Review and Discussion on different techniques of Anomaly Detection Based and Recent Work

Vaishali V. Khandagale
SVERI's College of Engineering, Pandharpur
Dist.Solapur State: Maharashtra

Yoginath Kalshetty
SVERI's College of Engineering,Pandharpur
Dist.Solapur State: Maharashtra

Abstract

In different application domains as well as areas of research anomaly detection is one of the well-studied problems. There are many techniques presented by different authors successfully and accurate anomaly detection by different researchers. Methods presented are specific to applications or some other domains of research. Some techniques presented are based on data mining and machine learning domains. The main aim of this paper is to summarize the different types of methods presented for anomaly detection. In this paper we have presented the key components of anomaly detection which will be helpful for researcher to understand the existing techniques of anomaly detection. First we will give the overview of intrusion detection systems, then the key components of anomaly detection system. Later we will discuss the different techniques of anomaly detection. In addition to this review, we will present the most recent technique presented over the anomaly detection in this paper.

1. Introduction

Generally the term anomalies are defined as the data pattern which is not conforming to normal behavior of well-definednotion. Hence anomaly detection is nothing but the problem of finding patterns in data that do not conform to expected behaviour. Anomaly detection can be used in a wide variety of applications such as fraud detection for credit cards, insurance or health care, intrusion detection for cyber - security, military surveillance and many others. So, to make it simple, when working with anomaly detection the goal is to define a region representing normal behaviour and declare any observation in the data which does not belong to this normal region as an anomaly.

Although it may sound easy, anomaly detection have certain challenges which make anomaly detection problem not so easy to solve. Some of the challenges are as follow:

1. Defining a normal region which encompasses every possible normal behaviour is very difficult, since the boundary between normal and anomalous data is often not precise.
2. Malicious applications often make anomalous data appear like normal, thereby making the task of defining normal data more difficult.
3. Availability of labelled data for training/validation purposes is often an issue.
4. Data often contains noise which tends to be similar to actual anomalies andtherefore it is difficult to distinguish and remove.

Anomaly detection combines concepts from many different disciplines such as machine learning, statistics, data mining, and information theoryand apply them to specific problem formulations.

The key part of any anomaly detection technique is the nature of input data. Each instance of data usually consists of one or more attributes. The natures of attributes determine the applicability of anomaly detection techniques.

Anomalies can be classified into three categories:

1. Point anomaly:

If an individual instance of data can be considered as an anomaly with respect to the rest of the data.

2. Contextual anomaly:

If a data instance is considered anomalous inside a specific context, then it can be classified as an anomaly.

3. Collective anomaly:

If a collection of data instances is anomalous with the respect to the entire data set. Individual data instances in this type of anomaly might not be considered anomalous.

An important part in any anomaly detection technique is data labelling which denotes whether some data instance is normal or anomalous. Based on the extent to which labels are available, anomaly detection techniques can operate in one of the following modes:

1. Supervised anomaly detection: These techniques assume the availability of a training data set which has labelled instances for normal and anomalous classes.
2. Semi - supervised anomaly detection: Techniques which operate in this mode assume that the training

data set has labelled instances for only the normal class.
3. Unsupervised anomaly detection: These techniques do not require training data. They assume that normal instances are far more frequent in the test data than the anomalies.
In this paper in section II we will discuss in detail about IDS, in section III we will discuss about different techniques of anomaly detection with their advantages and disadvantages, finally in section V we will present the conclusion over the same.

2. Introduction to IDS

A simple definition: It is the unrelenting active attempts in discovering or detecting the presence of intrusive activities. Intrusion Detection (ID) as it relates to computers and network infrastructure encompasses a far broader scope. It refers to all processes used in discovering unauthorized uses of network or computer devices. This is achieved through specifically designed software with a sole purpose of detecting unusual or abnormal activity. The beginning A USAF paper published in October 1972 written by James P. Anderson outlined the fact the USAF had "become increasingly aware of computer security problems. This problem was felt virtually in every aspect of USAF operations and administration". During that period of time, the USAF had the daunting tasks of providing shared used of their computer systems, which contained various levels of classifications in a need-to know environment with a user base holding various levels of security clearance. Thirty Years ago, this created a grave problem that is still with us today. The problem remains: How to safely secure separate classification domains on the same network without compromising security?
In 1980, James P. Anderson published a study outlining ways to improve computer security auditing and surveillance at customer sites. The original idea behind automated ID is often credited to him for his paper on "How to use accounting audit files to detect unauthorized access". This ID study paved the way as a form of misuse detection for mainframe systems. The first task was to define what threats existed. Before designing IDS, it was necessary to understand the types of threats and attacks that could be mounted against computers systems and how to recognize them in an audit data. In fact, he was probably referring to the need of a risk assessment plan to understand the threat (what the risks are or vulnerabilities, what the attacks might be or the means of penetrations) thus following with the creation of a security policy to protect the systems in place. Between 1984 and 1986, Dorothy Denning and Peter Neumann researched and developed the first model of real-time IDS. This prototype was named the Intrusion Detection Expert System (IDES). This IDES was initially a rule-based expert system trained to detect known malicious activity. This same system has been refined and enhanced to form what is known today as the Next-Generation Intrusion Detection Expert System (NIDES).
The report published by James P. Anderson and the work on the IDES was the start of much of the research on IDS throughout the 1980s and 1990s. During this period, the U.S. government funded most of this research. Projects like Discovery, Haystack, Multicast Intrusion Detection and Alerting System (MIDAS), Network Audit Director and Intrusion Reporter (NADIR) were all developed to detect intrusions. To better understand the terms used within the ID user and research community, some of the most commonly used terms are:
Host-Based: The data from a single host is used to detect signs of intrusion as the packets enters or exits the host.

