Stream Data Mining and Anomaly Detection

Mohammadjafar Esmaeili
Ph.D in Information Security
118 Sill Hall, Eastern Michigan University
Ypsilanti MI 48197 USA

Arwa Almadan
Ph.D in GIS
118 Sill Hall Eastern Michigan University
Ypsilanti MI 48197 USA

ABSTRACT
Detecting anomaly behaviors is one of the most challenging tasks for Information Systems (IS) administrators. The anomaly behavior is defined as any behavior from either inside or outside of the organization's information system that deviates from normal; this includes insider attacks as well as any behavior that threatens the confidentiality, integrity and availability of the organization's information systems. One of the strategies to detect an anomalous behavior is to create a clustering or classification model by utilizing data mining methodologies. The models could be generated from previous historical data or it could be based on current data. Although these models could identify normal and abnormal behavior, they couldn't satisfy the growing demand for better information security. The primary drawback of using these methods are a high rate of false positive; the model becomes outdated and there is high demand to maintain the models' integrity; and they have low response rate. This study attempts to overcome some of the disadvantages in the current data mining models, which have been used to detect anomaly behaviors. Moreover this research will attempt to introduce a model that utilizes stream data mining to actively monitor network traffic for anomaly detection.

General Terms
Information security, Network Security, and Stream Data Mining

Keywords
Stream Data Mining, Anomaly Detection, Data Mining, Supervised Data Mining, and Unsupervised Data Mining.

1. INTRODUCTION
According to Conorich, there is a critical need to detect anomalies in organizations' networks [2]. With a growing online connectivity among corporations and customers, detecting anomalies becomes very challenging. Securing organizations' information systems against hackers and internal threats is one of most challenging tasks in the field of information security. Although researchers are attempting to provide security for organizations' information systems, they have to make sure that they are not rejecting legitimate users or reducing the performance of the information systems. One of the systems that can be used to detect anomalies and monitor a network is Network Intrusion Detection System (NIDS). Conorich states that "good security cannot be obtained solely through the purchases and use of tools or technology" [2]. The NIDS systems will sit on the network and they log the network without taking proper action. In another word, the NIDS systems are passive and only collect data about the incidents in the networks and they cannot actively block the intrusions. Moreover, collected data by NIDS systems could be overwhelming and it is not an easy task for network administrators to go through the collected logs.

"Statistical techniques or a frequent episode mining" is one of the tools that can be used to analyze the collected data from NIDS systems [3]. The NIDS are primarily signature based and if they detect a behavior on the network that matches with a signature on their database they will raise an anomaly detection flag. After NIDS system detects an anomaly behavior the network administrator needs to take a next proper action to mitigate the possible risks. Intrusion systems are developing with same speed as the development of information security detection tools. In other words, the NIDS systems signature database needs to be updated more frequently to detect the anomalies and they are limited to existing signatures. Currently, some scientists are taking it one step further to make the process of anomaly detection systems more advanced so they actively detect anomalies [1, 4]. For example, Anderson, Selby, and Ramsey utilized unsupervised clustering methodology to create clusters that are based on the users' roles within an organization [4]. The researchers utilized historical data to create the clusters model. This model would later recognize any deviation from the clusters, known as "peer groups", as anomaly behavior. In other words, the proposed methodology used a data mining approach, clustering in this case, to model users' behavior (based on their roles), which are the clusters, and then try to recognize the outliers. Finally, the outliers will recognize as anomaly.

This study will present how stream data mining can be used in conjunction with other data mining techniques to detect anomalies in network traffic in near real time. Stream data mining has its own unique characteristics that need to be considered. Anomaly detection in high value streams is one such area. Faster detection of anomalies will lead to faster resolution. This study attempts to introduce a new approach that reduces false positives and continuously updates the anomaly detection models. One of the main goals of our approach is to train a better model for anomaly detection systems that can be used by network administrators.

2. BACKGROUND

2.1 Key Terminologies
Anomaly Behavior – Behavior from inside or outside the system that endanger the Confidentiality, Integrity, and Availability of organizations' information systems. Identifying mainstream instances versus outliers becomes a key initial step for defining anomaly behavior.

Sliding Window – Area in the memory that stores the data which is actively being mined, consists of a grouping of basic windows.

Basic Window – Portion of data that will be selected from network communications for further investigation.

