International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Review of Protection against Unauthorized Access Using Sub-Space Outliers Ranking

Rushikesh V. Mahalle[1], Parnal P. Pawade[2]

[1]ME (CSE) Scholar, Department of CSE, P R Patil College of Engineering & Technology, Amravati-444604, India
[2]Assistant Professor, Department of CSE, P R Patil Engineering & Technology, Amravati-444604, India

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. 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, UNADA

1. Introduction

Anomaly detection is recently a vital and active research problem in many fields and involved in numerous applications. Most of the existing methods are based on distance measure. Because of dynamic nature of the incoming data; declare an outlier often can lead us to a wrong decision. However, earlier research for the problem of outlier detection is suitable for disk resident datasets where the entire dataset is available in advance and algorithms can operate in more than single passes. But, outlier detection over data set is a challenging task because data is continuously updated and flowing.
Finding outliers in a collection of patterns is a very well-known problem in the data mining field. An outlier is a pattern which is dissimilar with respect to the rest of the patterns in the dataset. Depending upon the application domain, outliers are of particular interest. In some cases presence of outliers are adversely affect the conclusions drawn out of the analysis and hence need to be eliminated beforehand. There are varied reasons for outlier generation in the first place. For example outliers may be generated due to measurement impairments, rare normal events exhibiting entirely different characteristics, deliberate actions etc. Detecting outliers may lead to the discovery of truly unexpected behavior and help avoid wrong conclusions etc.

Outlier is the data point that does not conform to the normal points characterizing the data set. Detecting outliers has important applications in data cleaning as well as in the mining of abnormal points for fraud detection, stock market analysis, intrusion detection, marketing, network sensors. Finding anomalous points among the data points is the basic idea to find out an outlier. Distance based techniques use the distance function for relating each pair of objects of the data set. Distance based definition (these definitions are computationally efficient) [Knorr & Ng, 1998; Ramaswamy et al., 2000] represent useful tool for data analysis [Knorr & Ng, 1999].

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 unlabeled 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 are knowledge based. Our particular goal is to identify those attacks with the help of Robust Clustering Algorithm and make whole data secure.

2. Literature Survey

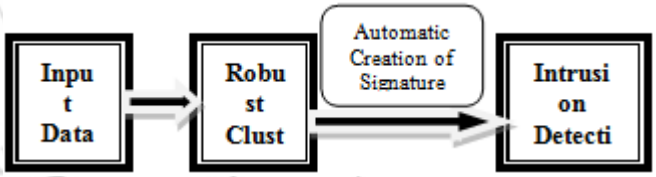
The problem of network anomaly detection has been extensively studied during the last decade. Most of the approaches analyse statistical variations of traffic volume descriptors (e.g., number of packets, bytes, or new flows) and particular traffic features (e.g., distribution of IP addresses and ports), 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. [4] [5]

2.1 Existing Approaches for the Detection of Network Attacks:

Two different approaches are by far dominant in current research literature and commercial detection systems: signature-based detection and anomaly detection. Signature-based detection is the de-facto approach used in standard security devices such as IDSs, IPSs, and firewalls. When an attack is discovered, generally after its occurrence during a diagnosis phase, the associated anomalous traffic pattern is coded as a signature by human experts, which is then used to detect a new occurrence of the same attack. Signature-based detection methods are highly effective to detect those attacks which they are programmed to alert on. However, they cannot defend the network against new attacks, simply because they cannot recognize what they do not know. In addition, building new signatures is a resources-consuming task, as it involves manual traffic inspection by human experts. On the other hand, 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. Nevertheless, anomaly detection requires training for profiling, which is time- consuming and depends on the availability of purely anomaly-free traffic data sets. Labelling traffic as anomaly-free is not only time consuming and expensive, but also prone to errors in the practice, since it is difficult to guarantee that no anomalies are buried inside the collected data. In addition, it is not easy to keep an accurate and up-to-date normal-operation profile. Our thesis is that these 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. To this aim paper 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.[6] [7]

Well in that their false-negative reports were negligible. These results indicate that traffic anomaly detection mechanisms based on deviation score techniques may be effective, however further development is necessary. [8]

3. Proposed System



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 an completely unsupervised fashion.

3.1 Architecture of the Unsupervised Detection of Attacks:

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 $\mathbf{Y} = \{\mathbf{y}, \ldots \mathbf{yn}\}$ be the set of n flows in the flagged time slot. Each flow $\mathbf{yi} \in \mathbf{Y}$ 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.

This approach falls within the unsupervised anomaly detection domain. The vast majority of the unsupervised detection schemes proposed in the literature are based on clustering and outliers detection [12]

4. Conclusions

The Unsupervised Network Anomaly Detection Algorithm that we have pro-posed presents many interesting advantages w. r. t. previous proposals in the field unsupervised anomaly detection. It uses exclusively unlabeled 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 avoids UNADA, 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.

References

[1] S. Hansman, R. Hunt "A Taxonomy of Network and Computer Attacks", in Computers and Security, vol. 24 (1), pp. 31-43, 2005.
[2] R. C´aceres, "Measurements of wide-area Internet traffic," Tech. Rep. UCB/CSD 89/550, Computer Science Department, University of California, Berkeley,1989.
[3] J. Brutlag, "Aberrant Behavior Detection in Time Series for Network Monitoring", in Proc. 14th Systems Administration Conference, 2000.
[4] K. Claffy, Internet Traffic Characterization, Ph.D. thesis, University of California, San Diego, 1994.
[5] L. Portnoy, E. Eskin, and S. Stolfo, "Intrusion Detection with Unlabeled Data Using Clustering", in Proc. ACM DMSA Workshop, 2001
[6] K. Claffy, Internet Traffic Characterization, Ph.D. thesis, University of California, San Diego,1998.
[7] R. Manajan, S. Bellovin, S. Floyd, V. Paxson, S. Shenker, and J. Ioannidis, "Controlling high bandwidth aggregates in the network," ACIRI Draft paper, February 2001.
[8] Paul Barford, Jeffery Kline, David Plonka and Amos Ron "A Signal Analysis of Network Traffic Anomalies" In procedeengs of ACM Sigcomminternet Measurment Workshop 2002
[9] A. Soule et al., "Combining Filtering and Statistical Methods for Anomaly Detection", in in Proc. ACM IMC, 2005
[10] Johan Mazel1,2, Pedro Casas1,2, and Philippe Owezarski1,2 1 CNRS; LAAS; 7 avenue du colonel Roche, F-31077 Toulouse Cedex 4, France 2 Universite de Toulouse; " Sub-Space Clustering and Evidence Accumulation for Unsupervised .Network Anomaly Detection" Springer-Verlag Berlin Heidelberg, 2011.
[11] Balachander Krishnamurthy, Subhabrata Sen, Yin, Zhang Yan Chen_ AT&T Labs– Research; "Sketch based Change Detection: Methods, Evaluation, and Applications"180 Park Avenue University of California IMC'03, October 27–29, 2003.
[12] Ana L.N. Fred Telecommunications Institute, Instituto Superior T´ecnico, Portugal and Anil K. Jain Dept. of Computer Science and Engineering Michigan State University, USA "Data Clustering Using Evidence Accumulation"2009