Network-Based: The data from a network is scrutinized against a database and it flags those who look suspicious. Audit data from one or several hosts may be used as well to detect signs of intrusions.

Anomaly detection model: The IDS has knowledge of normal behavior so it searches for anomalous behavior or deviations from the established baseline. While anomaly detection's most apparent drawback is its high false positive, it does offer detections of unknown intrusions and new exploits.

Misuse detection model: The IDS has knowledge of suspicious behavior and searches activity that violates stated policies. It also means looking for known malicious or unwanted behavior. In fact, its main features are its efficiency and comparably low false alarm rate. In the last few years, the ID field has grown considerably and therefore a large number of IDS have been developed to address specific needs4. The initial ID systems were once anomaly detection tools but today, misuse detection tools dominate the market. With an increasingly growing number of computer systems connected to networks, ID has become a necessity. In the mid 1990s, commercial products surfaced for the masses. [KF05]
Two of the most popular IDS in the mid 1990s were Wheel group's Net ranger and Internet Security Systems' Real Secure. Both of these companies started out with network-based IDS. Wheel group was formed in October 1995 to commercialize a security product initially prototyped by the U.S. Air Force then called Net ranger. This product "scans traffic for "signature of misuse", providing real-time alarm and details of the

furtive attacks that may plague a network".5 In February 1998, Wheel group was acquired by Cisco to eventually become an integral part of Cisco's security architecture.

Internet Security Systems, Inc. (ISS) was founded in April 1994 by Thomas Noonan and Christopher Klaus, after Mr. Klaus invented and released the first version of the Internet Scanner.6 on 9 December 1996, ISS announced the release of a tool to augment network security with real-time attack recognition called Real Secure. On the 19 Aug 1997, they announced the first commercial released of their IDS called Real Secure 1.0 for Windows NT 4.0 a new commercial breakthrough.

Another point to consider is most commercially available systems are knowledge-based, which means matching signatures of known attacks against changes in systems or streams of packets on a network. However, their major weaknesses are, they are often helpless against new attacks, so they must be continually updated with new knowledge for new attacks signatures. Despite the fact these false positives are common with behavior based IDS, so is its ability to detect a previously unreported attack.

To help solve the knowledge-based problems, workshops have been held every year for the past four years to share information related to ID. The research topics are quite varied every year and they cover a wide range of subjects such as Lesson Learned, IDS and Law, Modeling Attacks, Anomaly Detection, etc. These workshops main objective are to find new solutions to new and challenging problems. The problems, the research community is now facing are high-speed networks and switching.

Today, more vendors are advertising they can process at gigabit speed. To name a few, Internet Security Systems (ISS), Network ICE, and Intrusion.com advertise they can analyze and alert on gigabit traffic. As networks expand and get faster, network IDS may lose popularity. To address this problem, vendors have turned to the host. How can the host be part of the equation and provide data when it is directly probed for information? The solution was to install host based IDS. The advantages of this type of ID are: analysis of audit or data log, real-time and distributed processing. There are many forms such as host-based ID, TCP Wrappers, Tripwire, and a free tool such as Snort. An intrusion detection system (IDS) monitors network traffic and monitors for suspicious activity and alerts the system or network administrator. In some cases the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user or source IP address from accessing the network. IDS come in a variety of "flavors" and approach the goal of detecting suspicious traffic in different ways. There are network based (NIDS) and host based (HIDS) intrusion detection systems. There are IDS that detect based on looking for specific signatures of known threats- similar to the way antivirus software typically detects and protects against malware- and there are IDS that detect based on comparing traffic patterns against a baseline and looking for anomalies. There are IDS that simply monitor and alert and there are IDS that perform an action or actions in response to a detected threat. We'll cover each of these briefly.

