VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2016 ISSN(e): 2309-3978;ISSN(p): 2411-6246
Volume 4, Number 1, January-December, 2016

pp. 21-26

PRACTICAL NETWORK ANOMALY DETECTION USING DATA MINING TECHNIQUES

XIEJUN NI[1], D HE*[1], F AHMAD[2]

[1]School of Computer Science and Software Engineering, East China Normal University, Shanghai, China
[2]Department of Computer Science, University of Central Punjab lahore, Pakistan
Email: djhe@sei.ecnu.edu.cn

Revised July 2015

ABSTRACT. *Network anomaly detection is an effective way to detect intrusions which defends our computer systems or network from attackers on the Internet. In this paper, we introduce the current research works in network anomaly detection and consider several pratical solutions for this issue. Different from signature-based method, data mining techniques can automatically extract normal pattern from a large set of network data and distinguish them from each other. However, those data mining techniques, such as classification, clustering, association rules and feature selection, can not be applied into this problem directly due to the characteristic of network data and technique themselves. We analyze those unfitness and propose some adaptation to detect anomaly timely and accurately.*

Keywords*:* Anomaly detection; Data mining; Feature selection; Clustering Algorithm; Classifiers; Association Rules.

1. **Introduction.** Intrusion detection systems (IDSs) play a significant role to effectively defend our crucial computer systems or networks against attackers on the Internet.

According to the information source of data, an IDS can be either host or network-based. The host-based IDS mainly consider host related data such as process identifiers and system calls. While a network-based IDS analyzes network related data, such as traffic volume, source/destination IP address, protocol usage, and port number,etc. This paper mainly focuses on the network based type of IDS.

Depending on the type of analysis procedure, network IDS can be categorized into misuse detection and anomaly detection. Misuse detection schemes, such as Snort [1], work well in discovering patterns of known intrusion. Once network data is matched with known attack signatures, an alarm is generated. However, signature-based schemes are not capable of detecting novel and unfamiliar intrusions. In addition, building new signatures require human experts' manual inspection which is not only expensive, but also induces a significant period of vulnerability between the discovery of a new attack and the construction of its signatures.

Conversely, anomaly-based approaches, a subset of intrusion detection systems, is an effective way to detect intrusion, which can discover patterns that do not conform to expected behavior. It has the ability to detect previously unseen intrusion event, but the rate of false positives is usually higher than misuse detections methods.

Patcha et al.[2] further categorizes anomaly detection methods into three categories according to the principle of detection: statistics-based, data mining-based and machine learning-based. Statisticsbased method is difficult to adapt to the non-stationary variation of the network traffic, which leads to a high false positive rate [3]. To alleviates these shortcomings, a number of ADSs employ data mining techniques [4–8]. Data mining techniques aim to discover understandable patterns or models from given data sets [9]. It can efficiently identify profiles of normal network activities for anomaly detection, and build classifiers or clustersing models to detect attacks. Some earlier work show that these techniques can help to identify

abnormal network activities efficiently.

In this paper, we mainly consider anomaly detection methods based on data mining techniques, analyze advantages and disadvantages of those methods. Aiming at these disadvantages, several pratical solutions are given to obtain better detection result.

The paper is organized as follows. In Section 2 we describe the current network detecting condition and open issues and challenges in anomaly detection. In Section 3 we give detailed analysis of mainstream intrusions/attacks detection techniques based on data mining in computer networks and systems. We present the pratical solution and discussion in Section 4.

2. Current detecting condition: In this section, we explain the definition of current network anomaly traffic and describe current detection conditions.

The open nature of the Internet allow any device or software that fit the stardard technology to connect to the Internet, which brings great challenge for QoS(Quality of Service) and network security management. Network anomaly traffic is the traffic patterns that impact negative effect on normal usage of network so as to control or damage target host.

Anomaly traffic. There are mainly four common kinds of network anomaly traffic that do harm to the network or systems.

1. DoS/DDoS attack. DoS(Denial of Service) or DDoS(Distributed DoS), refers to the attack that attackers can occupy large amounts of system resources which leads to the server refusing normal users' request. Ping Flood, Teardrop attack and Smurf are frequently-used DoS attack. DDoS is the further development of DoS. It can coordinate multiple computers on the Internet to launch an attack. Network data shows that during DDoS, packets from a lot of source IP addresses to single target , which is more harmful and difficult to defend.

2. Port Scan. Port Scan always be the prelude of an attack. Attacker usually spy all available ports in order to make useful breakthrough. Network data show that in the period, pakcets from one source IP address are sent to a mass of target IP addresses or different and continous ports on one target IP.

