Autonomous Network Security For Detection Of Network Attacks using Cluster

Ms.Priti K.Doad
Department of CSE,
G.H.Raisoni College, Amravati. 
doad.priti@gmail.com

Mr.Mahim M.Bartere
Department of CSE,
G.H.Raisoni College, Amravati. 
mahip.bartere@raisoni.net

Abstract

Retrieval of information from the databases is now a day’s significant issues. The thrust of information for decision making is challenging one. To overcome this problem, different techniques have been developed for this purpose. One of techniques is clustering. Clustering is a significant task in data analysis and data mining applications. It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). The clustering is unsupervised learning. In this paper we propose a methodology for comparing clustering methods based on the quality of the result and the performance of the execution. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. Clustering has been widely used as a segmentation approach therefore, choosing an appropriate clustering method is very critical to achieve better results. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. There are different types of Clustering algorithms partition-based algorithms such as K-Means, KNN, density-based algorithms and SSC-EA-based algorithms. Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms.

Keywords:- Data Mining, Density Based, Partition Based clustering, Kalman Filter, UNADA.

1. Introduction

The unsupervised detection of network attacks represents an extremely challenging goal. The detection of network attacks is a paramount task for network operators in today’s Internet. Denial of Service attacks (DoS), Distributed DoS (DDoS), network/host scans, and spreading worms or viruses are examples of the different attacks that daily threaten the integrity and normal operation of the network. The principal challenge in automatically detecting and analysing network attacks is that these are a moving and ever-growing target. Unsupervised learning is a recent approach in knowledge exploration. It is widely used on/with unlabelled data, such as extracting relevance that exists in records. Unsupervised learning is an important supplementary method to category data since it could increase the precision of clustering results.

Unlike supervised learning, unsupervised learning attempts to find the most reasonable patterns by uncovering relationships best instead of using preferential classification labels. Because the idea behind unsupervised learning is to run an unsupervised algorithm on raw data, we consider the applications of data clustering and data reduction (including dimension reduction, size reduction, etc.) as two key issues in the framework of knowledge exploration. The use of an unsupervised learning method could save time in data processing by removing the matching and ranking process used for specified classes. While the majority of our work focused on identifying anomalies in aggregate traffic at the campus border router, the source and destination address in the IP flow data allows us to isolate anomalies at different points in the network. As you move closer to the source of an anomaly, the event typically becomes more pronounced in the data and thus easier to expose. However, if the event takes place at a point in the network where there is lower aggregation of traffic then there is typically more variability in the ambient traffic and, as a result, the task of isolating the anomaly signal becomes more difficult. We show that our methods work well whether the measurement point is close to or distant from the point of the anomaly.

Two different approaches are by far dominant in the literature and commercial security devices: signature-based detection and anomaly detection. Signature-based detection systems are highly effective to detect those attacks which they are programmed to alert on. However, they cannot defend the network against unknown attacks. Even more, building new signatures is expensive and time-consuming, as it involves manual inspection by human experts. Anomaly detection uses labelled data to build normal-operation-traffic profiles, detecting anomalies as activities that deviate from this baseline. Such methods can detect new kinds of network attacks not seen before. Signature engines also have their disadvantages. Because they only detect known attacks, a signature must be created...
for every attack, and novel attacks cannot be detected. Signature engines are also prone to false positives since they are commonly based on regular expressions and string matching. Both of these mechanisms merely look for strings within packets transmitting over the wire. A disadvantage of anomaly-detection engines is the difficulty of defining rules. Each protocol being analyzed must be defined, implemented and tested for accuracy. The rule development process is also compounded by differences in vendor implementations of the various protocols. Custom protocols traversing the network cannot be analyzed without great effort. Moreover detailed knowledge of normal network behaviour must be constructed and transferred into the engine memory for detection to occur correctly. On the other hand, once a protocol has been built and behaviour defined, the engine can scale more quickly and easily than the signature-based model because a new signature does not have to be created for every attack and potential variant. [1] [2]

Apart from detection, operators need to analyse and characterize network anomalies to take accurate countermeasures. The characterization of an anomaly is a hard and time-consuming task. The analysis may become a particular bottleneck when new anomalies are detected, because the network operator has to manually dig into many traffic descriptors to understand its nature. Even expert operators can be quickly overwhelmed if further information is not provided to prioritize the time spent in the analysis. Contrary to current supervised approaches, we develop in this work a completely unsupervised method to detect and characterize network anomalies, without relying on signatures, training, or labelled traffic of any kind. The proposed approach permits to detect both well-known as well as completely unknown anomalies, and to automatically produce easy-to-interpret signatures that characterize them.

