CLUSTERING BASED OUTLIER DETECTION METHOD FOR NETWORK BASED INTRUSION DETECTION

Deevi Radha Rani

ABSTRACT

The discovery of objects with exceptional behavior is an outstanding challenge from a knowledge discovery standpoint and has received considerable attention in many applications such as network attacks, fraud detection. This paper proposes a simple clustering based algorithm to detect outlying objects. The main problem for network intrusion detection system is the ability to exploit ambiguities in the traffic stream. Network-Based Intrusion Detection monitors network traffic for particular network segment and analyzes the network and application protocol activity to identify suspicious activity. There are several recently developed outlier detection schemes to detect attacks in a network. In this paper, the proposed algorithm is applied to network intrusion detection system to detect ambiguities or violations in the network traffic stream.

Keywords: Outlier Detection, Clustering, Network based Intrusion Detection

1. INTRODUCTION

Detecting outliers in data is a research field with variety of application domains, including fraud detection, criminal activities in e- commerce and detecting suspicious activities. Outlier detection approaches concentrate on detecting data objects that do not comply with the general behavior or model of the data and which are grossly different from or inconsistent with the remaining set of data.

Most of the outlier detection effort is on dataset with a small set of numerical or ordinal data but not on large dataset or either on categorical data. The proposed clustering based Outlier detection method is scalable to large set of data and Categorical data also where in some techniques are used to map categorical to numerical values.

Another issue is how to apply outlier detection scheme to Network Intrusion detection to detect ambiguities in the network traffic stream. Data mining based intrusion detection techniques generally fall into one of two categories; namely misuse detection and anomaly detection. In misuse detection approaches, each instance in a data set is labeled as normal or intrusion (attack) and a learning algorithm is trained over the labeled data. These approaches are able to automatically retrain intrusion detection models on different input data that

include new types of attacks as long as they have been labeled appropriately. The main advantage of misuse detection is that it can accurately detect known attacks, while its drawback is its inability to detect novel, previously unseen attacks.

Traditional anomaly detection approaches, on the other hand, build models of normal data and detect deviations from the normal model in observed data. Anomaly detection applied to intrusion detection and computer security has been an active area of research. Anomaly detection algorithms have the advantage that they can detect new types of intrusions as deviations from normal usage.

The proposed method is like signature based detection where in time and frequency of network usage of a user is considered as “normal” and time and frequency of network usage of intruder is considered as “intrusion” or “attack”. The intruder detection capability used is the Threshold value of frequency of network usage. In this paper, a test dataset is considered to apply the proposed algorithm to find intruders.

2. PREVIOUS WORK

There is no single universally applicable or generic outlier detection approach. Therefore, many approaches have been proposed to detect outliers. These approaches can be classified in [1] into four major categories based on the techniques used, which are: distribution-based, distance-based, density-based and deviation-based approaches.

Distribution-based approaches [7] develop statistical models (typically for the normal behavior) from the given data and then apply a statistical test to determine if an object belongs to this model or not. Objects that have low probability to belong to the statistical model are declared as outliers. However, Distribution-based approaches cannot be applied in multidimensional scenarios because they are univariate in nature. In addition, a prior knowledge of the data distribution is required, making the distribution-based approaches difficult to be used in practical applications.

In the **distance-based** approach [6], outliers are detected as follows. Given a distance measure on a feature space, a point q in a data set is an outlier with respect to the parameters M and d, if there are less than M points within the distance d from q, where the values of M and d are decided by the user. The problem with this approach is that it is difficult to determine the values of M and d.

Density-based approaches [5] compute the density of regions in the data and declare the objects in low dense regions as outliers. Outlier score is assigned to any given data point, known as Local Outlier Factor (LOF), depending on its distance from its local neighborhood.

Deviation-based approaches [1] identify outliers by examining main characteristics of objects in a group. Object that deviate from this description are considered outliers. The first approach is sequential exception technique which sequentially compares objects in a set, while the second employs an OLAP data cube approach.

