Applying Outlier Detection Techniques in Anomaly-based Network Intrusion Systems – A Theoretical Analysis

J. Rene Beulah
Research Scholar
Government College of Engineering, Tirunelveli

D. Shalini Punithavathani, Ph.D
Principal
Government College of Engineering, Tirunelveli

ABSTRACT

With the advent of the Internet, security has become a major concern. An intrusion detection system is used to enhance the security of networks by inspecting all inbound and outbound network activities and by identifying suspicious patterns as possible intrusions. For the past two decades, many researchers are working in Intrusion Detection Systems. In recent years, anomaly detection has gained popularity with its ability to detect novel attacks. Nowadays researchers focus on applying outlier detection techniques for anomaly detection because of its promising results in identifying true attacks and in reducing false alarm rate. In this paper, some of the works which applied outlier analysis in anomaly detection is studied and their results are analyzed.

General Terms

Network Security

Keywords

Outlier detection, anomaly detection, intrusion detection

1. INTRODUCTION

Intrusion detection has been at the center of intense research in the past decade owing to the rapid increase of sophisticated attacks on computer systems [1]. Intrusion is a set of actions aimed to compromise computer security goals such as confidentiality, integrity and availability of resources. Intrusion Detection System (IDS) is a device or a software application capable of monitoring different activities in a network and analyze them for signs of security threats. The goal of an IDS is to discover breaches of security, attempted breaches or open vulnerabilities that could lead to potential breaches.

Based on working methodology, Intrusion Detection Systems can be classified as Misuse or Signature Detection and Anomaly Detection. Methods using Misuse Detection maintain a database of signatures of previously known attacks and the data being analyzed is compared with the database. If the signatures are matched, an intrusion is identified. The advantage of such misuse detection methods is that false alarm rate is very low. The main drawback is that these methods can detect only attacks for which they are previously trained and they fail to detect new attacks or even variants of known attacks. On the other hand, Anomaly Detection techniques build a model for the normal network behavior and any abnormal behavior that deviates from this model is identified as an intrusion. The key advantage of such techniques is high detection rate because of the ability to detect previously unseen intrusions. But the rate of false positives is high. Since the first introduction of anomaly-based intrusion detection to the research community in 1987, the field has grown tremendously [1]. Anomaly-based Network Intrusion Detection Systems is currently a prime focus of research in the field of Intrusion Detection and a critical issue is to reduce false alarms.

A number of statistical based, knowledge based and machine learning based techniques have been used in the field of Anomaly based Network Intrusion Detection Systems. Outlier detection is one of the most successful anomaly detection approaches in intrusion detection. Outlier mining is more suitable to fulfill the task of anomaly intrusion detection [2]. In addition, compared with other approaches of anomaly detection, the anomaly detection approach based on outlier mining does not need training process, successfully overcoming the high frequency of false reports existing on other anomaly detection approaches [2]. An outlier detection technique is effective in reducing the false positive rate with a desirable and correct detection rate [3]. An outlier is a data point which is very different from the rest of the data based on some measure. Approaches for outlier detection can be categorized as statistical-based, distance-based, density-based, clustering-based and frequent pattern-based. Outlier detection is of interest in many practical applications like network intrusion detection, credit-card fraud detection, mobile phone fraud detection, insurance claim fraud detection, insider trading detection and image processing. Recently, outlier detection is finding its prominent place in the literature of Anomaly based Intrusion Detection Systems.

The rest of this paper is organized as follows. Section 2 discusses different outlier detection techniques used in Anomaly based Intrusion Detection Systems. Section 3 presents the performance of those techniques. Section 4 discusses the research issues and challenges while Section 5 gives concluding remarks. Section 6 lists the references.

