ISSN:0975-9646

Detection & Classification of Network Anomalies using SVM and Decision Tree

Mayank Nagar, Prof. Shraddha Pandit, Prof. JayPrakash Maurya

Department of Computer Science & Engineering, R.G.P.V.
Bhopal, India
IES College of Technology
Bhopal, India

Abstract: Here in this paper a new technique of detecting network anomalies in the traffic is implemented using the concept of Support vector machine and decision tree. The idea is to first apply clustering of the data traffic using support vector machine and then classifying the network traffic using vertical partition based id3 decision tree algorithm. The proposed technique implemented here provides high accuracy of detecting network anomalies as well as providing less time complexity and less error rate.

1. INTRODUCTION

Intrusion detection systems are software's used for recognizes the intended or unintended use of the system resources by unauthorized users. They can be classified into misuse detection systems and anomaly detection systems. Misuse detection systems representation attacks as a definite pattern and are more valuable in detecting known attack patterns. If the intrusion happens through learning, then the anomaly detection system may find out the intruder's behavior and for this reason may fail. Being more generalized and having a extensive possibility as compared to misuse detection systems, most of the present techniques focus on anomaly detection systems. Data mining approaches can be applied for both anomaly and misuse detection. Clustering techniques can be used to form clusters of data samples equivalent to the usual exploit of the method. Clustering based techniques can identify new attacks as compared to the classification based techniques.

Anomaly-based signatures are typically things to come across for network traffic that turns from what is distinguish on an average. The key trouble with this methodology is to first define what normal is. A few systems have hard-coded explanations of normal and in this case they could be think about heuristic-based systems. A variety of systems are built to study ordinary but the challenge with these methods is to reduce the opportunity of inappropriately classifying abnormal behavior as usual. In addition, if the traffic pattern that is being studied is assumed to be normal then the system has to challenge with how to differentiate between acceptable deviations and those that are not acceptable or that represent attack-based traffic. The work in this area has been typically limited to the academy, even if there are a few viable products that maintain to use anomaly-based detection methods. A subgroup of this type of detection is the profile-based detection methods. These systems based on their aware on transforms in the way that users or systems are interacting on the network. Interesting facts can be learned from development data and it is feasible to detect enduring attacks based on these algorithms.

Anomaly detection [1] is about finding the normal practice patterns from the audit data, while misuse detection is on the subject of encoding and matching the intrusion patterns via the audit data. The fundamental subject of our approach is to be appropriate data mining programs to the widely collected audit data to calculate models that precisely confine the authentic activities (i.e. patterns) of intrusions and normal activities. This ordinary approach removes the require to physically analyze and encode intrusion patterns, as well as the estimation in choosing statistical measures for normal procedure profiles. Anomaly detection [10-15]. The objective of such analysis is to imprison fine grained patterns in traffic distributions that easy volume based metrics cannot recognized.

In the last two decades, a fine number of anomaly based intrusion detection approaches[3-5][9] have been expanded, but a lot of them are common in nature, and thus reasonably uncomplicated. Due to the lack of papers that talk about a variety of information of incremental anomaly detection, we present a analysis in a arranged manner along with current research problems and challenges in this field.

2. LITERATURE REVIEW OF EXISTING TECHNIQUES

So this section deals with already existing anomaly detection based on classifying frequent traffic patterns.

George Nychis et.al proposed[2]entropy-based analysis of traffic distributions for anomaly detection there has been little effort to expansively recognize the detection power of using entropy-based analysis of various traffic distributions in combination with each other. We think that two classes of allocations: flow-header features (IP addresses, ports, and flow-sizes), and behavioral characteristics (degree distributions measuring the number of distinct destination/source IPs each host communicates with).Rather unexpectedly, here they examine that the entropies of the deal with and port allotments are forcefully shows a relationship with each other and make accessible very similar detection capabilities. The behavioral and flow size distributions are fewer correlated and detect incidents that do not show up as anomalies in the port and address distributions. Further analysis using synthetically generated

anomalies also recommends that the port and address distributions have inadequate effectiveness in detecting scan and bandwidth flood anomalies.
Here they give two suggestions: First, they should look beyond simple port and address based traffic distributions for fine-grained anomaly detection. In exacting, they should think about distributions that complement each other in their detection ability. Second, to keep away from the biases arising from uni-directional auditing, it is sensible to use bi-directional flow generalizations for computing traffic distributions whenever feasible. Two interesting directions of future work are: (1) exploring information-theoretic measures for deciding traffic features [6], and (2) leveraging the observed correlations to complement traditional entropy-based anomaly detection.