3. Worms/Virus. Worms/Virus take advantage of computer system vulnerability and spread themselves by several ways in the LAN(local area network) or WAN(wide area network). Virus or Worms, such as RedCide, Worm.Blaster and Worm Saaser scan the ports and launch attacks to vulnerable ones, in addition they are capable to spread themselves between computers by the Internet.

4. Uncontrolable P2P traffic. Most P2P(peer-to-peer) applications can transit firewall or proxy server ot obtain unrestricted access to the Internet. Some worms and virus adopt the principle of P2P to launch an attack. Some P2P applications are attached with other load, such as virus, worms, malicious plug-ins and harmful program codes.

So how to defense those threats and prevent our system and network from intrusion? As we know, intrusion actions such as spying, intrusion or attack rely on sending and receiving traffic, if we could detect those traffic and distinguish them from normal ones, then necessary measures would be carried out to stop or defense those intrusion.

Network data description. Nowadays, with the sharp development of the Internet, complex network structure and various network applications is make various security the network traffic data far more complex and large-scaled to analysis. And the famous 4V [10] concept of big data can also used to describe the network traffic:

1. Volume. The scale and complexity of network data is beyond the Moores law which means the amount of traffic to be detected in every terminal increases rapidly. String matching based signature method is a computational intensive task.

2. Variety. Network data usually is derived from various sources, where it is described in unstructured or semi-structured way. Proper integration is necessary to make uniform format.

3. Value. The value density of data is low. Anomaly detection problem usuallyfaces with high dimensional network data. Some features of these data are useless in identifying anomaly.

4. Velocity. The detection needs response in real-time in order to detect attack or anomaly in time.

There are some widely used network datasets with label information(normal or attack) for anomaly detection experiments and performance evaluation. For example, DARPA-99 intursion dataset [11] collects several weeks of traffic on the platform with simulation attacks. Those tcpdump format data should be open with packet analysis tools such like Wiresharks and then interesting fields of data would be extracted for running algorithm. Another famous datasets is KDDCup-99 datsets [12], it derived from DARPA-99 and

extract key features such as connection duration, protocol usage .etc to consist a 42-dimension network connection records set. KDD99 dataset is far more simple than former one, since one connection use a 42-dimension vector that take place of vast packets.

Briefly speaking, network anomaly detection has the requirement of dealing with large traffic in a short period so as to response or generate an alarm timely. In addition, the accurate of detection anomaly should be guaranteed since a high false negative rate and false positive rate can leads unnecessary alarms or potential damage.

3. Related Work. In this section, we focus on anomaly detection methods, especially data mining-based ones, and detailed analysis of those approaches are given.

Data mining is a subject that extract implicit, unknown in advance but potentially useful information and knowledge from large-scale, incomplete and noisy real-world application data. It has the advantage in mining features and rules from audit records and network traffic data that without too much priori knowledge, so it is suitable for anomaly detection. Wenke Lee [13] is the first person that applied data mining techniques into network anomaly detection, since then, a lot of related work is devoted to this research area. They can be classified into 5 types, respectively:

1. Classification. Classification is a primary analysis tools in data mining. The task of classification is learn from training dataset and then constructure a classification model or classification function that is capable to predict the category of data.

Before building classifiers, there should be a training dataset that contains label information which indicates the true category of the data. Then a classification model is built and used to analyze data without label. Commonly used algorithms are k-NN(k-Nearest Neighbor), Decision Tree classifier, Naïve Bayes classifier, ANN(Artificial Neural Network) and SVM(Support Vector Machines), etc.

Supervised anomaly intrusion detection approaches(classification) [8-10] highly rely ontraining data from normal activities, which are commonly used as data mining techniques. Since training data only contain historical activities, the profile of normal activities can only include the historical patterns of normal behavior. Therefore, new activities due to the change in the network environment or services are considered as deviations from the previously built profile, namely attacks.

2. Clustering. Clustering is a ubiquitous unsupervised learning method, aims to group objects into meaningful subclasses according their similarity. Members within the same cluster are similar to each other and members from dierent clusters are distinct from each other.

KMeans, a clustering method, is employed to detect unknown attacks and divide network data space e ectively in [14]. However the performance and computation complexity of KMean method are sensitive to the predefined number of clusters and initialized cluster centers. Wei et al. [15] employs improved FCM algorithms to obtain an optimal k. In [16], the authors proposed an anomaly detection method. This method utilizes a density-based clustering algorithm DBSCAN for modeling the normal activities of a user in a host.

