Survey on Anomaly Extraction System Using Featured Histogram and Mining Rules

Mr. Sandeep J. Kamble, Prof.Sachin Deshpande

Abstract—With the Progress in anomaly extraction in backbone network, the networking demand for finding out anomaly is growing that also increases demand for finding out root-cause analysis, network forensics, attack mitigation, and anomaly modeling especially in backbone network. Also when its scope enlarges to rich traffic and very small number of false positive there will be need to maintain best method for mining. Numerous Techniques have been developed for anomaly extraction and data mining and purpose of this paper is to categorize and evaluate these methods. For finding out one that is abnormal or exception the best method will be discussed. Paper Also Summarizes several methods to ensure highly extracted data which includes Apriori algorithm, Fp growth algorithm, Enhance FP growth algorithm such a mining methods , Also with the rich traffic data and false positive rate of methods with Several scenarios will discussed.

Index Terms— anomaly, association rules, computer network, data mining , detection algorithm .

I. INTRODUCTION

Detecting and identifying network anomalies are becoming an important factor to consider when taking care of network security, uptime and performance of ISPs and large scale networks. An anomaly is defined as a "Deviation or abnormality from the normal or common order. Anomalies such as denial of service attacks, port scans, worms, etc can be found at any time in the network traffic. These network anomalies waste network resource, cause performance degradation of network device and end hosts, and conduct to security issues concerning all internet users. Thus, a correctly detecting such anomalies has become an important problem for the network community to solve EASE OF USE.

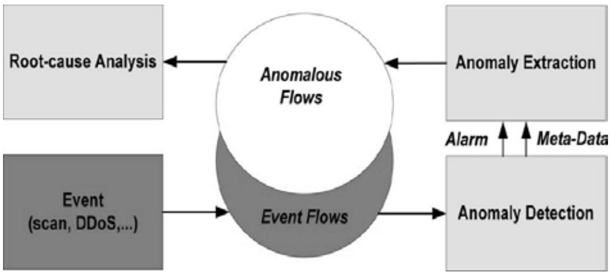

Fig. 1. Goal of Anomaly Extraction

A.Motivation

Identifying the network anomalies is critical for the timely mitigation of events, like failures or attacks that can affect the performance and security of a network. An anomaly is defined as a "Deviation or abnormality from the normal or common order. While studying anomalies affecting computer networks, we consider actions that differ from normal network behavior, such as significantly increased traffic, use of new protocols and malicious attacks. Anomaly detection techniques are the last line of defense when other approaches fail to detect the security threads and other security problems. The foremost challenge in identifying and detecting anomalies is the fact that they can be caused by a vast set of events. While studying, researchers have pose number of interesting research problems like modeling, involving statistics and efficient data structure. Nevertheless researchers have not yet gain widespread adaption, as a number of challenges, like calibration and reducing number of false positive rate remain to be unsolved. Motivated by this, the problem of identifying the traffic flows correlated with an anomaly during a time interval with an alarm. Anomaly extraction reflects the goal of gaining the more information about the anomaly alarm, which or without additional meta-data, is often meaningless for the network operator. Data mining techniques are used to identify anomalous behavior. Identified anomalous flows can be used for a number of applications, like network forensics, root-cause analysis of the event causing an anomaly, anomaly modeling, and improving anomaly detection accuracy.

II. ANOMALY EXTRACTION

The size of database has increased rapidly these days, Due to which the unknown attacks (anomalies) are also increasing. So anomaly detection has become an important area for both commercial interests as well as academic research. In [1] introduces an automated anomaly extraction system which extract features is going to affect the anomaly. Anomaly detection typically used from the perspectives of network monitoring and network security. This has led to a growing interest in the development of tools capable in the automatic extraction of knowledge from data.

A. Anomaly Extraction

The set of data points those are considerably different than the remainder of the data. Anomaly detection refers to detecting patterns in a given data set that do not conform to an established normal behavior. The abnormal patterns thus detected are called anomalies and translate to critical and actionable information in several application domains. Anomalies are also referred to as outlier, surprise deviation etc. Most anomaly detection algorithms require a set of purely normal data to train the model and they implicitly assume that anomalies can be treated as patterns not observed before. An anomaly detection system may provide meta-data. This meta-data is generated by histogram based anomaly detectors. To extract anomalous flows, one could build a model describing normal flow characteristics and use the model to identify deviating flows. Building such a microscopic model is very challenging due to the wide variability of flow characteristics. One could compare flows during an interval with flows from normal or past intervals and search for changes.

