Mining Anomaly using Association Rule

Mahadik Priyanka V.
Department of Computer Engineering ,
Smt. Kashibai Navale College of Engineering ,
Off Sinhgad Road,Vadgaon(Bk), Pune-411041, India.

Kosbatwar Shyam P.
Department of Computer Engineering ,
Smt. Kashibai Navale College of Engineering ,
Off Sinhgad Road,Vadgaon(Bk), Pune-411041, India.

ABSTRACT

In a world where critical equipments are connected to internet, hence protection against professional cyber criminals is important. Today network security, uptime and performance of network are important and serious issue in computer network. Anomaly is deviation from normal behavior which is factor that affects on network security. So Anomaly Extraction which detects and extracts anomalous flow from network is requirement of network operator. Using Histogram based detector to identify anomalies and then applying Association rule mining, anomalies will extracted. Apriori algorithm will use to generate the set of rule applied on metadata. Identification and Extraction of anomalous flow can be used for useful application e.g. Root cause analysis, Network forensics, Modeling anomalies etc.

General Terms

Aproiri Algorithm.

Keywords

Anomaly Detection, Anomaly Extraction, Association Rule, Data Mining.

1. INTRODUCTION

To extract anomaly first step is to detect them. For anomaly detection three different approaches are proposed. The first scheme is entropy based detection which applies entropy as its main feature to detect anomaly. The second is histogram based detection scheme which uses kullback-libeler distance and Apriory in order to detect and identify anomalies. The third scheme is SENATUS where robust version of Principle Component Analysis (PCA) is used for detection.

From high level view point anomaly detection is generally divided into three categories Supervised, semi supervised, unsupervised. In supervised technique knowledge about normal pattern and anomalous traffic pattern is essential according to which anomalies are classified and thus it lacks the ability to classify unknown anomalies. The Semi supervised technique requires knowledge of normal traffic pattern, if there is deviation from this pattern it generates alert. Main principle in unsupervised detection system is to compare the current interval to either previous reference interval to detect abnormal behavior.

Further detection technique view is classified into two groups first is volume based where number of flows, packets, bytes and other volume metrics are used and second is feature based system where field extracted from packet header field is used as metrics. It include source IP address, destination IP address, source port, destination port, protocols ,packet size, flow duration. Number of traffic feature distribution presents with anomalous events such as flash crowd, Alpha flows, denial of service, port scan, port sweeps.

An anomaly is defined as a "Deviation or departure from the normal or common order, form, or rule". Mining network anomalies are becoming an important factor for taking care of uptime, network security, and performance large scale networks. In this paper the aim is to take noteworthy steps toward a system that overcomes these criteria. Main thing is to try to find methods that are capable to detect a set of network anomalies, and to do this with high detection rate and low false alarm rate. By determining and interpreting the patterns present in the set of detected anomalies, it seeks to extract the anomalies from the data.

Table 1: A list of anomaly with definition

Anomaly Type	Definition
Flash Crowd	odd burst of traffic to single target, from a classic supply of sources
Alpha Flows	Unusually big volume end to end flow
DDoS	huge amount of traffic to one target from a number of sources
Port Scan	discover to many target ports on a small set of target addresses

The Table 1 shows the different types of anomalies and their respective definition.

The rest of paper is constructed as follows. Related work is described in section 2.Section 3 proposed architecture. Section 4 consists of Mathematical model. Used algorithm is described in section 5 finally section 6 concludes the paper.

2. RELATED WORK

A number of studies have focused on developing feature based anomaly detection schemes[4],[5] and number of studies have focused on extracting anomalies[3],[6]. D.Brauckhoff et al.[1] proposed anomaly extraction which first detects anomaly by using Histogram based detector and association rule is used to extract those anomaly. The technique that model flows and applies frequent item set mining to find large set of flow with identical values in one or more features. Four functional blocks are described named as Histogram cloning, Voting, Flow is pre-filtering and AR mining which extract anomalies.

Histogram based detector [4] which proposed by Andreas Kind is anomaly detector, generates metadata. It uses KL (Kullback-Leibler) distance to detect anomalies. It first Select features and construct histograms then map it into metric space. It then cluster and extract models and finally it classifies the anomalies. Use of histogram gives additional views of network.