2. OUTLIER DETECTION TECHNIQUES USED IN ANOMALY BASED IDS

The task of outlier detection is to find the small groups of data objects that are exceptional to the inherent behavior of the rest of the data. Outlier detection can be used to isolate suspicious or interesting patterns in the data [4]. The output of an outlier detection may be either a score about the level of outlierness of a data point or a binary label indicating whether a data point is an outlier or not. The anomaly detection problem is similar to the problem of finding outliers, specifically, in network intrusion detection. The key challenge for outlier detection in this domain is the huge volume of data. Outlier detection schemes need to be computationally efficient to handle these large-sized inputs [5]. Some of the outlier detection techniques used in Anomaly based IDS are discussed in this section.

2.1 Using Outlier Detection to Reduce False Positives

Reducing false positives in anomaly based techniques needs more attention of researchers. A novel method for reducing false alerts is proposed in [6] by Fu Xiao and Xie Li. The authors claim that outlier detection has been successfully applied in many fields but has not been used to reduce false positives so far and they are the first to introduce this technique into this field. In order to filter IDS alerts better, the authors have designed an improved frequent pattern-based outlier detection algorithm. An outlier score is assigned to each alert, which is calculated based on how many frequent attribute values the alert contains. If an alert has more frequent patterns, its score is high and it is more likely a false positive. In order to filter alerts in real time, they have designed a two-phrase framework. In the learning phrase, the feature set of false positives is built and based on this set, a threshold of true alerts is calculated. In the online filtering phrase, the outlier score of each new alert is compared with this threshold to determine whether it is false positive or not. Moreover the feature set is automatically updated in order to keep its accuracy.

2.2 Cost Distribution Based Outlier Detection

Cost Distribution-based outlier detection algorithm is proposed in [7] by Komsit Prakobphol and Justin Zhan. The notations used are given in Table 1.

Table 1. Notations used to define CDOF [7]

Notation	Denotation
d	Data point
N_k	Set of k-nearest neighbors of *d*
n	Number of k-nearest neighbors of *d*
o	Object in N_k
D	Data set consists of d_1 and N_k $D=(d_1, d_2, \ldots, d_n)$
SCD	Shortest connected path cost description SCD=$(c_1, c_2, \ldots, c_n)$
SWC	Average square weighted cost distribution
CDOF	Cost-distribution based outlier detection factor

SCD of d_1 is a sequence (c_1, c_2, …, c_n) on sequence D=(d_1,d_2, …, d_n) such that for each d_i, 0<i<n, c_i is the shortest distance from any data point d in {d_1, d_2, …, d_i} to d_{i+1}.

SWC of d1 is the weighted square sum of SCD of d1. SWC is defined as

$$SWC(d_1) = \frac{2}{n(n+1)\sum_{i=1}^{n} c_i} \sum_{i=1}^{n} (n+1-i)c_i$$

CDOF of d1 with respect to its k-neighbors is defined as

$$CDOF(d_1) = \frac{n(SWC(d_1))}{\sum_{o \in N_k} SWC(o)}$$

CDOF of d_1 is the ratio of SWC form d_1 to D and SWC of d_1's k-distance neighbors to their own k-distance neighbors. It indicates the degree of being an outlier. Points that lie deep in the cluster have CDOF close to 1 where outliers have higher CDOF. Outliers are points whose CDOF values exceed a given threshold.

2.3 Random-Forests-Based Outlier Detection

Random forests based Network IDS is proposed in [8] by Jiong Zhang, Mohammad Zulkernine and Anwar Haque. In their paper, a data mining algorithm called random forests was applied in misuse, anomaly and hybrid-network-based IDS. In the anomaly detection method, the IDS captures the network traffic and constructs dataset by preprocessing. Service-based patterns are built over the dataset using the random forests algorithm. With the built patterns, the outliers related to each pattern can be found. Two types of outliers are detected. The first type is an activity that deviates significantly from the others in the same network service. The second type is an activity whose pattern belongs to the services other than their own service. For instance, if a http activity is classified as a ftp service, the activity will be determined as an outlier. The proximity is one of the most useful tools in the random forests algorithm. After the forest is constructed, all cases in the dataset are put down each tree in the forest. If cases k and n are in the same leaf of a tree, their proximity is increased by one. Finally, the proximities are normalized by dividing by the number of the trees. If the outlier-ness of a case is large and the proximity is small, the case is determined as an outlier.