NIDS: Network Intrusion Detection Systems are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network. Ideally you would scan all inbound and outbound traffic; however doing so might create a bottleneck that would impair the overall speed of the network.

HIDS: Host Intrusion Detection Systems are run on individual hosts or devices on the network. A HIDS monitors the inbound and outbound packets from the device only and will alert the user or administrator of suspicious activity is detected.

Signature Based: A signature based IDS will monitor packets on the network and compare them against a database of signatures or attributes from known malicious threats. This is similar to the way most antivirus software detects malware. The issue is that there will be a lag between a new threat being discovered in the wild and the signature for detecting that threat being applied to your IDS. During that lag time your IDS would be unable to detect the new threat.

Anomaly Based: IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is "normal" for that network- what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and alert the administrator or user when traffic is detected which is anomalous, or significantly different, than the baseline.

3. Different Methods of Anomaly Detection

Following figure 1 showing the key components for anomaly detection based on which we categories the different types of anomaly detection techniques:

3.1 Machine Learning Based Techniques

In this technique a model is learnt from a set of labelled data instances. The learnt model can then be used to classify the given test instance into one of the classes. These techniques can be grouped into two

categories: *multi-class* and *one-class* anomaly detection techniques.

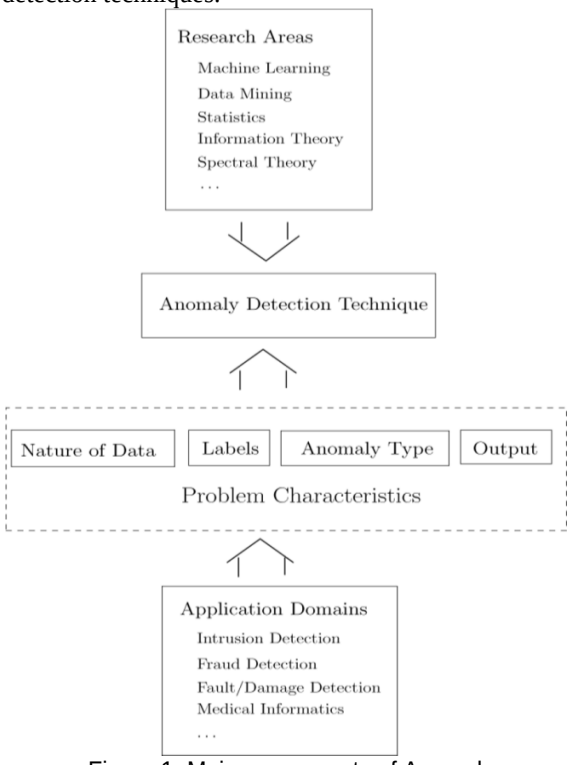

Figure 1: Main components of Anomaly Detection System

Multi-class classification based techniques assume that the training data contains labelled instances belonging to multiple normal classes, while the one-class techniques assume that all normal instances belong to one normal class. Some of the most popular classification based algorithms are: *Neural network based, Bayesian network based, Support vector machine based* and *Rule based.*

Advantages: the testing phase is fast since each test instance needs to be compared against the pre-computed model.

Disadvantage: these techniques can only label test instances, they cannot give meaningful anomaly score.

3.2 Nearest Neighbour Based Techniques

Nearest neighbour based anomaly detection techniques require a distance or similarity measure between two data instances. The key assumption in this technique is that normal instances occur in dense neighbourhoods, while anomalies occur far from their closest neighbours. These techniques can be grouped into two categories:

1. Using distance to k-th nearest neighbour: The anomaly score is defined as its distance to its k-th nearest neighbour in a given data set.

2. Using relative density: Density based techniques estimate the density of the neighbourhood of each data instance. An instance that lies in a neighbourhood with low density is declared to be anomalous, while an instance that lies in a dense neighbourhood is declared to be normal.

Advantages: These methods are unsupervised in nature. They are purely data driven. Also these techniques can be used on different data types, one only need to define an appropriate distance measure for the given data.

Disadvantages: if the data has normal instances that do not have enough close neighbours, the technique fails to label them correctly. Also, the computational complexity of the testing phase is a significant challenge since it involves computing the distance of each test instance with all instances belonging to either the test data or to the training data.

3.3 Clustering Based Techniques

Clustering anomaly detection technique is an unsupervised technique used to group similar data instances into clusters. Clustering techniques can be grouped into following three categories:

1. Normal data instances belong to a cluster, while anomalous data instances don't belong to any cluster. Some of the most popular algorithms in this category are *DBSCAN, ROCK* and *SNN clustering.*

2. Normal data instances lie close to their closest cluster centroid, while anomalous data instances lie far from their closest cluster centroid. An interesting algorithm from this category is *SOM* (*Self organizing maps*). This algorithm is widely used in intrusion detection systems.