D.Brauckhoff et al [5] used Principal Component Analysis (PCA) which has been first planned as a method for traffic anomaly detection technique. It consists essentially of two components, first is a decision component and second is an entropy reduction component. It applying statistical tests to a decision variable issued from the first step. D.Brauckhoff proposed a signal processing approach to anomaly detection. The challenge was to design filters that would be adapted to anomaly features.

For detection of anomalies Fernando Silveira Thomson and Christophe Diot Thomson introduced Unsupervised Root Cause Analysis (URCA).It isolates anomalous traffic and classifies alarms with least manual assistance and high correctness. URCA has an algorithm for each step of root cause analysis process, first is identifying the flows whose change has triggered the anomaly, and second classifying the root cause event from the characteristics of the identified flows [9].

Marc Ph. Stoecklin proposed technique to detect traffic anomalies and provide decision support. It is a Flow-based approach to detect network anomalies. The detection on 2 abstraction levels first was ability to expose anomalies of different natures and second was interpretable visualization and graphical reports of abnormal events [10].

Paul Barford, Jeffery Kline, David Plonka and Amos Ron [7] found a wavelet system that effectively isolates both short and long-lived traffic anomalies. In addition to this system, they developed the concept of a deviation score which considers signal variation in both the high and medium frequency bands. Drawback of this system is that it unable to evaluate the impact of additional features in deviation scores.

In the paper [8], D.Brauckhoff et al finds the Root cause of network anomalies which uses frequent item set mining to extract anomalies and summarize traffic flows which cause anomaly.

Entropy as a summarization device used by Anukool Lakhina, Mark Crovella, and Christophe Diot[6] to show that the analysis of feature distributions leads to important advantage on two faces first, it enables highly sensitive detection of a wide range of anomalies, augmenting detections by volume-based methods, and second it enables automatic categorization of anomalies by means of unsupervised learning.

In the paper [2],[3] the respective author suggested the importance of data mining to remove anomalies from the network. Rakesh Agrawal, Ramakrishnan Shrikant explained Algorithm Apriori, Apriori Candidate Generation, Algorithm AprioriTid, and Algorithm AprioriHybrid in detail with the essential features of the associative rule.

3. PROPOSED ARCHITECTURE

Architecture of the anomaly extraction problem is given in figure. To observe network traffic and detect anomalies in network a number of different histogram-based anomaly detectors are used. Upon detecting an anomaly, by taking union of meta-data provided by the histogram-based anomaly detectors, a set of suspicious flows is prefiltered. This prefiltering is essential since it reduce a large portion of the normal flows. By applying association rule mining, report of frequent item-sets in the set of suspicious flows is generated. Figure 1 illustrate metadata generation by voting from k histogram clones.How anomalous flows are summarized in item set by association rule mining is illustrated in figure 2.Functional blocks of anomaly extraction are histogram cloning, voting, flow prefiltering and association rule mining.

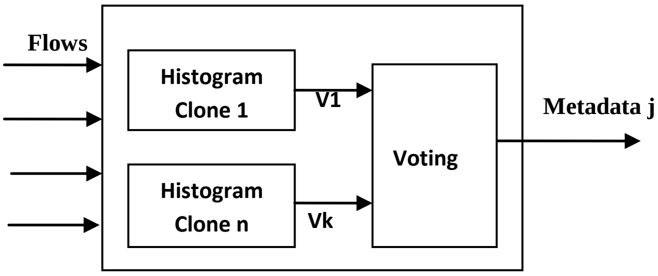

Figure 1:Anomaly Detection

3.1 Histogram Cloning

For Histogram Detector (HD), Histogram cloning is a challenging technique. The objective is to maintain number of randomized histograms of the same feature, therefore it will obtain additional views of network traffic. Corresponding to n different traffic features created by n histogram-based detectors is recognized by this method. for each of the n features there are m bins per time interval. Hash function is then applying to each of the clones. Each feature value is placed randomly into one of the m bins. For histogram detection, KL distance is computed between every newly created distribution and a reference distribution, if a KL distance goes beyond a given threshold, the algorithm creates an alarm on the set of time bins and the corresponding set of feature values within each time bin. Histogram based anomaly detector is used that applies histogram cloning which gives multiple randomize histogram to obtain additional views of network and uses the Kullback libeler (KL)distance which detect anomalies. If KL distance exceeds a given threshold then algorithm generates alarm.

