IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

Clustering Based Anomaly Detection for Intrusion Detection

Ms. Nayana VM

PG Scholar, Computer Science and Engineering, LBS College of engineering, Kasaragod, India

Abstract: The intrusion detection is done in the data mining by means of using the clustering technique. Due to the risks of average clustering ways for intrusion detection, I am performing a graph-based intrusion detection algorithm with the aid of utilizing outlier detection ways. Compared to other intrusion detection algorithm of clustering, this algorithm is mindless to preliminary cluster quantity. In the meantime, it is strong within the outlier's detection and capable to notice any shape of cluster alternatively that the circle one best. This paper makes use of graph-based cluster algorithm (GB) to get an initial partition of knowledge set to valid clusters by using an precision parameter. On the other hand, since of this intrusion detection mannequin is cantered on mixed training dataset, so it must have high label accuracy to assurance its efficiency. Hence, in labelling phrase, the algorithm imposes outlier detection algorithm to label the influence of GB algorithm once more. This measure is equipped to reinforce the labelling accuracy.

Keywords: clustering, intrusion detection, anomaly based, graph algorithm, DBSCAN

I. INTRODUCTION

Intrusion detection methods (IDSs) are monitoring contraptions which have been added to the wall of security with a view to avert malicious recreation on a approach. This work specializes in network intrusion detection techniques (NIDSs) customarily in view that they can detect the widest variety of assaults in comparison with different forms of IDSs. Network IDSs analyse traffic to notice on-going and incoming assaults on a network. Nowadays, business IDSs most commonly use a database of ideas, called signatures, to take a look at to realize attacks on a network or on a number laptop. Intrusion detection techniques are monitoring devices which might be used to detect intrusions on a pc or a network. An intrusion detection process is an imperative instrument for network directors due to the fact that without such a gadget, it might be impossible to research the big amount of packets traversing current networks each second. After greater than thirty years of intensive study on intrusion detection systems, the subject remains to be open to additional investigations specially concerning the accuracy of the detection. Moreover, versions of identified assaults as well as new attacks can most of the time go by means of the procedure without being detected.

The targets of the IDS furnish the specifications for the IDS policy. Competencies goals comprise:

- Detection of assaults
- Prevention of attacks
- Detection of coverage violations
- Enforcement of use policies
- Enforcement of connection policies
- Assortment of evidence
- Following are the explanations for the dimension of IDS.
- False positive (FP): Or false alarm, Corresponds to the quantity of detected attacks but it is in fact ordinary.
- False negative (FN): Corresponds to the quantity of detected typical occasions but it's simply attack, in other words these attacks are the goal of intrusion detection systems.
- true positive (TP): Corresponds to the quantity of detected assaults and it is in fact attack.
- true negative (TN): Corresponds to the number of detected average occasions and it is virtually normal.

Mainly there are two types of intrusion detection

Anomaly Detection

Anomaly based detection techniques rely on competencies of usual behaviour to realize any assaults. Hence attacks, together with new ones are detected as long as the assault habits deviates sufficiently from the usual conduct. However, if the assault is similar to the common behaviour, it may not be detected. Moreover, it is difficult to associate deviations with unique attacks. As the customers exchange their behavior, natural habits must be redefined..

DOI 10.17148/IJARCCE

Misuse detection

Misuse detection programs use a priori talents on assaults to appear for attack traces. In other words, they detect intrusions by using knowing what the misuse is. Signature (rule) situated methods are probably the most usual examples of the misuse detection methods. In signature founded detection, assault signatures are sought within the monitored useful resource.