2.4 Density Based Outlier Mining

In general, density based outlier detecting methods estimate the density of the neighborhood of each data point. An outlier is a point that lies in a neighborhood with low density. A new density based outlier mining algorithm which reduces computational time is proposed in [9] by Peng Yang and Biao Huang. D is a dataset, C is the core object and P, Q are points in the dataset. M(P) is the module of P which is defined as $M(P) = \sqrt{\sum_{i=1}^{d} P_i^2}$. Two lemmas are proposed. Lemma 1 says that "If a data object is a core object, all data objects within the ε-neighborhood of it are not outliers". It is inferred from Lemma 1 that for every object in the dataset, we need not judge whether there are core objects within the ε-neighborhood of it. If P is a core object, it and all data in its ε-neighborhood are marked non-outliers. Otherwise, find whether some core objects exist in ε-neighborhood of P. If so, then P is in the ε-neighborhood of some core object and thus P is not an outlier. Lemma 2 says that "If d(P,Q) denotes the Euclidean distance between P and Q, then d(P,Q)≥|M(P)-M(Q)|. From Lemma 2, it is inferred that there is no need to calculate the Euclidean distance between every two points. Instead, for a point Q, |M(P)-M(Q)| is calculated while checking the ε-neighborhood of point P and if |M(P)-M(Q)|> ε, then d(P,Q)> ε. Thus Q is not in the ε-neighborhood of P. Thus applying these two lemmas greatly reduces the number of computations. In addition, the algorithm proposed by them can effectively minimize the number of seeds to identify outliers.

2.5 Outlier Ensemble Detection

An intrusion detection method based on outlier ensemble detection is proposed in [2] by Bin HUANG, Wen-fang LI and De-li CHEN. The authors have used KDD data set for

Intrusion Detection. There are both numerical and character properties in the data set. So Code mapping is used to convert character property to a new property with Boolean values. If the character property has a large number of values, then the bits needed to complete coding is also lengthy. So Principal Component Analysis (PCA) is used to reduce dimension which results in a totally new data set which can represent the whole information of the initial data set. After this preliminary processing of data, an outlier mining algorithm based on Similar Coefficient Sum is applied to find the anomaly set. Then the outlier mining algorithm based on Kernel Density is executed to find the anomaly set. These two outlier mining algorithms are repeated M times and an ensemble of the results will give the final anomaly set according to Voting Mechanism. The authors conclude that the ensemble approach can give better results than applying single outlier mining algorithm.

2.6 Weighted Distance Based Outlier Detection

A new outlier detection algorithm called Weighted Distance Based Outlier Detection (WDBOD) algorithm to detect the intruders in wireless environment is proposed in [10] by S.Ganapathy, N.Jaisankar, P.Yogesh and A.Kannan. The algorithm has two phases: Training and Testing. In training phase, first, the weighted average distance is calculated. Then the number of nodes whose distance is greater than the weighted average is computed. The inner weighted average for the k-nearest inner nodes is computed. Finally the data is trained for inner and outer nodes. In the testing phase, weighted distance of the new arriving node is computed. If the distance of the newly arrived node is greater than the weighted average distance then the node is classified as abnormal else normal. The authors claim that this weighted distance based outlier detection algorithm detects unexpected entries in databases and networks with high accuracy and low false alarm rate.

2.7 Sub-Space Outliers Ranking

An Unsupervised Network Anomaly Detection Algorithm (UNADA) for knowledge-independent detection of anomalous traffic is proposed in [11] by Pedro Casas, Johan Mazel and Philippe Owezarski. UNADA relies on robust clustering algorithms to detect outlying traffic flows. UNADA runs in three consecutive steps:

Step 1: Detecting an anomalous time slot

1. Captured packets are first aggregated into multi-resolution traffic flows.
2. Different time-series are then built on top of these flows
3. Any generic change-detection algorithm based on time-series analysis is used to flag an anomalous change.