Hao Zhang et al [8] proposed User Intention-Based Traffic Dependence Analysis for Anomaly Detection. They explain a novel approach that can be used for detecting anomalous traffic on a host. This scheme investigates direct and indirect dependencies in how a user interacts with applications and how applications respond to the user's requests following the specifications of the applications. By enforcing an application's correct responses to user actions, they are capable to identify vagabond events. Vagabond events are nothing but to outbound network events that are not generated by any user actions and may hence be due to anomalies [8].This work aims to show the viability of user intention-based dependence analysis for detecting suspicious network connections of a host in a concrete web browser setting. They enforce correct system behaviors, as opposed to anomalous characteristics. Their user intention-based traffic dependence analysis produces structures in network events. These structures across outbound requests enable improved context-aware security analysis. Dependence analysis on network flows builds a traffic-dependency graph based on the observed network events and user actions [8].

Analyzing the dependencies between network traffic and user activities has not been systematically investigated as a general approach for anomaly detection. Traffic dependency graph captures the causal relations of user actions and network events for improving host integrity. Result indicated that the feasibility of enforcing HTTP traffic dependencies [8].

Monowar Hussain Bhuyan[7] propose a special type of IDSs, called Anomaly Detection Systems, develop models based on normal system or network performance, with the aim of distinguishing both known and unknown attacks. Anomaly detection systems face a lot of problems including high rate of false alarm, ability to work in online mode, and scalability. The technological developments, open problems, and challenges over anomaly detection using incremental approach are also discussed. Here they present a survey in a structured manner along with current research issues and challenges in this field.
Experimental analysis have shows that for different type of attacks, some anomaly detection approaches are more successful than others. As a result, sufficient possibility exists for working toward solutions that maintain high detection rates while lowering false alarm rates. Incremental learning approaches that combine data mining, neural network and threshold based analysis for the anomaly detection have shown great assure in this region.

Burbeck et al. [16]propose a new method based on existing framework ADWICE (Anomaly Detection With fast Incremental Clustering) uses the first phase of the BIRCH clustering framework [17] to realize fast, scalable and adaptive anomaly detection. It expands the original clustering algorithm and be appropriate the resulting detection mechanism for analysis of data from IP networks. The performance is shows on the KDD99 intrusion dataset with on data from a test network at a telecom company. Their experimental analysis show that good detection quality (95%) and acceptable false positives rate (2.8%) think about the online, real-time distinctive of the algorithm. The number of alarms is more reduced by application of the aggregation techniques implemented in the Safeguard architecture.

Rasoulifard et al. [18] they proposed an important to increase the detection rate for known intrusions and also to detect unknown intrusions at the identical time. It is also important to incrementally learn new unknown intrusions. Most current intrusion detection systems make use of either misuse detection or anomaly detection. Consecutively to employ these techniques effectively, the authors propose an incremental hybrid intrusion detection system. This structure combines incremental misuse detection and incremental anomaly detection. The structure can learn new classes of intrusion that do not exist in data used for training. The structure has low computational complexity, and so it is suitable for real-time or on-line learning. The authors use the KDDcup99 intrusion dataset to establish this method.

Kalle et.al.[19] proposed a mechanisms of Anomaly detection is very costly in real-time. First, to deal with enormous data volumes, one needs to have efficient data structures and indexing mechanisms. Second, the dynamic nature of today's information networks makes the classification of normal requests and services difficult. What is reflect on normal during some time interval may be classified as abnormal in a new framework, and vice versa. These aspects create many proposed data mining techniques less appropriate for real-time intrusion detection. The authors look at the inadequacy of ADWICE and propose a new grid index that improves detection performance while preserving efficiency in search. Additionally, they propose two mechanisms for adaptive development of the standard model: incremental extension with new elements of normal behaviour, and a new feature that make possible not remembering of outdated elements of normal behaviour. It estimates the technique for network-based intrusion detection using the KDD99 intrusion dataset with on data from a telecom IP test network. The experimental analysis give up good

recognition quality and act as proof-of-conception for alteration of normality.