Clustering-based approaches, consider clusters of small sizes as clustered outliers. In these approaches, small clusters (i.e., clusters containing significantly less points than other clusters) are considered outliers. The advantage of the clustering-based approaches is that they do not have to be supervised. Moreover, clustering-based techniques are capable of being used in an incremental mode (i.e., after learning the clusters, new points can be inserted into the system and tested for outliers).

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices. Incidents have many causes, such as malware (e.g., worms, spyware), attackers gaining unauthorized access to systems from the Internet, and authorized users of systems who misuse their privileges or attempt to gain additional privileges for which they are not authorized. Although many incidents are malicious in nature, many others are not; for example, a person might mistype the address of a computer and accidentally attempt to connect to a different system without authorization. An intrusion detection system (IDS) is software that automates the intrusion detection process.

Most Intrusion Detection System technologies use multiple detection methodologies, either separately or integrated, to provide more broad and accurate detection. The methodologies [4] are like Signature-Based Detection, Anomaly-Based Detection, Stateful Protocol Analysis. A signature is a pattern that corresponds to a known threat. **Signature-based detection** is the process of comparing signatures against observed events to identify possible incidents. **Anomaly-based detection** is the process of comparing definitions of what activity is considered normal against observed events to identify significant deviations. **Stateful protocol analysis** is the process of comparing predetermined profiles of

generally accepted definitions of benign protocol activity for each protocol state against observed events to identify deviations.

There are many types of Intrusion Detection System technologies [4]. They are divided into the following four groups based on the type of events that they monitor. **Network-Based**, which monitors network traffic for particular network segments or devices and analyzes the network and application protocol activity to identify suspicious activity. **Wireless**, which monitors wireless network traffic and analyzes it to identify suspicious activity involving the wireless networking protocols themselves. **Network Behavior Analysis** (NBA), which examines network traffic to identify threats that generate unusual traffic flows, such as DDoS attacks, scanning, and certain forms of malware. Host-Based, which monitors the characteristics of a single host and the events occurring within that host for suspicious activity.

3. PROPOSED METHOD

In this paper a simple clustering based approach is proposed to detect outliers. In this approach a set of properties of a normal data is taken as input and produces all data points that do not match with set of properties and these are called as outliers. The algorithm has 2 phases. In I Phase, the set of data points are considered as a Normal Cluster. Every data point in a cluster is compared with

Input:
D, Set of Data Points
N $\{n_1, n_2, \ldots n_i\}$, Set of properties of Normal Data
Output:
2 Clusters, Deviation Cluster O, Normal Cluster C
Outliers

Algorithm

Begin
Initially, suppose D as a Cluster C
Repeat
Select a Property n_i from N
Select a point p_i from C which has not been visited
Get the Property of p_i and compare with n_i
if Property matches
Keep it in C
Otherwise
Remove from C & Place it in O
Until All Properties are matched
if p_i in O do not match with t, Threshold Property in N then it is an outlier
End

Fig. 1: Proposed Algorithm

each property of a normal data. If a data point does not match with a property it is considered as Deviation. So, there form 2 clusters Normal Cluster and Deviation Cluster. In II Phase, Deviation cluster is considered and every data point in the cluster is processed. If a data point in deviation cluster does not match with particular threshold property it is considered as an outlier.

A threshold is a value that sets the limit between normal and abnormal behavior. Thresholds usually specify a maximum acceptable level, such as x failed connection attempts in 60 seconds, or x characters for a filename length.

The basic structure of the proposed method is given in figure 1.

This algorithm can be applied to a Network Intrusion Detection System to find intruders of the system. The properties of the normal users and intruders differ from each other in one or the other. The above proposed idea can be implemented to a Network to find violations or ambiguities in the network.

4. RESULTS AND DISCUSSION

In this section a test dataset is considered to investigate the performance of the proposed approach which finds the possible intruders based on frequency of network usage. The test dataset considered is the frequencies of each user of a particular network at a given moment. Here the dataset consist of 10 frequencies of 10 users at a moment.