Step 2: Building outliers ranking

1. Input: All the flows in the time slot flagged as anomalous
2. Outlying flows are identified using a robust multi-clustering algorithm, based on a combination of Sub-Space Clustering, Density-based Clustering and Evidence Accumulation Clustering techniques.
3. Build outliers ranking: The evidence of traffic structure provided by this clustering algorithm is used to rank the degree of abnormality of all the identified outlying flows

Step 3: Finding anomalies

1. Using a simple thresholding detection approach, the top-ranked outlying flows are flagged as anomalies.

2.8 Outlier Subspace Analysis

In [12], David Kershaw, Qigang Gao and Hai Wang proposed a method which employs SPOT (Stream Projected Outlier de-Tector) as a prototype system for anomaly-based intrusion detection. SPOT is an outlier detection approach capable of quickly processing high dimensional streaming data. A model of normal behavior is established by using the set of system calls generated during the execution of a UNIX process. This model is further defined by taking sequences of system calls of a certain length, known as the window length which is often set to six. Normal execution of a process generates a specific set of system call combinations. The normal behavior of a process is modeled using this collection. As the first step of offline learning, a Sparse Subspace Template (SST), a group of subspaces used to detect outliers is constructed. A multi-objective genetic algorithm is employed to produce the subspaces. In the detection stage, any arriving data point mapped to a cell with low density achieves a high abnormality value. The data points are placed in an “Outlier Repository”. An outlierness number is associated with each trace which helps to distinguish normal and intrusive traces.

2.9 Reference-Based Outlier Detection

NADO, Newtork Anomaly Detection using Outlier approach is proposed in [3] by Monowar H Bhuyan, D K Bhattacharyya and J K Kalita. It is an effective reference-based outlier detection approach for detecting anomalies in networks. It works by identifying reference points and by ranking outlier scores of candidate objects. First, a variant of the k-means clustering technique is applied to partition the dataset X into k clusters C1, C2,… Cm. Then, for each cluster, calculate reference points. Also build mean-based profile for each cluster. Outlier score is calculated for each candidate data object and the candidate data objects are ranked according to their score value. The data objects are sorted based on score values and the anomalies are reported with respect to a user-defined threshold.

3. PERFORMANCE COMPARISON

Among the 9 methods discussed above, 6 papers have discussed the experimental results for KDDcup99 intrusion dataset. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment. Their results are summarized in Table 2. Detection rate (DR) and False Positive Rate (FPR) are the most commonly used metrics. Detection rate is the ratio between the number of correctly detected attacks and the total number of attacks. FP rate is the ratio between the number of normal connections that are incorrectly misclassified as attacks and the total number of normal connections [1].

Table 2. Performance Tabulation

Method	Detection Rate (DR)	False Positive Rate (FPR)

Cost Distribution Based Outlier Detection [7]	96% 99%	0.2% 0.3%
Random Forests Based Outlier Detection [8]	65%	1%
Density Based Outlier Mining [9]	>75%	Not Analyzed
Outlier Ensemble Detection [2]	94.237%	1.63%
Weighted Distance Based Outlier Detection [10]	99.5%	2%
Reference Based Outlier Detection [3]	89.49%	Not Analyzed

Paper [6] focuses on reducing false positives. The authors have applied their method for DARPA 2000 data set. Reduction rate is 86% and one alert filtered is a true alert and so completeness is 98% and soundness is 71%. They also repeated their experiments on real-world IDS alerts. They got 92% reduction rate. All alerts filtered were false positives, so completeness is 100% and soundness is 50%.

The authors of paper [11] verified the effectiveness of UNADA to detect real single source-destination and distributed network attacks in real traffic traces from different networks. They show that their detection results provide a stronger evidence of the accuracy of UNADA.

The authors of paper [12] used the UNM datasets to evaluate SPOT's effectiveness on ordered system call data. They have used two measurements for evaluation: false positive percentage (false positive rate) and true positive percentage (detection rate). Detection rate is above 90% when false alarm rate is about 2%.