3.2 Voting

After histogram cloning, if clone generates alarm on suspicious flow then histogram bin undergoes in interactive process. If sufficient amount of histogram clones generate an alarm on same traffic feature value then it will be included in metadata. A voting strategy between each histogram clone for a specific feature generates meta-data. It will undergo an iterative process if a clone generates an alarm on a histogram bin that removes suspicious flows until no alarm is generated. The voting scheme is introduced at this stage, and it will be included in the metadata for additional pre-filtering, if an enough quantity of histogram clones generates an alarm on the same feature.

3.3 Flow Profiltering

The goal of flow prefiltering is, from the set of all metadata, metadata that match is filter and then by taking union of such metadata smaller dataset is generated. Smaller dataset will lead to decrease in processing time and increase in detection rate. Before applying AR mining one wants to dataset based on the set of candidate flows is filter. Each flow record in the candidate flows consists of at least two feature values, and at most four, e.g. [source AS, destination AS, source port, destination port]. The motivation behind this is to filter only the flows that match the union of a flow record in the set of all flows, hence generating a smaller dataset. Another and even more important reason for pre-filtering is the impact on the detection rate. It will most likely result in a higher rate of FP item sets, if one includes all flows without any filtering. Therefore, it is advantageous to only apply the Apriori algorithm to a dataset containing a small amount of normal flow.

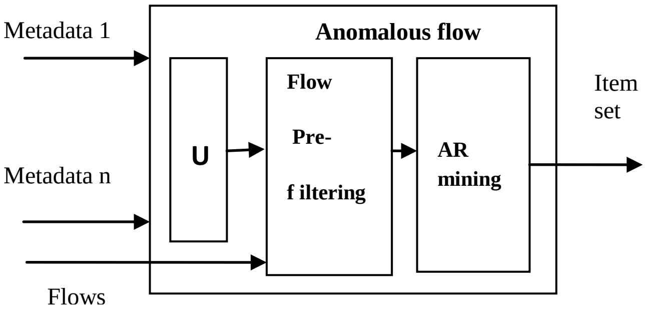

Figure 2:Anomaly Extraction

3.4 Association rule mining

Association rule mining also known as learning is a data mining technique that grew rapidly in popularity partly due to an article authored by R. Agrawal et. Al.The technique was originally proposed to find relations between products in a large scale database consisting of customer transactions. The association rules could further on be employed as a tool to increase profit, in particular how to design coupons and what to put on sale. For example, a rule might say that 80 percent of customers purchasing beer also purchase potato chips. This particular fact may be exploited by placing both products next to each other on the shelves, and will most likely lead to an increase in sales.

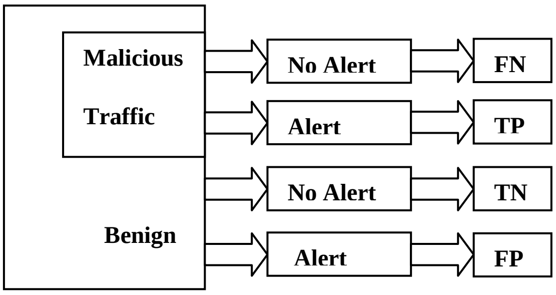

Figure 3:Possible Output of Anomaly Identification

When identifying an anomalous event it will be classified as either a true positive (TP) or a false negative (FN). The former means that the anomaly was due to a malicious activity, whereas the latter corresponds to a malicious event that did not trigger an alarm. Events that are not malicious are often referred to as benign activities. A benign event is either classified as a true negative (TN) or a false positive (FP). A TN is when a benign event does not trigger an alarm, whereas benign activity that generates an alarm is classified as a FP. An illustration of this concept is shown in Figure 3.

3.5 MATHEMATICAL MODEL

1. Each clone includes an anomalous feature value in the set V_k with probability pa , while a normal feature value is selected only if it collides on one of the selected bins and has, thus, a selection probability of $p_n = \frac{b}{m}$, where m is the total number of bins[1].

$$Pa \geq \sum_{i=1}^{k} \binom{k}{i} p_a^i \ (1 - P_a)(k - i) \text{………} \ (1)$$

2. And an upper bound for the probability that an anomalous feature value is missed [1]

$$p_{\bar{a}} \leq \sum_{i=1}^{k} \binom{k}{i} p_a^i \ (1 - P_a)(k - i) \text{……..} \ (2)$$