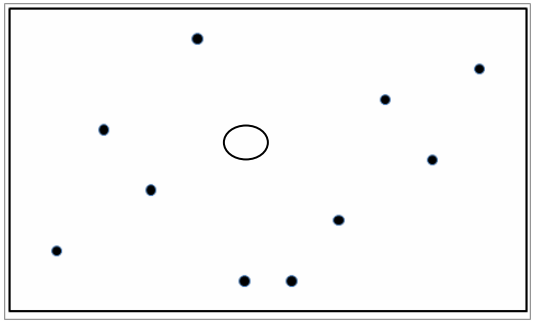

Fig 2: Data Set

The point with is circled is considered as threshold property of a user. When the above proposed algorithm is applied, 2 clusters will be formed as Deviation Cluster (intuders) and Normal Cluster (authorized users). When I Phase of Algorithm is applied the following result is observed.

The above cluster which is filled is a deviation cluster, where the frequency of the data points in this cluster will not match with the nomal cluster property.

The data point with red color is the frequency of user2 to access the network. The frequency will not match with the normal cluster property. So initially it is treated as ambiquity or intruder. When II Phase of algorithm is applied the following result is observed.

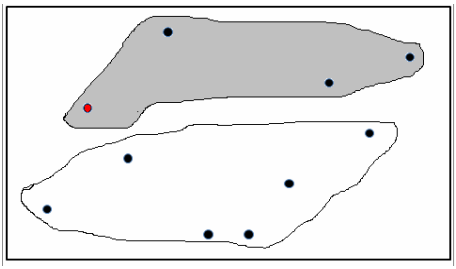

Fig 3: After applying I Phase Algorithm

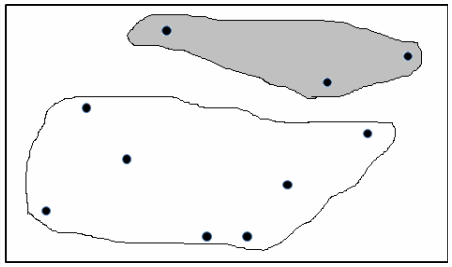

Fig 4: Result: After II Phase Algorithm

In the II phase of the algorithm the data points in the deviation cluster is processed. If any point nearly matches with threshold property it is not considered as outlier. So, here due to some reasons user2 frequency changes from normal frequency but user2 is authorized to use the network. So, in the II phase of algorithm user2 is not considered as violation or intruder. Only user4, user8 and user10 are intruders.

5. CONCLUSION

In this paper, a simple clustering based algorithm for outlier detection is proposed. First this algorithm is performed to find deviation cluster and normal cluster by comparing the properties of dataset and normal property. Finally the deviation cluster is compared with threshold property and outliers are found. This algorithm is applied to network based intrusion detection system to find intruders or ambiguities or violations in the network. The test results show that the proposed approach gave effective results when applied to different data sets.

REFERENCES:

1. [Han, J. and M. Kamber, 2006, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2nd Ed.
2. Moh'd Belal Al- Zoubi, "An Effective Clustering-Based Approach for Outlier Detection", European Journal of Scientific Research, ISSN 1450-216X Vol.28 No.2 (2009), pp.310-316
3. Anna Koufakou, Jimmy SEcretan, John Reeder, Kelvin Cardona, and Michael Georgiopoulos, "Fast Parallel Outlier Detection for Categorical Datasets using MapReduce", 2008 International Joint Conference on Neural Networks.

4. Karen Scarfone, Peter Mell, "Guide to Intrusion Detection and Prevention Systems (IDPS)", Recommendations of the National Institute of Standards and Technology, February2007.
5. Breunig, M. M., Kriegel, H.P., Ng, R. T., and Sander, J., "LOF: Identifying density-based Local Outliers", Proc. Of the ACM SIGMOD International Conference on Management of Data, 2007
6. Knorr, E., Ng, R, and Tucakov, V., "Distance-based Outliers: Algorithms and Applications", Very Large Databases VLDB Journal, 2000
7. Barnett, V., Lewis, T. "Outliers in Statistical Data", John Wiley, 1994. Opim Salim Sitompul and Shahrul Azman Noah, A Transformation-oriented Methodology to Knowledge-based Conceptual Data Warehouse Design ,Journal of Computer Science 2 (5): 460-465, 2006.