To extract anomalous flows, one could build a model describing normal flow characteristics and use the model to dentify deviating flows. However, building such a microscopic model is very challenging due to the wide variability of flow characteristics. Similarly, one could compare flows during an interval with flows from normal or past intervals and search for changes, like new flows that were not previously observed or flows with significant increase/decrease in their volume [2], [3]. Such approaches essentially perform anomaly detection at the level of individual flows and could be used to identify anomalous flows.

B. Data Mining

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The term data mining or knowledge discovery in database has been adopted for a field of research dealing with the automatic discovery of implicit information or knowledge within the databases. The implicit information within databases, mainly the interesting association relationships among sets of objects that lead to association rules may disclose useful patterns for market basket analysis, decision support, financial forecast, medical diagnosis, protein sequences etc. Data mining comprises data, information and knowledge.

C. Anomaly Detection Techiniques

The labels associated with a data instance denote that instance is normal or anomalous. Labeling is often done manually by a human expert and hence requires substantial effort to obtain the labeled training data set. The anomalous behavior is dynamic in nature, so new type of anomalies arises, for which there is no labeled training data available. Anomaly detection techniques can operate in one of the following three modes:

1) **Supervised anomaly detection**: In supervised anomaly detection, predictive model is built for normal as well as anomaly classes. Techniques trained in supervised model assume the availability of a training data set which has labeled instances for normal and anomaly class. Each instance is compared with the model either it belongs to normal class or anomaly class.

2) **Semi-Supervised anomaly detection**: In semi- supervised anomaly detection model, assume the training data has labeled instances for only the normal class. These techniques are widely applicable than supervised techniques. The approach used in such techniques is to build a model for the class corresponding to normal class, and use the model to identify anomalies in the test data. These techniques are not commonly used, because it is difficult to obtain a training data set which covers every possible anomalous behavior that can occur in the data set.

3)**Unsupervised anomaly detection**: In unsupervised anomaly detection techniques the training data is not required. So these techniques are widely used. The techniques in this category make the implicit assumption that normal instances are more frequent than anomalies in the test data set. Sometimes this assumption is not true then such techniques suffer from high false alarm rate.

III. FEATURED HISTOGRAM-BASED TRAFFIC ANOMALY DETECTION

Feature-based anomaly detection examines the range of network traffic features, instead relying solely on traffic volume. Common network traffic features are IP header fields, like source/destination IP address, source/destination port numbers and TCP flags. Feature-based detection techniques rely on the observation that under normal conditions traffic features exhibit regular patterns that may be deviated by anomalies. A histogram is a distribution of the number of flows, bytes or packets over possible values of a traffic features. Histogram gives detailed characteristic of traffic features. For example a histogram could give the number of flows associated with different destination IP address over period of time. Each histogram is then embedded into a metric space so that dissimilar histograms are positioned close apart in the space [4]. Data mining techniques are used to identify patterns of typical behavior. These patterns are finally compared to the online behavior of a network to identify deviations and trigger anomaly alarms.

Histogram-based anomaly detectors have been shown to work well for detecting anomalous behavior and changes in traffic distribution. Histogram based detectors built for our evaluation that uses Kullback-Leibler (KL) distance to detect anomalies. Each histogram detectors monitors a flow feature distribution, like the distribution of source ports or destination ports or destination IP addresses etc. At the end of each time interval, it computes for each histogram the KL distance between the distribution of the current interval and a reference interval.

Compared to the few feature-based detection techniques presently available, the Histogram-Based scheme has two main differences [4]. Firstly, it models the detailed characteristics of histograms, which enables us to identify a wider range of anomalies. In contrast, previous anomaly detection approaches only coarsely model the characteristics of traffic features, which render certain anomalies undetectable. Secondly, the Histogram-Based scheme mainly focuses on identification of anomalies in enterprise networks rather than large backbone networks, which have been the focus of most previous studies [4].
The two simple observations encompass the main ideas of anomaly detection approach.