Unsupervised Detection of Network Attacks specifies that the objective of clustering is to partition a set of unlabelled elements into homogeneous groups of “similar” characteristics, based on some similarity measure. Different from other techniques for unsupervised data analysis (e.g. density estimation, dimensionality reduction, etc.), clustering permits to work with multiple-classes problems without modifying the characteristics of the analysed traffic, hence it represents an attractive means for unsupervised detection of attacks.

Our approach relies on robust clustering algorithms to detect both well-known as well as completely unknown attacks, and to automatically produce easy-to-interpret signatures to characterize them, both in an on-line basis. The analysis is performed on packet-level traffic, captured in consecutive time slots of fixed length ΔT and aggregated in IP flows (standard 5-tuples). IP flows are additionally aggregated at 9 different flow levels l. These include: source IPs, destination IPs, source Network Prefixes, destination Network Prefixes, and traffic per Time Slot. The complete detection and characterization algorithm runs in three successive stages. The first step consists in detecting an anomalous time slot where an attack might be hidden. The unsupervised detection and characterization algorithm begins in the second stage, using as input the set of IP flows captured in the flagged time slot. In the third stage, the evidence of traffic structure provided by the clustering algorithms is used to produce filtering rules that characterize the detected attack and simplify its analysis. The characterization of an attack can be a hard and Time-consuming task, particularly when dealing with unknown attacks. Even expert operators can be quickly overwhelmed if simple and easy-to-intercept information is not provided to prioritize the time spent in the analysis. To alleviate this issue, the most relevant filtering rules are combined into a new traffic signature that characterizes the attack in simple terms. This signature can ultimately be integrated to any standard security device to detect the attack in the future, which constitutes a major step towards autonomous security: in a nutshell, our algorithm automatically produces new signatures without any previous data about traffic or knowledge about the attack. [3] [4]. Current network monitoring systems rely strongly on signature-based and supervised-learning-based detection methods to hunt out network attacks and anomalies. Despite being opposite in nature, both approaches share a common downside: they require the knowledge provided by an expert system, either in terms of anomaly signatures, or as normal-operation profiles. In a diametrically opposite perspective we introduce UNADA, an Unsupervised Network Anomaly Detection Algorithm for knowledge-independent detection of anomalous traffic. It uses a novel clustering technique based on Sub-Space-Density clustering to identify clusters and outliers in multiple low-dimensional spaces. The evidence of traffic structure provided by these multiple clustering’s then combined to produce an abnormality ranking of traffic flows, using a correlation-distance-based approach. We evaluate the ability of UNADA to discover network attacks in real traffic without relying on signatures, learning, or labelled traffic. Additionally, we compare its performance against previous unsupervised detection methods using traffic from two different networks.

1.1. Motivation

Our motivation of this work is to build an efficient and effective robust clustering based algorithm for detection of network attack by allowing training with unlabelled data. Its efficiency and effectiveness will be the higher detection rate and the lower false detection comparing to the existing approaches of unsupervised detection of network attacks.

The two knowledge-based approaches are not sufficient to tackle the anomaly detection problem, and that a holistic solution should also include knowledge-independent analysis techniques. There are some algorithms, and it becomes critical in the case of
unsupervised detection, because there is no additional information to select the most relevant set some approaches can be easily extended to detect other types of attacks, considering different sets of traffic features. In fact, more features can be added to any standard list to improve detection and characterization results. To this aim we propose UNADA, an Unsupervised Network Anomaly Detection Algorithm that detects network traffic anomalies without relying on signatures, training, or labelled traffic of any kind. Based on the observation that network traffic anomalies are, by definition, sparse events that deviate markedly from the majority of the traffic, UNADA relies on robust clustering algorithms to detect outlying traffic flows.

1.2. Objective
The objective of Knowledge Independent Detection of Network Attack is simply to detect the attacks which are completely unknown to us. There is no previous knowledge about that data. There are some algorithms in existence which are used for network security but they are inefficient as they are knowledge based (Signature Based and Anomaly Based) whenever there is a vast amount of continuous incoming data then it is a big risk regarding the network attacks which attacks with the help of Robust Clustering Algorithm and make whole data secure.

