Unsupervised Learning for anomaly Intrusion detection

Seminar 1 - Report

Mohamed EL Fadly
5-17-2015
Table of Contents

Introduction .................................................................................................................................................. 2
Motivation .................................................................................................................................................... 2
Intrusion detection for cyber-security ........................................................................................................ 2
Machine Learning ........................................................................................................................................ 4
Machine Learning in Anomaly detection Systems ....................................................................................... 4
Problem definition ....................................................................................................................................... 5
Objective .................................................................................................................................................... 5
Unsupervised Learning ............................................................................................................................... 6
Challenges .................................................................................................................................................. 6
Proposed Approach .................................................................................................................................... 6
  Nature of input data: ................................................................................................................................ 7
  Availability of supervision ....................................................................................................................... 8
  Type of anomaly: point, contextual, structural .......................................................................................... 8
Output of anomaly detection ....................................................................................................................... 10
Evaluation of anomaly detection techniques ................................................................................................ 10
Clustering based techniques ....................................................................................................................... 12
Related Work ............................................................................................................................................. 12
  A Near Real-Time Algorithm for Autonomous Identification and Characterization of Honeypot Attacks .......................................................................................................................... 12
  Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection ......................... 14
  An Unsupervised Anomaly Detection Engine with an Efficient Feature set for AODV ......................... 15
Conclusion ................................................................................................................................................ 16
References .................................................................................................................................................. 17
Introduction

Anomaly detection is an important problem that has been researched within diverse research areas and application domains. Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains. Of these, anomalies and outliers are two terms used most commonly in the context of anomaly detection; sometimes interchangeably. Many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic [11].

Motivation

Anomaly detection finds extensive use in a wide variety of applications such as fraud detection for credit cards, insurance, or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities. The importance of anomaly detection is due to the fact that anomalies in data translate to significant, and often critical, actionable information in a wide variety of application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination or anomalies in credit card transaction data could indicate credit card or identity theft. One of the crucial domain where anomaly detection is heavily used is Intrusion detection for cyber-security.

Intrusion detection for cyber-security

An Intrusion Detection system (IDS) intelligently monitors activities that occur in a computing resource, e.g., network traffic and computer usage, to analyze the events and to generate reactions. In IDSs, it is always assumed that an intrusion will manifest itself in a trace of these events, and the trace of an intrusion is different from traces left by normal behaviors. To achieve this purpose, network packets are collected, and the rule violation is checked with pattern recognition methods. An IDS system usually monitors and analyzes user and system
activities, accesses the integrity of the system and data, recognizes malicious activity patterns, generates reactions to intrusions, and reports the outcome of detection.

The activities that the IDSs trace can form a variety of patterns or come from a variety of sources. According to the detection principles, the intrusion detection can be classified into the following modules [1]:

1. **Misuse/Signature detection**,  
2. **Anomaly Detection**,  
3. **Hybrid Detection**,  
4. **Scan detector and profiling modules**.

1. **Misuse/Signature Detection**: is an IDS triggering method that generates alarms when a known cyber misuse occurs. A signature detection technique measures the similarity between input events and the signatures of known intrusions. It flags behavior that shares similarities with a predefined pattern of intrusion. Thus, known attacks can be detected immediately and realizable with a lower false-positive rate [1].

2. **Anomaly Detection**: Anomaly detection triggers alarms when the detected object behaves significantly differently from the predefined normal patterns. Hence, anomaly detection techniques are designed to detect patterns that deviate from an expected normal model built for the data [1].

3. **Hybrid Detection**: Combining both anomaly and misuse detection techniques to overcome their drawbacks. Since both of these methods have drawbacks: misuse detection techniques lack the ability to detect unknown intrusions, while anomaly detection techniques usually produce a high percentage of false alarms [1].

4. **Scan Detection and Profiling Module**: Scan detection generates alerts when attackers scan services or computer components in network systems before launching attacks. The Profiling modules group similar network connections and search for dominant behaviors using clustering algorithms [1].

However, Traditional signature-based intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats. Moreover, there is a substantial latency in deployment of newly created signatures across the computer system. That’s why anomaly detection can alleviate these limitations.
Most of the anomaly detection solutions depend heavily on Machine learning approach to find solutions to cybersecurity problems.

**Machine Learning**

In 1959, Arthur Samuel defined machine learning as the “Field of study that gives computers the ability to learn without being explicitly programmed.”[5] Yet, Tom M. Mitchell provided a widely quoted, more formal definition for machine learning stating that: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E” [6]. Machine learning is one of the cornerstone fields in Artificial Intelligence, where machines learn to act autonomously, and react to new situations without being pre-programmed. It is about designing algorithms that allow computers to learn.