Egilmez et al.[17] proposed a novel spectral anomaly detection method by developing a graph-based framework over wireless sensor networks. In their method, graphs are chosen to capture useful proximity information of measured data and employed to project the graph signals into normal and anomaly subspaces.

Kingsly et al.[18] presented a new density-based and grid-based unsupervised anomaly detection method . The method uses the adaptive grid algorithm to reduce the total number of potential dense units by merging small 1-dimensional clusters that have similar densities. However this method does not scale well with an increase in dimensions.

Clustering algorithms separate data points into different clusters according to their distance and density, with the growth of dimension number, the distance or distance could not work very well. Feature selection is necessary to decrease the number of useless features and improve the detection efficiency.

3. Outlier analysis. Outlier point is significantly different from most other observed values. Outlier analysis also called outlier mining, which refers to the process of discovering outlier points and handle them.

Outlier detection and clustering are closely related. Objects that outside the clusters of a dataset may represent intrusions. There are a lot of measurement of similarity between two different data points in multiple dimension space, such as Euclidean distance, Manhattan distance and Mahalanobis distance. If the distance between one point to other points are greater than a predefined D or there is less than N data points in it's neighborhood.

Ramaswamy et al. [19] proposed a scheme which computes the Euclidean distance of the k^{th} nearest neighbor from a point O. And lastest outlier detection algorithms [19-21] are using the full dimensional distances between two points as well as the densities of local neighborhoods. However, with high

dimensionality, it becomes difficult to estimate the similarity between data points precisely. And some outlier detection algorithm suffers from the drawback of high false positive rate.

4. Association analysis. Association analysis are commonly used in finding potential relationship between items. For example, mining association rules is able to find relationships between data, or sequence analysis is able to discover precedence relationship between events.

Given a database D of transactions where each transcation $T \in D$ denotes a set of items in which T is a set of items. An association rule is an implication in the format like X->Y, X and Y are items in T and antecedent and consequent of an association rule respectively. When dealing with association rules, confidence and support are two important concepts that show the universalityof a rule.

Association rules have been commonly used in mining normal patterns from various kinds of network data for anomaly detection [22,23]. They are particularly important in the domain of anomaly detection because association rules can be used to construct a profile of anomalous connections detected by the intrusion detection system. Usually association rules constructs the normal profile of network data from attack-free, however attack free data are difficult to obtain.

4. Solutions

Since most approaches have it's weak side, necessary adaptation are made to make it easier for detection and improve the detection result. In this section, we briefly propose two hybrid detecting models to reducing the difficulty in applying those algorithms and to achieve better precision or shorter detection time.

1. Feature selection before clustering. As aforementioned, clustering algorithms show its low effectiveness in detecting high dimensional network data, and among those features, considerable features are useless since they could not provide valuable information for classification. Traditional feature selection, is skilled in dealing with network with discrete features, however network data contains both continuous or discrete features. One idea is take discretization operation to those discrete features and change them into discrete values. Another way is to use selection approaches that unified both features, such as Maximal Information Coefficient [24], to evaluate the importance of one feature to the labels, and decide which one to be get rid of from the whole feature set.

Step1. Extract features from original tcpdump

Step2. Run Maximal Information Coefficient algorithm, calculate the information coefficient betweet from one feature to each other.

Step3. Get rid of feature with low information value and remain features with high information value.

Step4. Run any clustering algorithm on the reduced data. After get clusters, extract examples from every cluster and label them by expert analysis.

Step5. Use label result of Step4 to represent the other data in the same clusters.

2. Outlier detection with association rules. Outlier detection use a simple schemes that regards outlier points as anomaly data and others as normal. However, the result outlier judgement is not so accurate according to the paramters and leads to high false positive and false negative rate. To reduce those rates, association rules can be used to construct normal network profiles. The schems is, when a point is judged as an outlier point, we compare it with normal rules mined in advance. If the point matches one of the rules, which indicates that it may not be an anomaly point, if it is unmatched, then it would be labled with 'anomaly'. The association rules provides extra guarantee in case that outlier scheme makes a mistake.

Step1. Mining association rules from dataset by using Apriori or FP-Grow Tree algorithm. If a rule has the 'normal' field, store it in an string array.

Step2. After normal profile constructed, running outlier detection algorithms over data.

Step3. For every record, if it is regarded as an outlier point, then match it with every rule in the normal profile. If successfully machted, label it as normal data, otherwise anomaly.