Clustering

Data clustering is primarily a description that is employed as a standard technique for data analysis in varied fields like machine learning, data processing, pattern reorganization, image analysis and bio-informatics. Cluster analysis is additionally recognized as a very important technique for classifying information, finding clusters of a dataset supported similarities within the same cluster and dissimilarities between completely different clusters . putt every purpose of the dataset to exactly one cluster is that the basic of the standard clustering technique wherever as clustering rule really partitions untagged set of knowledge into completely different teams in step with the similarity. As compare to information classification, information clustering thought is taken into account as unsupervised learning method that doesn't need any tagged information set as training data and also the performance of knowledge clustering rule is mostly considered the maximum amount poorer. Though information classification is healthier performance adjusted however it needs a tagged information set as training information and much classification of tagged data is mostly terribly troublesome additionally as valuable. There square measure several algorithms that square measure planned to enhance the clustering performance. clustering is largely thought-about as classification of comparable objects or in alternative words, it's exactly partitioning of information sets into clusters so data in every cluster shares some common attribute. The hierarchic, partitioning and mixture model ways square measure the 3 major varieties of clustering processes that square measure applied for organizing information. The selection of application of a specific technique typically depends on the kind of output desired, the known performance of the strategy with specific form of information, out there hardware and software facilities and size of the dataset.

II. EXISTING SYSTEM

There are too many clustering algorithms that widely in use. From that k means clustering is most popular because of the less time complexity [1]. Fuzzy c means [2] clustering is an another type clustering algorithm that have high detection rate and low false positive rate than k means and it allow an item to be belongs to more than one cluster. But the main drawback of these systems are they needs fixed number of cluster and before starting the process the user should specifies the number of cluster. It not at all a good approach for intrusion detection. So this paper is going to do works on the cluster that accept any number of clusters .So initially this paper deals with two algorithms

1. DBSCAN clustering
2. GB clustering

1. DBSCAN clustering

DBSCAN (Density Based Spatial Clustering of Application with Noise) is a density based clustering algorithm . Here regions with high density into clusters and discovers cluster of arbitrary shape in spatial databases with noise. It defines as maximal set of density-connected points. Density means number of points within a specified radius called as Eps. A point is a core point if it has more than specified number of points called as (Min Pts) within Eps.

The advantage of this scheme is it does not require a-priori specification of number of clusters and it is able to identify noise data while clustering. DBSCAN algorithm is able to find arbitrarily size and arbitrarily shaped clusters. But DBSCAN algorithm fails in case of varying density clusters. Also fails in case of neck type of dataset. It does not work well with the high dimensional data So I am chosen graph based clustering technique to my approach.

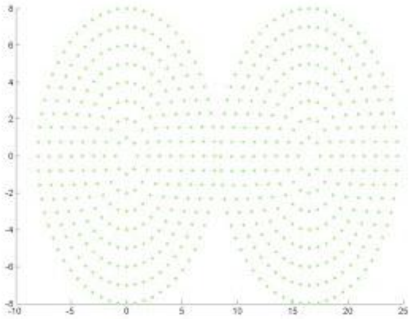


figure 1

2.GRAPH-BASE CLUSTER ALGORIGHM

First, Graph-based clustering algorithm [3] 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 is computed by a parameter of cluster precision .

$$\alpha = dismin + (dismax - dismin) \times ClusterPrecision \qquad □ □ □$$

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.

Here 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.

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 as shown in figure 2. Each sub-graph represents a cluster. Finally, outliers are processing.