Machine learning algorithms are categorized, based on the desired outcome of the algorithm; those algorithms include [8]:

**Supervised learning:** The machine is trained with labeled data, where the algorithm generates a function that maps inputs to desired outputs. Although this method is widely used, obtaining labeled data is always difficult and expensive. Popular categorizations include artificial neural network (ANN), support vector machine (SVM), and decision trees.

In **unsupervised learning**, no target or label is given in sample data. Unsupervised learning methods are designed to summarize the key features of the data and to form the natural clusters of input patterns given a particular cost function. The most famous unsupervised learning methods include $k$-means clustering, hierarchical clustering, and self-organization map. Unsupervised learning is difficult to evaluate, because it does not have an explicit teacher and, thus, does not have labeled data for testing.

**Machine Learning in Anomaly detection Systems**

A typical anomaly detection system consists of the following steps [1].

1- **Data collection:** the volume of data is extremely large, and it requires data reduction in data preprocessing. Also, most of the data in the network are streaming data, and requires another step of data reduction.
2- **Data preprocessing**: including feature selection, feature extraction, or a dimensionality reduction technique, and an information-theoretic method.

3- **Building normal profiles**: where machine learning plays an important part in this step

4- **Anomaly detection**: labeled data corresponding to normal behavior are usually available, while labeled data for anomaly behavior are not. Supervised machine-learning methods need attack-free training data. However, this kind of training data is difficult to obtain in real-world network environments. This lack of training data leads to the well-known unbalanced data distribution in machine learning.

**Problem definition**

In anomaly detection, labeled data corresponding to normal behavior are usually available, while labeled data for anomaly behavior are not. In addition, supervised machine-learning methods need attack-free training data. However, this kind of training data is difficult to obtain in real-world network environments. This lack of training data leads to the well-known unbalanced data distribution in machine learning. In the huge volume of network data, the same malicious data repeatedly occur while the number of similar malicious data is much smaller than the number of normal data. The imbalanced data distribution of normal and anomaly data induces a high false-positive rates (FPRs) of supervised intrusion detection systems (IDSs). Unsupervised machine learning methods outperform supervised machine-learning methods in updating rules intelligently while the detection rates downgrade. Thus anomaly detection systems can potentially find new attacks, but they generally have a lower accuracy rate for detection and a higher FAR.

The problem always remains is how to minimize the false negative and false positive rates while keeping higher accuracy rates.

**Objective**

To propose an unsupervised anomaly detection technique that will produce low false positive rates and to overcome challenges in using labeled data sets for supervised learning, such as time consumption, expensiveness, limitation of expertise, and the accuracy of labels in collecting labeled data.
Unsupervised Learning

Unsupervised learning is used to overcome the problem of using attack-free training data required by supervised learning. Moreover, with the changing network environment or services, patterns of normal traffic will change. The differences between the training and actual data can lead to high FPRs of supervised IDSs. To address these problems, unsupervised anomaly detection emerges to take unlabeled data as input.

The main goal of the unsupervised anomaly detection is to find malicious information buried in cyberinfrastructure even without prior knowledge about the data labels and new attacks. Subsequently, unsupervised anomaly detection methods rely on the following assumptions: normal data covers majority while anomaly data are minor in network traffic flow or audit logs.

That’s why most of the solutions to unsupervised anomaly detection are clustering-based anomaly/outlier detection techniques.

Challenges

Accurate detection of these malicious behaviors encounters several challenges.

1. The key challenge is that the huge volume of data with high-dimensional feature space is difficult to manually analyze and monitor. Such analysis and monitoring requires highly efficient computational algorithms in data processing and pattern learning.
2. Much of the data is streaming data, which requires online analysis.
3. It is also difficult to define a representative normal region or the boundary between normal and outlying behavior. As the concept of an anomaly/outlier varies among application domains
4. The labeled anomalies are not available for training/validation.
5. Training and testing data might contain unknown noises

Proposed Approach
Currently there is no well-defined written approach; however, there are aspects I should consider when choosing my approach
A key aspect of any anomaly detection technique is the nature of the input data. Each data instance can be described using a set of attributes (also referred to as variable, characteristic, feature, field, or dimension). The attributes can be of different types such as binary, categorical, or continuous. Each data instance might consist of only one attribute (univariate) or multiple attributes (multivariate). In the case of multivariate data instances, all attributes might be of the same type or might be a mixture of different data types [11]. The nature of attributes determines the applicability of anomaly detection techniques. For example, for statistical techniques different statistical models have to be used for continuous and categorical data. Similarly, for nearest-neighbor-based techniques, the nature of attributes would determine the distance measure to be used.
Availability of supervision

There are three types of supervision:

1- Supervised Anomaly Detection: Labels are available for both normal data and anomalies

2- Unsupervised Anomaly Detection: No labels are assumed; based on the assumption that anomalies are very rare compared to normal data

3- Semi-supervised Anomaly Detection: Labels are available only for normal data and a modified classification model is used to learn the normal behaviour and then detect any deviations from normal behaviour as anomalous

In my research I will be focusing on either fully unsupervised or the semi-supervised

Type of anomaly: point, contextual, structural

1- **Point Anomalies**: If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed a point anomaly. This is the simplest type of anomaly and is the focus of majority of research on anomaly detection. For example, in the below figure, points o1 and o2, as well as points in region O3, lie outside the boundary of the normal regions, and hence are point anomalies since they are different from normal data points.