Intrusion Detection System – is used to detect malicious and anomaly activity inside of the network or on a specific host. Gregory, states "there are two types IDS: Network-based IDS

(NIDS) and Hosted-based IDS (HIDS)" [3]. NIDS monitors traffic on the network while HIDS monitors software programs that run on the servers (although this is where it is typically used, HIDS may monitor programs on a host). Both send alerts when unwanted activities such as viruses, worms, denial of service attacks, etc. are detected.

Stream Data Mining – A method of data mining that can be used on extremely large data sets or on data that is actively streaming and changing like network communication.

2.2 Background

Anderson, Selby, and Ramsey research focus on creating a data mining model by utilizing clustering techniques [1]. The users in an organization categorized in "peer-groups", which means they have a similar behavior and role in organizations. This study assumes that people with same functionality in organizations will have same access pattern to an organization's information system. For example there is no need for a help desk group to have access to financial documents and if a person form this group tries to access any resources out of his/her group permissions the system will raise a insider attack flag. In other words, the people will cluster in defined groups based on their functionality and their normal behavior and any deviation will be considered intrusion. One of the main disadvantages of this system is false positive and false negative, which is rejecting legitimate users and accepting unauthorized access. Moreover, Scarfone and Mell state that Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS) by themselves have some disadvantages [6]. For instance, IDS are passive and they cannot prevent intrusion, and IPSs have high "positive false rate", which they block legitimate activities.

For these reasons, there is a demand for some complimentary technologies such as: network forensic analysis tools, anti-malware technologies (antivirus software and antispyware software), firewalls and routers. Since IDS and IPS are looking to find the pattern that matches their signature data base they are vulnerable to getting outdated. In order to actively detect anomalies IDS and IPS needs to continuously get updated with new signature database. In our study we tried to introduce a model that overcomes outdated signature database and false positives.

Stream data mining is a growing and complex facet within the data mining field, but it has very specific uses. The first primary use is handling data in streams that is changing continuously. Examples of streaming data are: credit fraud protection, mining E-commerce data, web mining, stock analysis, network intrusion detection, telecommunications data, and homeland security analysis. The second main use of stream data mining is for large data set that cannot be stored in memory all at once so the system streams the data and performs a single pass mining of that data.

Stream data mining is not a technique in and of itself but rather a function that is combined with other data mining techniques to allow for quicker results and dealing with streams of data. Types of data mining techniques that are combined with stream mining include: clustering, classification, frequent pattern mining, sequential pattern mining, mining outliers, and mining unusual patterns. Zhang, Liu, and Wang state "stream data are different from persistent data in that (a) they are transient, (b) usually they can only be read once, and (c) any systems working on them have no control over the order in which data streams arrive. Mining such data requires a new breed of systems operating continuously and indefinitely, and incorporating examples as they arrive. Ideally, such systems, even operating under stringent memory and time constraints, should be able to perform a variety of data mining tasks (e.g., clustering and classification of data streams, mining frequent and sequential patterns, mining partial periodicity, mining notable gradients, and mining outliers and unusual patterns)" [5].

Because of these reasons doing data intense mining is difficult and methods must be used to trim data sets to a usable size without affecting the results of the data. The article points to several methods of controlling data sets that are being researched they include reduction based on low-memory factored representation, scalable supervised algorithms for attribute reduction, and false negative using error parameters. Due to the reliance on limited resources of time and memory, stream data mining has to use unique methods to overcome the limitations. The possible solutions are: scalable attribute reduction; data set reduction using a calculation to determine if instance is relevant to the whole set; pruning instances using error parameters; or placing a controlled amount of data into what is known as a sliding window.

The sliding window technique which we use in our model consist of storing data from a specific length of time or to a specific volume in a basic window, which is one unit of a sliding window. Then when a new basic window is full it is added to the sliding window, while the oldest of the basic windows is discarded. At this point the data mining technique being performed on the streaming data is performed on all of the information actively in the sliding window. The sliding window provides a way to control or limit the amount of information in the system at any given time and provides a way to scale to the systems capabilities. One of the key strengths of the sliding window method is that detailed analysis can be performed on current data, which in most of the times, is more relevant/precious than older data.

3. INTRODUCED MODEL

Li and Deng introduce a model that uses stream data mining to look for frequent patterns in network traffic to improve network performance [7]. The model they had generated was the starting point for our development; however our approach is different than theirs and focused more on anomaly detection.

