sans.org/reading_room/whitepapers/detection/understanding-intrusion-detection-systems_337
2. Denning, D.E.: An intrusion-detection model. In: IEEE Computer Society Symposium on Research Security and Privacy, pp. 118–131 (1987)
3. Savage, S., Wetherall, D., Karlin, A., Anderson, T.: Network Support for IP traceback. IEEE/ACM Transactions on Networking, 226–237 (2001)
4. Dasgupta, D., Gonzalez, F.: An Immunity-Based Technique to Characterize Intrusions in Computer Networks. IEEE Trans. Evol. Comput. 6(3), 1081–1088 (2002)
5. Kaufan, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York (1990)
6. Ni, L., Zheng, H.-Y.: An unsupervised intrusion detection method combined clustering with chaos simulated annealing. In: Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, vol. 1922 (August 2007)
7. Guohui, W., Guoyuan, L.: Intrusion detection method based on graph clustering algorithm. Journal of Computer Applications, 1888–1900 (July 2011)
8. Mingqiang, Z., Hui, H., Qian, W.: A Graph-based Clustering Algorithm Intrusion Detection. In: The 7th International Conference on Computer Science & Education (ICCSE), pp. 1311–1314 (2012)