![Diagram of point anomalies](image)

2- **Contextual Anomalies**: If a data instance is anomalous in a specific context, but not otherwise, then it is termed a contextual anomaly. Contextual anomalies have been
most commonly explored in time-series data and spatial data. The below figure shows one such example for a temperature time-series that shows the monthly temperature of an area over the last few years. A temperature of 35°F might be normal during the winter (at time $t_1$) at that place, but the same value during the summer (at time $t_2$) would be an anomaly.

3- **Collective Anomalies**: If a collection of related data instances is anomalous with respect to the entire data set, it is termed a collective anomaly. The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous. The following figure is an example that shows a human electrocardiogram output. The highlighted region denotes an anomaly because the same low value exists for an abnormally long time.
Output of anomaly detection

An important aspect for any anomaly detection technique is the manner in which the anomalies are reported. Typically, the outputs produced by anomaly detection techniques are one of the following two types:

1. **Scores.** Each instance in the test data is assigned a score depending on the degree to which that instance is considered an anomaly. Thus the output of such techniques is a ranked list of anomalies. An analyst may choose to either analyze the top few anomalies or use a cutoff threshold to select the anomalies.

2. **Labels.** Each test instance is given a *normal* or *anomaly* label.

Evaluation of anomaly detection techniques

Accuracy is not sufficient metric for evaluation for example: network traffic data set with 99.9% of normal data and 0.1% of intrusions. Besides a trivial classifier that labels everything with the normal class can achieve 99.9% accuracy.

Anomaly class => C  
the normal class => NC

<table>
<thead>
<tr>
<th>Confusion matrix</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NC</td>
</tr>
<tr>
<td>Actual class</td>
<td></td>
</tr>
<tr>
<td>NC</td>
<td>TN</td>
</tr>
<tr>
<td>C</td>
<td>FN</td>
</tr>
</tbody>
</table>

Thus I will use **Standard measures for evaluating anomaly detection problems:**

1. **Recall (Detection rate)** - ratio between the number of correctly detected anomalies and the total number of anomalies

2. **False alarm (false positive) rate** – ratio between the number of data records from normal class that are misclassified as anomalies and the total number of data records from normal class

3. **ROC Curve** is a trade-off between detection rate and false alarm rate

4. **Area under the ROC curve (AUC)** is computed using a trapezoid rule
Possible approach

Anomaly Detection

- Classification Based
  - Rule Based
  - Neural Networks Based
  - SVM Based

- Nearest Neighbor Based
  - Distance Based

- Clustering Based
  - Parametric
  - Non-parametric

- Statistical
  - Information Theory Based
  - Spectral Decomposition Based
  - Visualization Based

- Others

Point Anomaly Detection

- Contextual Anomaly Detection

- Collective Anomaly Detection

- Online Anomaly Detection

- Distributed Anomaly Detection
Clustering based techniques

Our Key assumption is that normal data records belong to large and dense clusters, while anomalies belong do not belong to any of the clusters or form very small clusters. Categorization will be according to labels

- Semi-supervised – cluster normal data to create modes of normal behaviour. If a new instance does not belong to any of the clusters or it is not close to any cluster, is anomaly

- Unsupervised – post-processing is needed after a clustering step to determine the size of the clusters and the distance from the clusters is required from the point to be anomaly

Anomalies detected using clustering based methods can be either Data records that do not fit into any cluster (residuals from clustering), small clusters or low density clusters or local anomalies (far from other points within the same cluster)

The advantages for using clustering techniques are:

- No need to be supervised
- Easily adaptable to on-line / incremental mode suitable for anomaly detection from temporal data

However the Drawbacks are:

- Computationally expensive : Using indexing structures (k-d tree, R* tree) may alleviate this problem
- In high dimensional spaces, data is sparse and distances between any two data records may become quite similar. Clustering algorithms may not give any meaningful clusters

Related Work

A Near Real-Time Algorithm for Autonomous Identification and Characterization of Honeypot Attacks

Published in ASIA CCS '15 Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security

Owezarski presents an unsupervised algorithm - called UNADA for Unsupervised Network Anomaly Detection Algorithm - for identification and characterization of security related
anomalies and attacks occurring in honeypots. What is interested in their research that their method does not need any attack signature database, learning phase, or labeled traffic [12].