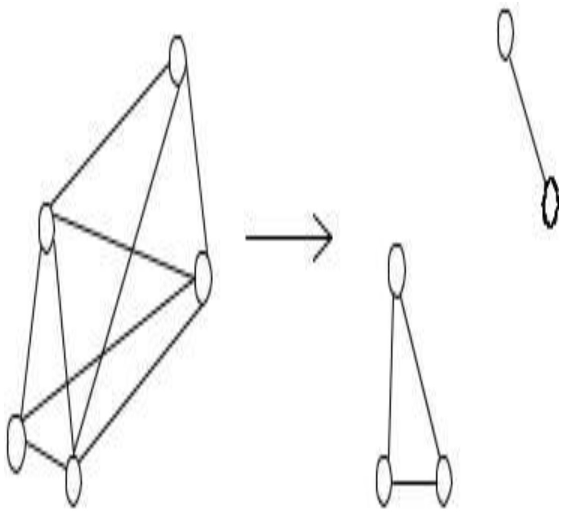


Figure 2

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.PROPOSED SYSTEM

In a graphical representation, each record has a hub vertex representing the record itself, along with a vertex for each attribute, each of which is connected to the hub. We can then represent the entire table by combining all the records into a single graph. To find anomalous database records, we would be interested in determining how unusual each of the separate star configurations was.

Application of anomalous cluster detection to attribute-value databases is straightforward, but many other classes of databases can benefit from this technique; it can be applied to any graph in which the vertices can be grouped in a meaningful way. An example is web click stream data. This type of data has a natural graph-based representation: vertices correspond to web pages, and directed edges correspond to navigational links selected by the user. Furthermore, the vertices can be grouped into clusters in a meaningful way – organized by user. Cluster A would contain all the click stream data from user A, cluster B would contain data from user B, and so forth. By performing anomalous cluster detection on the overall graph, we would then be testing each user for unusual web-navigation patterns.

Here I prefer the method for anomalous cluster detection using Subdue. First, some background is necessary. Subdue can be set to run multiple iterations on a single graph. After each iteration, the graph is compressed using the discovered substructure; in other words, every instance of the substructure is replaced by a single vertex. The next iteration of Subdue will then operate on the newly compressed graph. This multiple-iteration capability is used in my approach. It is important to realize that the best substructures will be discovered in the first several iterations, while later substructures will become less and less valuable (i.e., less common). Here I am assuming that Subdue halts once the graph contains no substructure with more than one instance.

The rationale for my method lies in the idea that clusters containing many common substructures are generally less anomalous than clusters with few common substructures. This is related to the underlying idea behind anomalous substructure detection – that common substructures are, in a loose sense, the opposite of anomalous substructures. On each iteration, Subdue discovers the best substructure (in the MDL sense), and then compresses the graph with it. It stands to reason, then, that anomalous clusters tend to experience less compression than other clusters, since they contain few common patterns.

The method can be described as by using usual clustering algorithm it can be easy to make clusters and in these clustering I am going to do outlier processing.

It can be described as,

```
1 For all i in RECORDSET DO{

2 Put note i into GRAPH
3 Repeat {
4 Calculate threshold by function
$α=dismin+(dismax-dismin)×ClusterPrecision$
 5 cut down all the edges whose value is greater than the threshold
6    transverse GRAPH, label all the sub-graphs.
7    outlier processing
          } until the outlier is processed completely
```

Outlier processing is done by using the concept of subdue system. , Subdue is an algorithm for detecting repetitive patterns (substructures) within graphs. For the purposes of this paper, a graph consists of a set of vertices and a set of edges, which may be directed or undirected. Furthermore, each vertex and edge contains a label to identify its type, which need not be unique. A substructure is a connected sub graph of the overall graph. Subdue keeps an ordered list of discovered substructures called the parent list; at the beginning, this list simply holds 1-vertex substructures for each unique vertex label. Subdue repeatedly removes all the substructures from the parent list, generates and evaluates their extensions, and inserts the extensions onto the list. An extension of a substructure is generated by adding either a new vertex (and its corresponding edge), or just a single edge within the substructure. As new substructures are being generated, a second list is maintained holding the best substructures discovered so far. When this process is finished, the substructure with the top value is reported, and possibly used to compress the graph before the next iteration begins. Compressing the graph refers to replacing each instance of the substructure with a new vertex representing that substructure. Each substructure is evaluated using the Minimum Description Length heuristic. The minimum description length (or simply description length) is the lowest number of bits needed to encode a piece of data; Subdue contains an algorithm that will approximate this value for any given graph. Using this heuristic, the best substructure to be the one that minimizes the following value:

$$F1\ (S\ ,G)\ =DL\ (G|S)+DL(S)$$

where G is the entire graph, S is the substructure, DL(G|S) is the description length of G after compressing it using S, and DL(S) is the description length of the substructure.