Detection of anomalies, particularly those that are a threat, in network traffic is crucial to the protection of your critical IT infrastructure. The problem is time; most techniques require significant periods of time before anomaly detection provides useful information. When data is streaming through your network it has small pieces of information attached to it to direct it to its destination called a header. To determine what traffic is actually an anomaly, you must capture and separate the informational elements of the header to create the attributes for each instance you perform data mining on. For the purposes of network anomaly detection the system will need, from the header file including but not limited to, the source IP address, destination IP address, Differentiated Service Code Point, and protocol, type of information contained, and other related attributes that can be collected to expand the model and create a more complete profile. As with any data mining exercise there are a significant number of methodologies that can be used individually or in combination to provide an effective result, in this paper we will be focusing on creating a methodology that will actually compare the results of both supervised and unsupervised learning with the original model that is created based on the historical data.

Our system as it is illustrated in figure-1 would process the data using a sliding window and run the information through the following techniques. First it will classify each instance as being associated with a particular role type: engineer, executive, administrative, administrator, helpdesk or IT, based on IT roles for accessing information. The system will use the same data set of activity for each group to create clusters that additional data will be added to. These new instances will be either added to existing clusters or be outliers.

The system actually creates an output for each methodology: one is which cluster of activity did the instance fall into; and second how the instance was classified. The decision engine takes these two outputs as inputs as well as the actual role of the individual from the server. If the role, cluster and classification do not match, or the instance does not fall into a cluster; then the system would flag it as an anomaly and provide the system administrator with this information. The administrator would investigate the anomaly to determine if it is an intrusion or misuse of company resources. The administrator also has the option of adding false positives into the training set, with the correct output variable in the model generator, so that as the models used for classifying and clustering can be updated with a higher accuracy to create less false positives.

Figure 1 is a flow chart that demonstrates how the model would work. One of our greatest concerns is that with the complexity of our model it will not provide an efficient enough result for use with stream data mining. The goal of our model is to increase performance by using multiple methods to determine behavior characteristics and compare them to the actual role.