- Feature histograms exhibit regular patterns that reflect the normal behavior of the network.
- Network anomalies may deviate the normal patterns of one or more features.

Histogram-Based Detection consists of following steps:
a) **Select features and construct histograms**:
A number of traffic features are used to identify anomalies in a network. Features like source/destination IP address, source/destination port numbers, TCP flags, protocol number, flow duration etc. Additional features are extracted from packet header. According to number of flows or occurrence of features a histogram is constructed.

b) **Map into metric space**:
For each selected feature, they prepare a set of training histograms. Then they map the vectors of training histogram into metric space. So that (1) two similar histogram are close in space, (2) two dissimilar histogram are far away. A number of different approaches are used to quantify how similar two histogram are.

c) **Clustering**:
Clustering is needed for identifying and modeling patterns of normal network behavior. K-means and hierarchical clustering algorithms are used. Hierarchical clustering iteratively groups clusters forming new clusters. Having performed clustering, it is necessary to distinguish the clusters that correspond to the anomalous and normal behavior.

d) **Classification**:
They call the set of clusters that model the normal behavior of a network baseline. In the detection process, the operation of a network is monitored and vectors for the different features are constructed. These vectors are, then, compared to the constructed models to measure how the observed online behavior differs. Two design choices are critical here. The first is how one computes the distance between a vector and a cluster of arbitrary shape. The second design option is the selection of necessary threshold(s) for raising an anomaly alarm.

IV. MINING RULES

A. Apriori Mining Rule

The standard algorithm for discovering frequent item-sets is the Apriori algorithm. Apriori [5] computes in each round the support for all candidate -item-sets. At the end of each round, the item-sets with frequency above the minimum support parameter are selected. The frequent item-sets of round are used in the next round to construct candidate -item-sets. The algorithm stops when no -item-sets with frequency above the minimum support are found.

Apriori uses level-wise search where k-itemsets (an itemset that contains k-items) are used to explore (k+1)-itemsets. In the beginning, the set of frequent 1-itemsets is found. This set contains items that satisfy minimum support and isKaur et al., In each subsequent pass, we begin with a set of itemsets found to be frequent in the previous pass. This set is used for generating new itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, we determine which of the candidate itemsets are actually frequent and they are used in the next pass. Therefore, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. An important property called Apriori property is used to reduce the search space which is described as: "All nonempty subsets of a frequent itemset must also be frequent" [8]. How Lk-1 is used to find Lk is consisting of two steps, join and prune actions as followed: 1. Join Step: Join Lk-1 with itself to obtain the candidate itemset Ck. 2. Prune Step: Scan the database to determine the count of each candidate in Ck. When the count is less than the minimum support count, it should be delete from the candidate itemsets. Meanwhile, if any (k-1) subset of candidate k-itemset is not in Lk-1 then the candidate cannot be frequent either and so can be removed. After this, we get k-itemset which is denoted by Lk.

B. FP-Growth Mining Rule

FP is a compressed Representation of the original database because only frequent items are used to construct tree other irrelevant information are pruned. It uses divide and conquer method that considerably reduces the size of subsequent conditional FP tree. It scans the database twice and performance is not influenced by the support factor

The process of constructing FP-Tree is as follows: First create root of the tree labeled with "null". Scan database second time as we scanned first time to create 1-itemset. Process items in each transaction in decreasing order of their frequency. A new branch is created for each transaction with the corresponding support. If the same node is encountered in another transaction, just increment the support count of common node. Each item points to the occurrence in the tree using the chain of node-link by maintaining the header table. Now the problem of mining frequent patterns in database is transformed to that of mining the FP-Tree. The constructed FP-tree is mined as: 1. Start from each frequent length-1 pattern (as an initial suffix pattern), construct its conditional pattern base. 2. Then, construct its conditional FP-Tree and perform mining on such a tree. 3. The pattern growth is achieved by concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-Tree. 4. The union of frequent pattern (generated by step 3) gives the required frequent item set [9]. A. Advantages of FP-Growth Algorithm The main advantages of FP-Growth algorithm are: FP-Tree is a compressed representation of the original database because only frequent items are used to construct• the tree, other irrelevant information are pruned. This algorithm only scans the database twice.• FP-Tree uses a divide and conquers method that considerably reduced the size of the subsequent conditional FPTree• [10].