The compression technique is done by, for each cluster, assign a value A; the higher A is, the more anomalous the cluster. A is given by the formula

$$A = 1 - \frac{1}{n} \sum_{i=1}^{N} (n - i + 1) * c_i$$

Where n is the no of iterations and ci is the percentage of the cluster that is compressed away on the i^{th} iteration. ci is more rigorously it is defined as,

$$\frac{{}^{DL}i-1\ \ (G)\ -DL_i\,(G)}{DL_0\,(G)},$$

Where D(L|G) is the description length of the cluster after j iterations. Some explanation of the above formula for A is in order. The idea is that all clusters begin with an A-value of 1 (i.e., completely anomalous), and the values drop off as portions of the clusters are compressed away during the iterations. The ci term will vary from 0 to 1; a value of 0 means that the cluster was not changed on the i th iteration, while a value of 1 means that the entire cluster was compressed away. The (n – i + 1) term will vary from n to 1 as i increases; this causes A to drop off more sharply for compressions that occur early on. The (1/n) term guarantees that the final value will be between 0 and 1, since the maximum possible value for the summation is n.

By performing the all process it easily to find the anomalous cluster with good efficiency.

And general description for the described method is consider the example graph figure 3 , suppose that the three triangles represent three separate clusters. (This is acceptable, even though edges connect the clusters; the edges running between are not considered to be part of any cluster.)

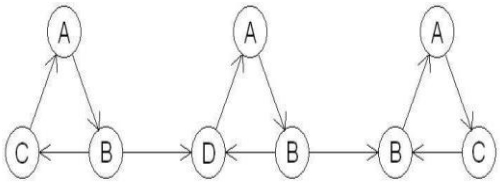
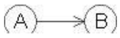


Figure 3

Recall that after one iteration, the substructure  is ranked highest, and will be used to compress the entire graph. If this is the only iteration under consideration, then we would consider the third cluster to be the most anomalous; it will not be compressed at all, whereas two of the three vertices will be compressed away in each of the first two clusters.

This method can be test using the 1999 KDD Cup network intrusion dataset [4]. The data consists of connection records, each of which is labelled as normal or as one of 37 different attack types. Each record contains 41 features describing the connection (duration, protocol type, number of data bytes, etc.); some of these features are continuous, others discrete. In the original competition, the dataset was split into two sections: the training data and the test data. Participants were able to train their detectors with the training data, and were then judged based on their performance on the test data. It is very useful method for network based intrusion detection. Since in network the intrusion is of different size so it work well there.

IV.CONCLUSION

Intrusion detection process founded on data mining increases the intelligence and reliability of network. Most likely, by means of clustering procedure, intrusion detection could also be carried out. This algorithm awarded in this paper could overcome some hazards of the normal cluster algorithm for intrusion detection and can obtain comparative sufficient efficiency of intrusion detection. However, there are nonetheless have many deficiencies needed to be expanded. A different disadvantage will have to be elevated is that the initial percentage of irregular and average files want guide control to search out the clusters, it suspicious roughly influences efficiency of this algorithm and the complexity is also higher in this method.

REFERENCES

1. Rakhlin and A. Caponnetto, “Stability of K-Means clustering”, Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2007, pp. 216–222.
2. A. Rui and J. M. C. Sousa, “Comparison of fuzzy clustering algorithms for Classification”, International Symposium on Evolving Fuzzy Systems, 2006 , pp. 112-117.
3. Yang Liu. GB-Cluster: a Graph-based Clustering Algorithm. Compu-ter science,2002
4. KDD.KDDCup1999Data.http://kdd.ics.uci.edu/databases/kddcup99/kdd-cup99.html, 1999