4. RESEARCH ISSUES

- Presently, anomaly detection techniques based on outlier mining is yet not competent enough to fulfill real-time detection [2]. Designing appropriate outlier detection techniques which can be applied for real-time detection is a challenging task.
- Intrusion data may contain numeric, categorical and mixed type of data. Effective preprocessing techniques must be used. Data normalization and proper feature selection play an important role and it affects the performance of the system.
- Design of an appropriate scoring function that can handle all types of data still remains a challenging task.
- Clear definition of the term outlier in the context of the research work is necessary
- The data may contain noise similar to actual outliers. Distinguishing and removing noise is a difficult task.
- Generally, datasets used for evaluation of IDSs are very large. Most of the works do not use the entire data set. A data reduction technique called sampling is used. More accurate results can be obtained by doing experiments with the whole data set.

5. CONCLUSION

Some of the works using outlier analysis for anomaly based network intrusion detection has been studied. All the works have shown exemplary results. Outlier detection is significant in anomaly based intrusion detection systems. Nowadays, much research is being carried out in the field of outlier analysis. Some hybrid outlier detection approaches giving effective results have been proposed in the literature. Development of an effective outlier detection technique for evolving network data is a challenging task. Our future work will focus on designing and applying such an effective outlier detection algorithm for IDS. Further research in this direction will be fruitful in achieving high detection rate and low false positive rate.

6. REFERENCES

[1] Mahbod Tavallaee, Natalia Stakhanova, Ali Akbar Ghorbani, "Toward Credible Evaluation of Anomaly-Based Intrusion-Detection Methods", IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews, Vol.40, No.5, September 2010.

[2] Bin HUANG, Wen-fang LI, De-li CHEN, Liang SHI, "An Intrusion Detection Method Based on Outlier Ensemble Detection", IEEE International Conference on Networks Security, Wireless Communications and Trusted Computing, 2009.

[3] Manowar H Bhuyan, D K Bhattacharyya, J K Kalita, "NADO: Network Anomaly Detection Using Outlier Approach", ICCCS'11 February 12-14, 2011, Rourkela, Odisha, India Copyright © 2011 ACM 978-1-4503-0464-1/11/02.

[4] A Mira, D K Bhattacharyya, S Saharia, "RODHA: Robust Outlier Detection using Hybrid Approach", American Journal of Intelligent Systems 2012, 2(5): 129-140.

[5] Prasanta Godoi, D K Bhattacharyya, B Borah, Jugal K Kalita, "A Survey of Outlier Detection Methods in Network Anomaly Identification", The Computer Journal, Vol.54 No.4, 2011.

[6] Fu Xiao, Xie Li. 2008, "Using Outlier Detection to Reduce False Positives in Intrusion Detection", IEEE IFIP International Conference on Network and Parallel Computing, 2008.

[7] Komsit Prakobphol, Justin Zhan, "A Novel Outlier Detection Scheme for Network Intrusion Detection Systems", IEEE International Conference on Information Security and Assurance, 2008.

[8] Jiong Zhang, Mohammad Zulkernine, Anwar Haque, "Random-Forests-Based Network Intrusion Detection Systems", IEEE Transactions on Systems, Man, Cybernetics - Part C: Applications and Reviews, Vol.38, No.5, September 2008.

[9] Peng Yang, Biao Huang, "Density Based Outlier Mining Algorithm with Application to Intrusion Detection", IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008.

[10] S. Ganapathy, N.Jaisankar, P.Yogesh, A.Kannan, "An Intelligent System for Intrusion Detection Using Outlier Detection", IEEE International Conference on Recent Trends in Information Technology, 2011.

[11] Pedro Casas, Johan Mazel, Philippe Owezarski, "UNADA: Unsupervised Network Anomaly Detection Using Sub-space Outliers Ranking

[12] David Kershaw, Qignag Gao, Hai Wang, "Anomaly-Based Network Intrusion Detection Using Outlier Subspace Analysis: A Case Study", Canadian AI 2001, LNAI 6657, pp.234-239, 2011 © Springer-Verlag Berlin Heidelberg 2011.