Zhong et al.[20] paper here an incremental clustering algorithm for intrusion detection using clonal selection based on a partitioning approach. It partitions the dataset into initial clusters by comparing the distance from data to cluster centroid with the size of cluster radius, and analyses the clustering data with mixed attributes by using an get better definition of distance compute. The purpose of this function optimizes clustering results by applying a clonal selection algorithm [21], and then labels clusters as normal or anomalous as suitable. The authors make an effort to find better cluster centroids to develop the partitioning quality which is calculated by the objective function. If the value of objective function is small, the sum of the distances from data to the cluster centers is also small and the objects in the same cluster are more close to each other. The technique efforts to optimize cluster results from one iteration to the next using the clonal selection algorithm [21]. The authors begin this incremental technique in requisites of high detection rate and low false positive rate.

3. PROPOSED METHODOLOGY

Here we proposed solution algorithm for support Vector Machine (SVM) to classify the data set in to number of clusters.

Algorithm for SVM:

1: Input: (x1,y1) ,.....(xn,yn),C,
2. Si← for all i=1,.....n
3. repeat
4. for i=1,.....n do
5. H(y) =
6. compute Ÿ=argmaxy
7. compute £i=max{0,maxy
8. if H(Ÿ)>£i +
9. Si←Si
10. w←optimize primal over S=
11. end if
12. end for
13. until no Si has changed during iteration.

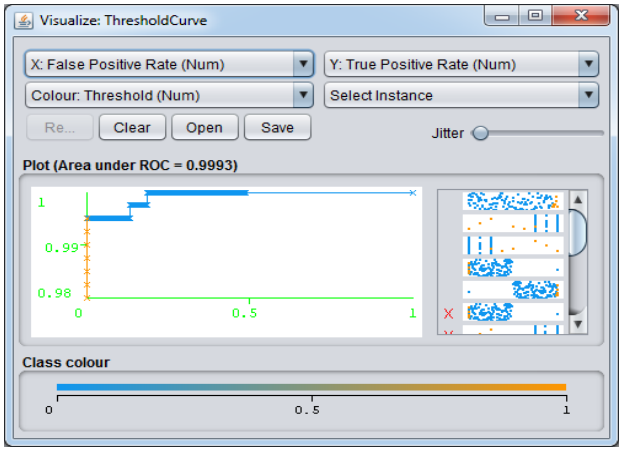
Decision Tree

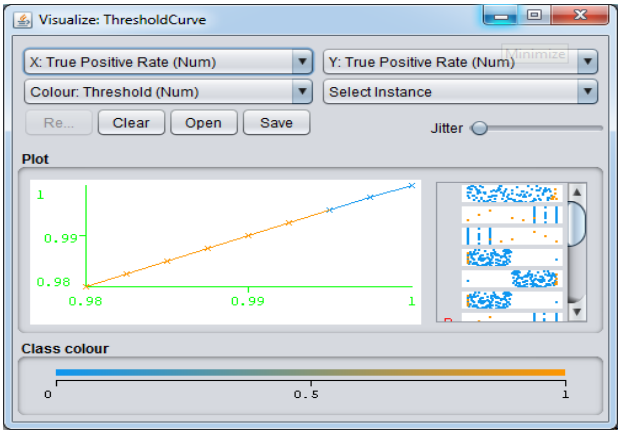
- **Vertical Partition based id3 decision tree**
- **Input Layer:**
- Define P_1, P_2… P_n Parties. (Vertically partitioned).
- Each Party contains R set of attributes A_1, A_2, …., A_R.
- C the class attributes contains c class values C_1, C_2, …., C_c.
- For party P_i where i = 1 to n do
- If R is Empty Then
- Return a leaf node with class value
- Else If all transaction in $T(P_i)$ have the same class Then
- Return a leaf node with the class value
- Else
- Calculate Expected Information classify the given sample for each party P_i individually.
- Calculate Entropy for each attribute (A_1, A_2, …., A_R) of each party P_i.
- Calculate Information Gain for each attribute (A_1, A_2,…., A_R) of each party P_i
- Calculate Total Information Gain for each attribute of all parties (TotalInformationGain( )).
- $A_{BestAttribute}$ ← MaxInformationGain( )
- Let V_1, V_2, …., V_m be the value of attributes. $A_{BestAttribute}$ partitioned P_1, P_2,…., P_n parties into m parties
- $P_1(V_1)$, $P_1(V_2)$, …., $P_1(V_m)$
- $P_2(V_1)$, $P_2(V_2)$, …., $P_2(V_m)$
- . .
- . .
- $P_n(V_1)$, $P_n(V_2)$, …., $P_n(V_m)$
- Return the Tree whose Root is labelled $A_{BestAttribute}$ and has m edges labelled V_1, V_2, …., V_m. Such that for every i the edge Vi goes to the Tree
- NPPID3(R – $A_{BestAttribute}$, C, ($P_1(V_i)$, $P_2(V_i)$, …., $P_n(V_i)$))
- End.