2. Literature Review
The problem of network anomaly detection has been extensively studied during the last decade. Most of the approaches analyse statistical variations of traffic volume descriptors (e.g., number of packets, bytes, or new flows) and particular traffic features (e.g., distribution of IP addresses and ports), using either single-link measurements or network-wide data. A non-exhaustive list of standard methods includes the use of signal processing techniques on single-link traffic measurements, Kalman filters for network-wide anomaly detection, and Sketches applied to IP-flows. Our approach falls within the unsupervised anomaly detection domain. The vast majority of the unsupervised detection schemes proposed in the literature are based on clustering. It reports improved results in the same data-set, using three different clustering algorithms: Fixed-Width clustering, an optimized version of k-NN. It presents a combined density-grid-based clustering algorithm to improve computational complexity, obtaining similar detection results. PCA (parallel clustering algorithm) and the sub-space approach is another well-known unsupervised anomaly detection technique, used in to detect network-wide traffic anomalies in highly aggregated traffic flows. UNADA (Unsupervised Network Anomaly Detection Algorithm) presents several advantages w.r.t. current state of the art. [5] [6]

2.1. Existing Approaches for the Detection of Network Attacks
Applying unsupervised anomaly detection in network intrusion detection is a new research area that has already drawn interest in the academic community. Eskin, et al. (Eskin et al. 2002) investigated the effectiveness of three algorithms in intrusion detection: The xed-width clustering algorithm, an optimised version of the k-nearest neighbour algorithm, and the one class support vector machine algorithm. He carried out further research based on the clustering method and showed improvements in accuracy when the clusters are adaptive to changing track patterns. [7]

Different approach using a quarter-square support vector machine is proposed in (Laskov, Schafer & Kotenko 2004), with moderate success. In (Eskin 2000), a mixture model for explaining the presence of anomalies is presented, and machine learning techniques are used to estimate the probability distributions of the mixture to detect anomalies. In (Zanero & Savaresi 2004), a novel two-tier IDS is proposed. The rst tier uses unsupervised clustering to classify the packets and compresses the information within the payload, and the second tier used an anomaly detection algorithm and the information from the rst tier for intrusion detection. Lane and Brodley (Lane & Brodley 1997) evaluated unlabelled data by looking at user profiles and comparing the activity during an intrusion to the activity during normal use. Supervised anomaly detection in network intrusion detection, which uses purely normal instances as training data, has been studied extensively in the academic community. A comprehensive survey of various techniques is given. An approach for modelling normal trc using self-organising maps is presented in (Gonzalez & Dasgupta 2002), while another one uses principal component classiers to obtain the model (Shyu, Chen, Sarinnapakorn & Chang 2003). One approach uses graphs for modelling the normal data and detects the irregularities in the graph for anomalies. Another approach uses the normal data to generate abnormal data and uses it as input for a classification algorithm. [8][9]

3. Proposed System Planning and Design
The proposed system will be identified to provide a solution to the problem of anomaly detection which is completely Knowledge Independent. In the Knowledge Independent Unsupervised Detection of Network Attack. We evaluate the ability of UNADA to discover network attacks in real traffic without relying on signatures, learning, or labelled traffic. Additionally, we compare its performance against previous unsupervised detection methods using traffic from two different networks.
3.1. System Design

In the system design input data at first that contain the data packets. A data set is an ordered sequence of object, this may contain anomaly and we have to detect anomalies in the data set to detect that anomalies in the huge dataset we have to apply robust clustering approach which will create automatic signature. In my proposed work I am going to implement completely blind approach so for that no any previous knowledge about the anomaly and to detect such types of blind attack. I am going to apply robust clustering approach for the detection of network anomaly in a completely unsupervised fashion.


The unsupervised detection stage takes as input all the IP flows in the anomalous time slot, aggregated according to one of the different aggregation levels used in the first stage. Let \( Y = \{y_1, \ldots, y_n\} \) be the set of \( n \) flows in the flagged time slot. Each flow \( y_i \in Y \) is described by a set of \( m \) traffic attributes or features on which the analysis is performed. The selection of these features is a key issue to any anomaly detection algorithm, and it becomes critical in the case of unsupervised detection, because there is no additional information to select the most relevant set. In this we shall limit our study to detect and characterize well-known attacks, using a set of standard traffic features widely used in the literature. However, the reader should note that the approach can be easily extended to detect other types of attacks, considering different sets of traffic features. In fact, more features can be added to any standard list to improve detection and characterization results. The set that we shall use here includes the following \( m = 9 \) traffic features: number of source/destination IP addresses and ports, ratio of number of sources to number of destinations, packet rate, ratio of packets to number of destinations, and fraction of ICMP and SYN packets. According to previous work on signature-based anomaly characterization, such simple traffic descriptors permit
to describe standard network attacks such as DoS, DDoS, scans, and spreading worms/virus. The algorithm is based on clustering techniques applied to data set. The objective of clustering is to partition a set of unlabeled elements into homogeneous groups of similar characteristics, based on some measure of similarity. Our goal is to identify in the different aggregated flows that may compose the attack. For doing so, the reader should not necessary to have that an attack may consist of either outliers (i.e., single isolated flows) or compact small size clusters, depending on the aggregation level of flows in Y.