Step4. If it is not regarded as an outlier point, label it as normal data.

Conclusion. In this paper, we considered network anomaly detection problem and briefly introduce mainstream data mining techniques applied in anomaly detection. A detailed analysis of advantages and disvantages of those schems are provided. We also propose several hybrid models that combines different data mining technologies as complementary and are capable to obtain better detection result in time or accuracy performance.

REFERENCES

[1] Roesch M. (1999).Snort: Lightweight Intrusion Detection for NetworksLISA. 99(1):229-238.

[2] Patcha A, Park J M. (2007).An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks, 51(12): 3448-3470.

[3] Luo Y B, Wang B S, Sun Y P, et al. (2013) FL-LPVG: An approach for anomaly detection based on low-level limited penetrable visibility graph.

[4] Tran Q A, Duan H, Li X. (2004).One-class support vector machine for anomaly network traffic detection. China Education and Research Network (CERNET), Tsinghua University, Main Building, 310.

[5] Hu W, Hu W. (2005).Network-based intrusion detection using Adaboost algorithmWeb Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on. IEEE, 2005: 712-717.

[6] Zhou Q, Gu L, Wang C, et al. (2006).Using an improved C4. 5 for imbalanced dataset of intrusion. Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services. ACM, 67.

[7] Zhang J, Zulkernine M, Haque A. (2008).Random-forests-based network intrusion detection systems. Systems, Man, and Cybernetics, Part C: Applications and Reviews,IEEE Transactions on, 38(5): 649-659.

[8] Tong X, Wang Z, Yu H. (2009).A research using hybrid RBF/Elman neural networks for intrusion detection system secure model. Computer physics communications, 180(10): 1795-1801.

[9] Hand D J, Mannila H, Smyth P. (2001).Principles of data mining. MIT press.

[10] Camacho J, Macia-Fernandez G, Diaz-Verdejo J, et al.(2014). Tackling the Big Data 4 vs for anomaly detection. Computer Communications Workshops (INFOCOMWKSHPS), 2014 IEEE Conference on. IEEE, 500-505.

[11] Lippmann R, Haines J W, Fried D J, et al.(2000).The 1999 DARPA off-line intrusion detection evaluation. Computer networks. 34(4): 579-595.

[12] Tavallaee M, Bagheri E, Lu W, et al. (2009).A detailed analysis of the KDD CUP 99 data set. Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009.

[13] Lee W, Stolfo S J. (1009).Data mining approaches for intrusion detection. Usenix security.

[14] Jianliang M, Haikun S, Ling B. (2009).The application on intrusion detection based on k-means cluster algorithm. Information Technology and Applications, 2009. IFITA'09. International Forum on. IEEE, 1: 150-152.

[15] Jiang W, Yao M, Yan J. (2008).Intrusion detection based on improved fuzzy c-means algorithm. Information Science and Engineering, 2008. ISISE'08. International Symposium on. IEEE, 2: 326-329.

[16] Oh S H, Lee W S. (2003).An anomaly intrusion detection method by clustering normal user behavior. Computers & Security, 22(7): 596-612.

[17] Egilmez H E, Ortega A. (2014).Spectral anomaly detection using graph-based filtering for wireless sensor networks. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 1085-1089.

[18] Leung K, Leckie C. (2005).Unsupervised anomaly detection in network intrusion detection using clusters. Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38. Australian Computer Society, Inc. 333-342.

[19] Ramaswamy S, Rastogi R, Shim K. (2000).Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Record. ACM, 29(2): 427-438.

[20] Breunig M M, Kriegel H P, Ng R T, et al. (2000).LOF: identifying density-based local outliers. ACM sigmod record. ACM, 29(2): 93-104.

[21] Knox E M, Ng R T. (1998).Algorithms for mining distancebased outliers in large datasets. Proceedings of the International Conference on Very Large Data Bases. 392-403.

[22] W. Lee, S.J. Stolfo, K.W. (1999). Mok, A data mining framework for building intrusion detection models, in: Proceedings of the IEEE Symposium on Security and Privacy, Oakland,CA, pp. 120–132.

[23] D. Barbara′ , J. Couto, S. Jajodia, N. Wu. (2001). ADAM: a testbed for exploring the use of data mining in intrusion detection,ACM SIGMOD Record: SPECIAL ISSUE: Special section on data mining for intrusion detection and threat analysis 30:15–24.

[24] Peng H, Long F, Ding C. (2005).Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(8): 1226-1238.