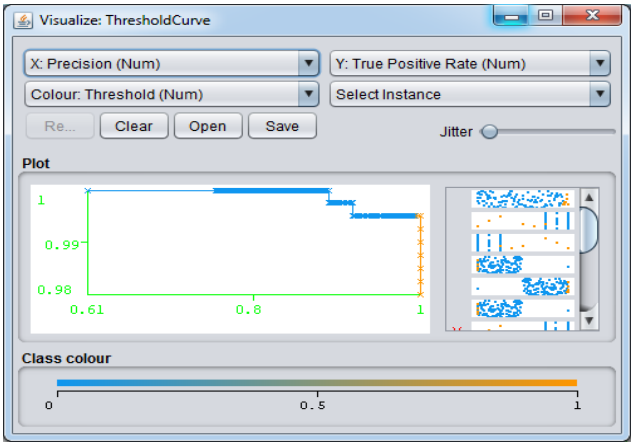
4. RESULT ANALYSIS

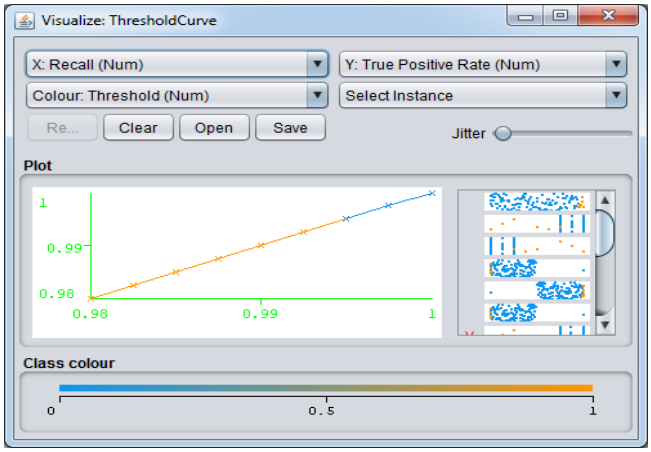
Number of instances	ID3_Mean absolute error	Vertical Partition Mean absolute error
20	0.2860	0.1167
25	0.280	0.276
50	0.310	0.290
100	0.350	0.298
200	0.380	0.310

Number of instances	ID3_time(milli sec)	Vertical Partition time(milli sec)
20	80	17
25	97	18
50	115	19
100	135	33
200	160	37









5. CONCLUSION

Network traffic anomaly refers to the status that the traffic behaviors deviated from its normal behaviors. It can bring great damage to networks and network equipments in a short time. The proposed technique implemented here provides high efficiency as compared to other existing technique for network anomaly detection.

REFERENCES

[1]. Wenke Lee,Salvatore J. Stolfo, Kui W. MokA, "Data Mining Framework for Building Intrusion Detection Models" DARPA (F30602-96-1-0311) and NSF (IRI-96-32225 and CDA-96-25374).

[2]. George Nychis, Vyas Sekar, David G. Andersen, Hyong Kim, Hui Zhang, " An Empirical Evaluation of Entropy-based Traffic Anomaly Detection" .