3. The probability that a normal feature value is included by more than k clones, on the other hand, is given by [1]

$$Pn \geq \sum_{i=1}^{k} \binom{k}{i} p_a^i \ (1 - P_n)(k - i) \text{…….}(3)$$

3.6 ALGORITHM USED

3.6.1 Apriori algorithm

```
Step 1-$L_1$ :={ frequent 1-itemsets};
Step 2-k:=2; // k represents the pass number.
Step 3-While ($L_{k-1}$≠ф) do begin
Step4-$c_k$: = New candidates of size k generated from $L_{k-1}$;
Step 5-forall transactions TЄD do begin
Step 6-Increment the count of all candidates in $c_k$ that are contained in T.
Step 7-end
Step 8-$L_k$: = All candidates in $c_k$ with minimum support.
Step 9-K:= k+1;
Step 10-end
```

3.6.2 Apriori Candidate Generation

Gives L_{k-1}, the set of all frequent (k-1)-itemsets, the algorithm returns a superset of the set of all frequent k-itemsets, The function works as follows. First, in the join step, it join L_{k-1} with L_{k-1}

```
Step 1-insert into $c_k$
Step 2-select p.item1, p.item2… p.itemk-1, q.itemk-1;
Step 3-from $L_{k-1}$ p, $L_{k-1}$ q
Step 4-where p.item1 = q.item1… p.itemk-2=q...Itemk-2, p.itemk-1<q.itemk-1;
Step 5-Next, in the prune step, delete all itemsets c Є $c_k$such that some (k-1) subset of c is not in Lk-1:
Step 6-forall itemsets, c Є $c_k$ do
Step 7-forall (k-1)-subsets s of c do
Step 8-if (s Є $L_{k-1}$) then
Step 9-delete c from $c_k$;
```

4. CONCLUSION

The paper presented problem of Anomaly Mining. Anomaly Mining takes traffic flow as input and finds anomalous flow as anomaly. Study of histogram based detector that provide Meta data for filtering suspicious flow and association rule for extracting anomaly is done. For detecting and classifying anomalies, traffic feature distributions are a rich source of information. For discovering association rules between items in a large database of transactions Apiori algorithm plays significant role. This work could provide useful additional features for network traffic monitoring.

5. REFERENCES

[1] D.Brauckhoff, X.Dimitropoulos, AWanger, and K.Salamatian: Anomaly Extraction in Backbone network using Association Rule.IEEE 2012.

[2] W.Lee and S.J.Stolfo, Data mining approaches for intrusion detection, in proc.7thUSENIX Security Symp., 1998, vol.7, p.6.

[3] Ramakrishnan Srikant: Fast algorithm for mining association rule and sequential pattern, at university of Wisconsin, 1996.

[4] A.Kind, M.P.Stoecklin, X.Dimitropoulos.Histogram-based traffic anomaly detection, IEEE Trans. Netw Service Manage., voi.6, no.2, pp.110-121, Jun.2009

[5] D. Brauckhoff, M. May, and K. Salamatian, Applying PCA for Traffic Anomaly Detection: Problems and Solutions, in IEEE INFOCOM MiniConference, 2009.

[6] A. Lakhina, M. Crovella, and C. Diot, Mining anomalies using traffic feature distributions, in Proc.ACM SIGCOMM, 2005, PP.217-228.

[7] Paul Barford, Jeffery Kline, David Plonka and Amos Ron.A Signal Analysis of Network Traffic Anomalies, IMW' 02, Nov. 6-8, 2002, Marseille, France

[8] I.Paredes Oliva,X.mitropoulos,M. M. Dante,P.Barlet-Ros, D. Brauckhoff, Automating Root-Cause Analysis of Network Anomalies using Frequent Itemset Mining,SIGCOMM'10,Aug 30-Sept 3,2010.

[9] Fernando Silveira Thomson and Christophe Diot Thomson, UPMC Paris Universitas, URCA: Pulling out Anomalies by their Root Causes, in proc.IEEE INFOCOM, Mar.2010.pp.1-9.

[10] M. P. Stoecklin, J.-Y. L. Boudec, and A. Kind,A two-layered anomaly detection technique based on multi-modal flow behavior models, in PAM: Proceedings of 9th International Conference on Passive and Active Measurement, ser. Lecture Notes in Computer Science. Springer, 2008,pp. 212–221.