3. Normal data instances belong to large and dense clusters, while anomalous data instances belong to small and or sparse clusters. A notable algorithm in this category is *CBLOF* (*Cluster Based Local Outlier Factor*). This algorithm captures the size of the cluster to which the data instance belongs, as well as the distance of the data instance to its cluster centroid.

Advantages: They can operate in an unsupervised mode. Also, the testing phases in these techniques are fast because each data instances needs to be compared to a small number of clusters.

Disadvantages: they are computationally very complex and they are only effective if the anomalies don't form significant clusters among themselves.

3.4 Statistical Techniques

The main assumption in statistical anomaly detection techniques is that normal data instances occur in high probability regions of a stochastic model, while anomalies occur in low probability regions of a stochastic model. Statistical anomaly detection techniques can be split into two categories: *parametric* and *nonparametric* techniques.

Parametric techniques assume that the normal data instances are generated by a parametric distribution such as a Gaussian distribution. The parameters are

estimated using *Maximum Likelihood Estimates (MLE).* The anomaly score of a data instance is the distance of that data instance to the estimated mean.
Apart Gaussian distribution based, other parametric techniques are *Regression model based* and *Mixture of parametric distributions based.* Non-parametric techniques use nonparametric statistical models. In these techniques the model structure is determined from the given data. Also, those techniques make fewer assumptions regarding data when compared to parametric techniques. One of the most popular non-parametric statistical based anomaly detection techniques is to use histograms to maintain a profile of the normal data. The histogram based techniques are widely used among intrusion detection developers, which makes them very interesting to me based on my area of research.
Advantage: if the assumptions regarding the data distribution hold true, statistical techniques provide a statistically justifiable solution for anomaly detection.
Disadvantage: they highly rely on an assumption that the data is generated from a particular distribution. This assumption is usually incorrect. Anomaly detection is one very large and complex area of research and the goal of this post were to introduce the reader to some basic parts of it. There are several directions for further research in anomaly detection, where the most interesting ones are in contextual and collective anomaly detection techniques, which are beginning to find increasing applicability in several domains.

4.Recent Work

Recently in [1], authors presented an online anomaly detection method based on over-sample PCA. Authors successfully proved that the osPCA with LOO strategy will amplify the effect of outliers, and thus we can successfully use the variation of the dominant principal direction to identify the presence of rare but abnormal data [1]. When When oversampling a data instance, our proposed online updating technique enables the osPCA to efficiently update the principal direction without solving eigenvalue decomposition problems [1]. We have identified few limitations of this paper which we can further think for improvement to this paper.

5.Conclusion and Future Work

In this paper we have presented the generalized definition of anomaly detection, its different methods, aspects and techniques of anomaly detection. During this paper we have studied the different types of intrusion detection systems with the brief introduction of each category of anomaly detection methods along with their advantages and disadvantages. In addition to this we have presented the details for recently presented technique which is based on over sample PCA for the online anomaly detection. For the future work we will suggest to present the investigation over the same technique and claims its efficiency against existing methods.

6. References

[1] Yuh-Jye Lee, Yi-RenYeh, and Yu-Chiang Frank Wang, "Anomaly Detection via Online OverSampling Principal Component Analysis", JOURNAL OF IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,
2012.

[2] S. Agarwala and et. al. E2EProf: Automated Endto-End Performance Management for Enterprise Systems, DSN, 2007.

[3] S. Agarwala and K. Schwan, SysProf: Online Distributed Behavior Diagnosis through Fine-grain System Monitoring, ICDCS, 2006.

[4] Ira Cohen and et. al. Correlating Instrumentation Data to System States: A Building Block for Automated Diagnosis and Control, OSDI, 2004.

[5] Bahl, Paramvir and et. al. "Towards highly reliable enterprise network services via inference of multi-level dependencies, SIGCOMM, 2007.

[6] Chen, Mike Y. and et. al. "Pinpoint: Problem Determination in Large, Dynamic Internet Services, DSN, 2002

[7] Song, X., Wu, M., Jermaine, C., and Ranka, S. 2007. Conditional anomaly detection. IEEE Transactions on Knowledge and Data Engineering 19, 5, 631-645.

[8] Theiler, J. and Cai, D. M. 2003. Resampling approach for anomaly detection in multispectral images. In Proceedings of SPIE 5093, 230-240, Ed.

[9] Phoha, V. V. 2002. The Springer Internet Security Dictionary. Springer-Verlag.

[10] Wong, W.-K., Moore, A., Cooper, G., and Wagner, M. 2003. Bayesian network anomaly pattern detection for disease outbreaks. In Proceedings of the 20th International Conference on Machine Learning.
AAAI Press, Menlo Park, California, 808-815.