Their algorithm has several advantages

1. It works in a completely unsupervised manner, what makes it able to work on top of any monitoring system, and directly usable, without preliminary configuration or knowledge.

2. It combines robust clustering techniques to avoid classical issues of clustering algorithms, e.g. sensitivity to initial configuration, the required a priori indication of the number of clusters to be identified, or the sensitivity of results when using less pertinent features.

3. It automatically builds simple and small signatures fully characterizing attacks; theses signature can then be used in a filtering security device.

4. It is designed to run in real time by making possible to take advantage of the parallelism of their clustering approach.

**Evaluation**

They run their algorithm on the honeypot traffic traces gathered at the University of Maryland and they compare the performance of UNADA against three previous approaches for unsupervised anomaly detection: (DBSCAN-based, k-means-based, and PCA-based outlier’s detection).
Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection

published in Proceeding ODD '13 Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description

Amer et.al have applied two modifications in order to make one-class SVMs more suitable for unsupervised anomaly detection: Robust one class SVMs and eta one-class SVMs [13]. The key idea of both modifications is that outliers should contribute less to the decision boundary as normal instances. Experiments performed on datasets from UCI machine learning repository show that their modifications are very promising: Comparing with other standard unsupervised anomaly detection algorithms, the enhanced one-class SVMs are superior on two out of four datasets. In particular, the proposed eta one class SVM has shown the most promising results.

Experiments and results

The experiments showed that the proposed SVM based algorithms are well suited for the unsupervised anomaly detection problem. In two out of four datasets, SVM based algorithms are even superior. They constantly outperform all clustering based algorithms. In general, they perform at least average on unsupervised anomaly detection problems.

When comparing the SVM based algorithms with each other, the eta one-class SVM seems to be the most promising one. On average, it produces a sparse solution and it also performs best in terms of AUC. In general, the robust one-class SVM produces a sparser solution than the standard one-class SVM, but in term of performance, there is no significant improvement. In terms of time efficiency, for larger datasets the enhanced algorithms are more efficient due to the sparsity property. As for the computational effort, SVM based algorithms have in general less than a quadratic time complexity due to the sparsity property.

In addition, they introduced a method for calculating an outlier score based on the distance to the decision boundary. In contrast to the binary label assigned by standard one class SVMs, it allows to rank the outliers, which is often essential in an unsupervised anomaly detection setup [13].
An Unsupervised Anomaly Detection Engine with an Efficient Feature set for AODV

Published in: Information Security and Cryptology (ISCISC), 2013 10th International ISC Conference. They proposed an anomaly detection engine to detect malicious attacks via intrusion detection systems by collecting decent features and applying robust PCA on the data set [14]. They have collected a feature set from some state-of-the-art works in the literature. Their simulation shows that the feature set detect anomaly behavior more accurate. In addition, they used robust PCA for analyzing the data set instead of PCA. Robust PCA means using an unsupervised algorithm versus semi-supervised provided by PCA. In contrast to PCA, their results show robust PCA cannot be affected by outlier data within the network. They simulated normal and attack states and the results were analyzed.

The results showed their features can detect much more attacks either by applying PCA or by applying robust PCA. Their contribution was in using the Robust PCA, through having an unsupervised algorithm that detect anomaly more accurate. The robust PCA can form the baseline profile even by existence of malicious nodes in the learning phase. Their results show that robust PCA cannot be affected by outlier data within the network.
Conclusion

Machine-learning methods play key roles in building normal profiles and intrusion detection in anomaly detection systems. In anomaly detection, labeled data corresponding to normal behavior are usually available, while labeled data for anomaly behavior are not. Supervised machine-learning methods need attack-free training data. However, this kind of training data is difficult to obtain in real-world network environments. This leads to the well-known unbalanced data distribution in machine learning. Moreover, the differences between training and actual (test) data lead to high false-positive rates (FPRs) of supervised intrusion detection systems (IDSs). Unsupervised anomaly detection can overcome the drawbacks of supervised anomaly detection. Thus, semi-supervised and unsupervised machine-learning methods are employed.

Because these methods flag any significant deviation from the baseline as an intrusion, it is likely that nonintrusive behavior that falls outside the normal range will also be labeled as an intrusion, resulting in a false positive. Another disadvantage of anomaly detection approaches is that hackers often modify malicious codes or data to make them similar to normal patterns. The problem I must solve is how to minimize the false negative and false positive rates and to determine what the effects are if I use the unsupervised or semi-supervised methods.
References

1- Sumeet Dua and Xian Du, Data Mining and Machine Learning in cybersecurity. Auerbach Publications, April 25, 2011