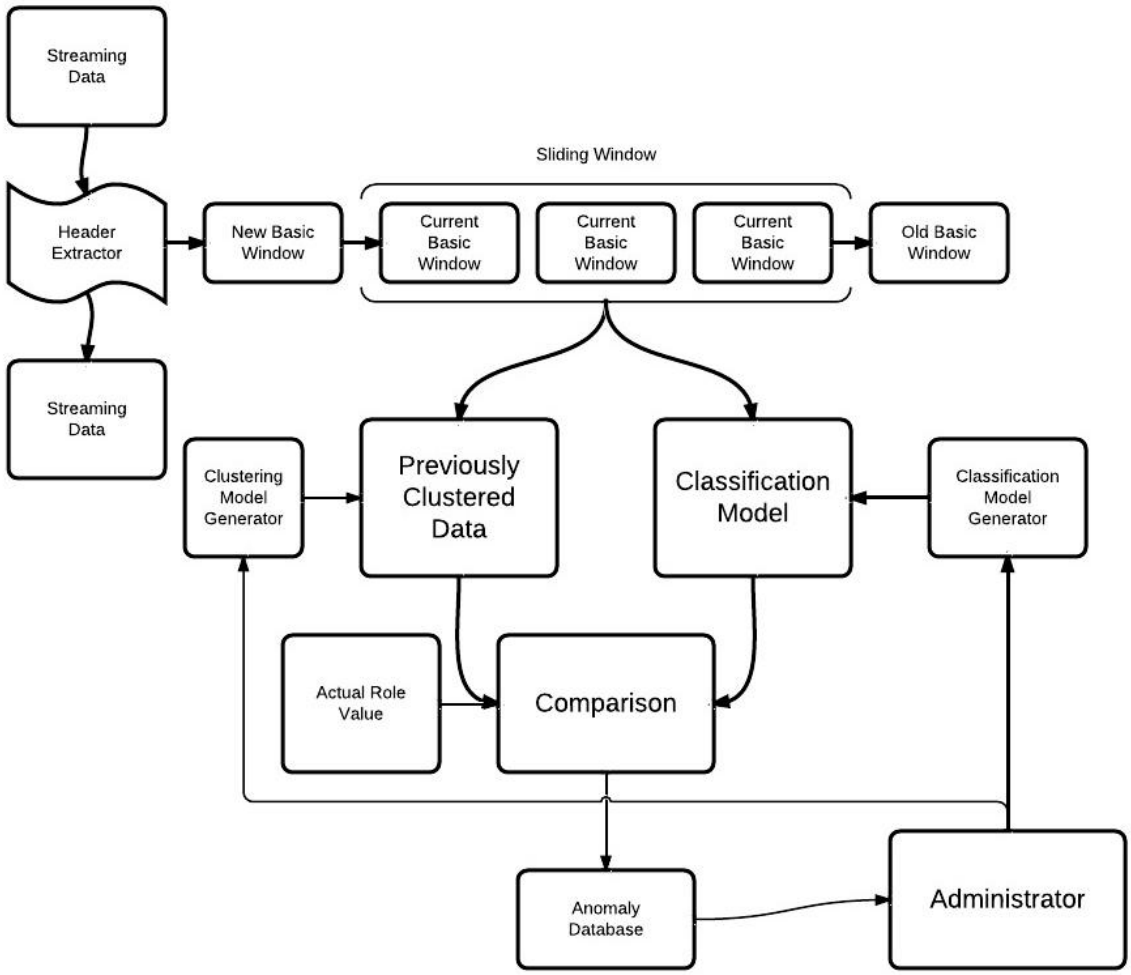

Figure 1: Anomaly Detection Model

Most methods provide for original model generation but do not typically incorporate model correction over time. Our goal of using administrator outputs to add instances to the model generators would allow for model improvement over time by correcting false positives (the legitimate instances, which are labeled as anomalies). We realize that this could be a time consuming so to reduce the amount of information that the administrator has to look at we have incorporated a database of anomalies which will presort the anomalies based on signatures stored there. In future research we will be looking for opportunities to simplify even further and automate this process completely or as much as possible reducing human involvement and replace him/her by an expert system. To accomplish this, anomalies that don't match signatures would then be run through an expert system that would reproduce the knowledge of the administrator to reclassify instances that are marked incorrectly or provide solutions new rules to handle threatening behavior.

4. CONCLUSION

The primary drawback of using technologies such as IDS systems and data mining to detect anomalies are a high rate of false positive; the model becomes outdated and there is high

demand to maintain the models' integrity; and they have low response rate. This study attempts to overcome some of the disadvantages in the current data mining models, which have been used to detect anomaly behaviors. Moreover this research attempts to introduce a model that utilizes stream data mining to actively monitor network traffic for anomaly detection. The introduced anomaly detection model in this study would provide a high level of accuracy over time by reducing the false positives and continuously updating the anomaly signature databases. Moreover, this model continuously updates the data mining models by updating its anomaly databases. however as we researched and discussed further the time and memory restrictions of stream data mining are too great for us to overcome and continue to use this model. So removing the streaming data segment and using stored log data would still provide our model with the ability to actively update on demand, produce higher quality results based on the comparison of the different methodologies compared to the instances actual role. Our model still has the weaknesses of having to store large amounts of data, require human input for machine training, and would be a complex implementation.

There are still advantages over current systems such as: the ability to update the model, to adjust what is defined as normal, on-demand and not having to wait for a new signature database to be made available. The systems model is based solely on your system providing you the greatest accuracy possible.

5. REFERENCES

[1] Anderson, G. F., Selby, D. A., & Ramsey, M. (2007, May). Insider Attack and Real-time Data Mining of User behavior. IBM Journal of Research and Development, 3(4), 465-475.

[2] Conorich, D. (2004, May). Monitoring Intrusion Detection Systems: From Data to Knowledge. Information Systems Security, 13(2), 19-30.

[3] Gregory, P. (2009). CISSP Guid to Security [Essenstioals] (, pp. 1-512). Course Technology.

[4] Hyun Oh, S., & Suk Lee, W. (2003). An Anomaly Intrusion Detection Method by Clustering Normal User Behavior. Computers & Security, 22(7), 596-612.

[5] Zhang, J., Liu, H., & Wang, P. P. (2006, July 22). Some current issues of streaming data mining. Information Science, 176(14).

[6] Scarfone, K., & Mell, P. (2007, February). Guide to Intrusion Detection and Prevention Systems (IDPS). National Institute of Standards and Technology, 1-127.

[7] Li, X., & Deng, Z. (2010, December). Mining frequent patterns from network flows for monitoring network. Expert Systems with Applications, 37(12).