C. Enhance FP-Growth Mining Rule

A study on the performance of the FP-growth method shows that it is efficient and scalable and is about an order of magnitude faster than the Apriori algorithm. However, it has its own problems, during mining frequent itemsets, mass of conditional pattern trees are generated recursively. It cost a lot of time to generate and release these trees. Besides, FP-tree and the conditional pattern tree need to traverse from top to down, but mining frequent pattern is in an opposite way. Finally, traversing in couple ways and a larger memory space is needed to main these tree.

FP stands for frequent pattern. The main disadvantage of FP growth is that it is very difficult to implement because of its complex data structure. To overcome these problems, I introduced Enhanced-FP. The main strength of Enhanced-FP is its simplicity and its based on single link list [6] Enhance-FP growth reduce the complexity associate with FP-growth algorithm. The Enhance-FP growth process the transaction directly.

An enhanced FP-growth algorithm based on compound single linked list. The algorithm introduces the compound single linked list to improve the structure of the FP-tree. The improved FP-growth is mined in one direction, using the header table in the former FP-tree, storing them in a sequence table, ordering the frequent item sets in descending sequence according to the min_sup, then a compound single linked list is formed. Through traversing each transaction's frequent item sets stored in its single linked list, mining the frequent patterns directly without generating conditional FPtree.

The steps to construct the compound single linked list

1. The first scan of database is the same as the FP-tree. The scan of the database derives the set of frequent items (1- item sets) and their support counts (frequencies). The set of frequent items is sorted in the order of descending support count. This resulting set or list is denoted L and an item header table is built.

2. The second scan of database is different from the FPgrowth. It is processing the items in each transaction in L order, and then inserting the items in each transaction into the single linked list recursively. The items' order in each single linked list is according to L order.

So using compound single linked list mining the frequent patterns directly without generating conditional FPtree. The new one will improve the algorithm both in runtime and the main memory consumption.

V. CONCLUSION

In this paper we propose a study of new method to Enhanced the FP algorithm. In which a data structure used is, called compound single linked list. The improved FP-growth is mined in one direction, using the header table in the original FP-tree, and a compound single linked list is constructed. The study shows that the improved algorithm is better than the Apriori & FP-growth, both in runtime and the main memory consumption.

REFERENCES

[1] D. Baruckhoff, X. Dimitropoulos, A. Wagner, and K. Salamatian, " Anomaly Extraction In Backbone Networks Using Association Rules" , in proc. IEEE ACM TRANSACTION ON NETWORKING, VOL 20. NO 6, DECEMBER 2012.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen, "Sketch-based change detection: Methods, evaluation, and applications," in Proc. 3rd ACM SIGCOMM IMC, 2003, pp. 234–247.

[3] G. Cormode and S. Muthukrishnan, "What's new: Finding significant differences in network data streams," IEEE/ACM Trans. Netw., vol. 13, no. 6, pp. 1219–1232, Dec. 2005.

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

[5] Ding Zhenguo, Wei Qinqin, Ding Xianhua "An Improved FP-growth Algorithm Based on Compound Single Linked List". In Proc. IEEE 2009.

[6] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in Proc. 20th VLDB, Santiago de Chile, Chile, Sep. 12–15, 1994, pp. 487–499.

[7] X. Li, F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone, and A. Lakhina, "Detection and identification of network anomalies using sketch subspaces," in Proc. 6th ACM SIGCOMM IMC, 2006, pp. 147–152

[8] Han J. and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, 2nd Edition.

[9] Pinki Sharma and Rakesh Sharma, "Study of Mining Frequent Patterns at Various Levels of Abstraction", International Journal of Advanced Research in Computer Science, Volume 1, No. 2, pp. 197-201, July-August 2010.

[10] Sotiris Kotsiantis and Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview", International Transactions on Computer Science and Engineering, Volume 32 (1), pp. 71-82, 2006