[3]. A. Patcha and J. M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Computer Networks (Elsevier Science), vol. 51, no. 12, pp. 3448–3470, August 22, 2007. [Online].Available: http://10.1016/j.comnet.2007.02.001

[4]. S. Kumar and E. H. Spafford, "An application of pattern matching in intrusion detection," The COAST Project, Department of Computer Sciences, Purdue University, West Lafayette, IN, USA, Tech. Rep.CSD-TR-94-013, June 17, 1994.

[5]. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol.41, no. 3, pp. 1–58, July 2009. [Online].
Available: http://doi.acm.org/10.1145/1541880.1541882.

[6]. Koller, D., and Sahami, M. Toward optimal feature selection. In Proc. of ICML (1996).

[7]. Monowar Hussain Bhuyan, D K Bhattacharyya and J K Kalita, "Survey on Incremental Approaches for Network Anomaly Detection "International Journal of Communication Networks and Information Security (IJCNIS),Vol. 3, No. 3,2011.

[8]. Hao Zhang, William Banick, Danfeng Yao and Naren Ramakrishnan "User Intention-Based Traffic Dependence Analysis for Anomaly Detection", IEEE Symposium on Security and Privacy Workshops (SPW-2012), pp. 104 – 112, 2012.

[9]. M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "NADO : Network anomaly detection using outlier approach," in ICCCS'11. ACM, February 2011, pp. 531–536.

[10]. Brauckhoff, D., Tellenbach, B., Wagner, A.,Lakhina, A., and May, M. Impact of traffic sampling on anomaly detection metrics. In Proc. Of ACM/USENIX IMC (2006).

[11]. Feinstein, L., Schnackenberg, D., Balupari, R., and Kindred, D. Statistical Approaches to DDoS Attack Detection and Response. In Proc. of DARPA Information Survivability Conference and Exposition (2003).

[12]. Karamcheti, V., Geiger, D., Kedem, Z., and Muthukrishnan, S. Detecting malicious network traffic using inverse distributions of packet contents.In Proc. of ACM SIGCOMM MineNet 2005 (2005).

[13]. Lakhina, A., Crovella, M., and Diot, C. Mining anomalies using traffic feature distributions. In Proc.of ACM SIGCOMM (2005).

[14]. Lee, W., and Xiang, D. Information-theoretic measures for anomaly detection. In Proc. of IEEE Symposium on Security and Privacy (2001).

[15]. Wagner, A., and Plattner, B. Entropy Based Worm and Anomaly Detection in Fast IP Networks.In 14th IEEE International Workshops on Enabling Technologies, Infrastructures for Collaborative Enterprises (WET ICE 2005) (2005).

[16]. K. Burbeck and S. Nadjm-tehrani, "ADWICE - anomaly detection with real-time incremental clustering," in In Proceedings of the 7th International Conference on Information Security and Cryptology, Seoul, Korea. Springer Verlag, pp. 4007-424, 2004.

[17]. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," SIGMOD Rec., vol. 25, no. 2, pp. 103–114, 1996. [Online]. Available: http://doi.acm.org/10.1145/235968.233324.

[18]. A. Rasoulifard, A. G. Bafghi, and M. Kahani, Incremental Hybrid Intrusion Detection Using Ensemble of Weak Classifiers, in Communications in Computer and Information Science. Springer Berlin Heidelberg, November 23 2008, vol. 6, pp. 577–584. [Online]. Available: http://10.1007/978-3-540 89985-3.

[19]. K. Burbeck and S. Nadjm-Tehrani, "Adaptive real-time anomaly detection with incremental clustering," Inf. Secur. Tech. Rep., vol. 12, no. 1, pp. 56–67, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.istr.2007.02.004.

[20]. C. Zhong and N. Li, "Incremental clustering algorithm for intrusion detection using clonal selection," in Proceedings of the 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application. Washington, DC, USA: IEEE Computer Society, 2008, pp. 326–331. [Online]. Available: http://dx.doi.org/10.1109/PACIIA.2008.256.

[21]. J. Li, X. Gao, and L. Jiao, "A novel clustering method with network structure based on clonal algorithm," in Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE Press, 2004, pp. 793–796.