4. System Implementation

4.1. Implementation Steps

Step 1:- To capture the packet of data which takes as input all the IP flows in flagged time slot by using analyzer i.e. Create Log file.

Step 2:- IP flows are additionally aggregated at different flow-resolution levels using different aggregation keys and apply sliding time windowing scheme for every 1 sec.

Step 3:- Create the feature space matrix by using following formula

\[ x(1) = [sipadd \ dipadd \ sport \ dport \ nsipadd/ndipadd \ y(1)/ndipadd] \]

Similarly, I have to create feature space matrices (i.e.clusters) for all time windows data set i.e., \( X=\sum(x_1,x_2,.....,x_n) \) and then apply Clustering algorithm and declare smallest group of cluster as outlier.

Step 4:- Detect anomalies using k-means clustering algorithm, evidence accumulation and outliers ranking.

Step 5:- Create a signature. Signature will be logged and updated in the signature table. Signature table can be used in for online detection anomalous flow.

Step 6:- To detect the attack in the future this signature can ultimately be integrated to any standard security device. There is filtering rules are combined into a new traffic signature that characterizes the attack in simple terms.

4.2. K-Means Algorithm

K-means algorithm as the underlying clustering algorithm to produce clustering ensembles. First, the data is split into a large number of compact and small clusters; different decompositions are obtained by random initializations of the K-means algorithm. The data organization present in the multiple clustering is mapped into a co-association matrix which provides a measure of similarity between patterns. The final data partition is obtained by clustering this new similarity matrix.

The main steps of K-means algorithm are as follows:-

1. Select an initial partition with N clusters; repeat steps 2 and 2 until cluster membership stabilizes.

2. Generate a new partition by assigning each pattern to its closest cluster center.

3. Compute new cluster centers.

4.3. Evidence accumulation and outliers ranking

The idea of evidence accumulation-based clustering is to combine the results of multiple clustering into a single data partition, by viewing each clustering result as an independent evidence of data organization. There are several possible ways to accumulate evidence in the context of unsupervised learning:

1. Combine results of different clustering algorithms.
2. Produce different results by resampling the data.
3. Running a given algorithm many times with different parameters or initializations.

5. Advantages the Proposed System on the Existing Approaches

1. Our unsupervised algorithm has several terms w. r. t. the state of the art:

2. First and most important, it works in a completely unsupervised fashion, which means that it can be directly plugged-in to any monitoring system and start to work from scratch, without any kind of calibration or previous knowledge.

3. It combines robust clustering techniques to avoid general clustering problems such as sensitivity to initialization, specification of number of clusters, or structure-masking by irrelevant features.

4. It automatically builds compact and easy-to-interpret signatures to characterize attacks, which can be directly integrated into any traditional security device.
5. It is designed to work on-line, using the parallel structure of the proposed clustering approach.

6. Application

In any organization such as bank, IT sector, software companies. There is transmission of lot of secure data. Each data is critical and important so it is necessary each data was be access only authorized person and if an unauthorized person try to access those data, then it can be detect by using unsupervised detection.

7. Conclusion

The Unsupervised Network Anomaly Detection Algorithm that we have proposed presents many interesting advantages w.r.t previous proposals in the field unsupervised anomaly detection. It uses exclusively unlabelled data to detect traffic anomalies, without assuming any particular model or any canonical data distribution, and without using signatures of anomalies or training. Despite using ordinary clustering techniques to identify anomalies, UNADA uses robust clustering technique, an Unsupervised Network Anomaly Detection Algorithm have the lack of robustness of general clustering approaches, by combining the notions of Sub-Space Clustering, Density-based Clustering, and multiple Evidence Accumulation.

We have verified the effectiveness of UNADA to detect real single source-destination and distributed network attacks in real traffic traces from different networks, all in a completely blind fashion, without assuming any particular traffic model, clustering parameters, or even clusters structure beyond a basic definition of what an anomaly is. Additionally, we have shown detection results that outperform traditional approaches for outlier’s detection, providing a stronger evidence of the accuracy of UNADA to detect network anomalies.

8. References