Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

A multi-level intrusion detection method for abnormal network behaviors

Soo-Yeon Ji [a,*], Bong-Keun Jeong [b], Seonho Choi [a], Dong Hyun Jeong [c,**]

[a] *Department of Computer Science, Bowie State University, 14000 Jericho Park Road, Bowie, MD 20715, USA*
[b] *Department of Computer Information Systems, Metropolitan State University of Denver, 1201 5th Street, Denver, CO 80204, USA*
[c] *Department of Computer Science and Information Technology, University of the District of Columbia, 4200 Connecticut Avenue NW, Washington, DC 20008, USA*

ARTICLE INFO

Article history:
Received 30 August 2015
Received in revised form
5 December 2015
Accepted 18 December 2015
Available online 30 December 2015

Keywords:
Network traffic analysis
Discrete wavelet transform
Visual analytics
Support vector machine

ABSTRACT

Abnormal network traffic analysis has become an increasingly important research topic to protect computing infrastructures from intruders. Yet, it is challenging to accurately discover threats due to the high volume of network traffic. To have better knowledge about network intrusions, this paper focuses on designing a multi-level network detection method. Mainly, it is composed of three steps as (1) understanding hidden underlying patterns from network traffic data by creating reliable rules to identify network abnormality, (2) generating a predictive model to determine exact attack categories, and (3) integrating a visual analytics tool to conduct an interactive visual analysis and validate the identified intrusions with transparent reasons.

To verify our approach, a broadly known intrusion dataset (i.e. NSL-KDD) is used. We found that the generated rules maintain a high performance rate and provide clear explanations. The proposed predictive model resulted about 96% of accuracy in detecting exact attack categories. With the interactive visual analysis, a significant difference among the attack categories was discovered by visually representing attacks in separated clusters. Overall, our multi-level detection method is well-suited for identifying hidden underlying patterns and attack categories by revealing the relationship among the features of network traffic data.

1. Introduction

Due to the advancement of Internet technologies, applications, and protocols, network traffic analysis has become more difficult since it deals with extreme amount of network traffic data. Because of the network complexity, network traffic analysis to detect unauthorized network intruders is also considered as one of the increasingly important research topics in network security.

To address the issue of protecting computing infrastructures by detecting network intruders, numerous intrusion detection (ID) techniques have been proposed. A traditionally known ID system discovers threats by analyzing traffic data at the network layer. The intrusion detection system (called host-based IDS) identifies threats on computer hosts by monitoring computer system logs, system calls, network events, and files (Das and Sarkar, 2014). To detect any abnormal behaviors, it monitors network packets to find possible attack signatures and compare them to known attack patterns. Although the host-based IDS is designed to prevent intruders by changing computer system security policies, it cannot monitor network traffic effectively because it only detects intrusions based on the analysis of information such as logs or packets (Bace, 1999). The system may detect threats based on known attack signatures, but new attacks cannot be discovered (Rubin et al., 2004).

Most analysis approaches are designed to detect intrusions by conducting misuse detection and anomaly detection. The misuse detection searches for events (i.e. known attacks) that are matched to predefined signatures (Kumar and Spafford, 1994). The anomaly detection identifies abnormal behaviors on hosts or networks based on the assumption that each attack shows different behaviors compared to normal activity. Therefore, it is possible to identify any abnormal attacks without having specific knowledge. Due to this advantage, the anomaly detection is used for designing various applications in other areas such as credit cards fraud

[*] Principal Corresponding author.
[**] Corresponding author.
E-mail addresses: sji@bowiestate.edu (S.-Y. Ji), bjeong@msudenver.edu (B.-K. Jeong), schoi@bowiestate.edu (S. Choi), djeong@udc.edu (D.H. Jeong).

http://dx.doi.org/10.1016/j.jnca.2015.12.004
1084-8045/© 2015 Elsevier Ltd. All rights reserved.

detection (Kou et al., 2004), fault detection in safety critical systems (Worden and Dulieu-Barton, 2004), and any domains that aim to detect abnormal activities including a medical field (Duftschmid and Miksch, 2001). However, the anomaly detection method may provide a high false alarm rate, and require extensive training sets to achieve a reliable performance result (Chandola et al., 2009; Eskin et al., 2002).

Abnormal behaviors are considered as different patterns if they do not match to a well-defined model representing normal behaviors. To discover abnormal behaviors (i.e. intrusions or attacks), understanding their trends or patterns is essential. ID can help us to minimize further damages by providing early warnings. In this paper, we extended our two previous studies by focusing on (1) generating simple and reliable rules to identify intrusions, (2) building a predictive model to determine exact attack categories by utilizing a signal processing technique (i.e. DWT) and Support Vector Machine (SVM), and (3) visually representing the input data to support an interactive visual analysis. For the visual analysis, a visual analytics tool called iPCA (Jeong et al., 2009) was used. With this tool, an interactive visual analysis was conducted to understand the intrusions and their relationships.

The rest of this paper begins with explaining related work in Section 2, our approach including a description of the data (i.e. NSL-KDD) and methods in Section 3. Study results are provided in Section 4. Lastly, Section 5 presents implications of this study and avenue for future research.

2. Related work

Researchers have applied various algorithms or theories such as statistics, machine learning, data mining, information theory, and spectral theory to extract patterns from attacks and design better anomaly detection techniques. Machine Learning (ML) is one of the broadly used algorithms in anomaly detection. ML techniques develop classifiers to determine possible attacks. Markou and Singh (2003a,b) proposed a detection technique with utilizing neural networks and statistical approaches. Rule-based anomaly detection techniques are introduced to capture rules that can identify network behaviors using Fuzzy (Chadha and Jain, 2015; Amini et al., 2015) or decision trees (Lee et al., 2008; Kruegel and Toth, 2003; Stein et al., 2005; Jain and Abouzakhar, 2013). Also, clustering technique (Lin et al., 2015) and SVM (Kuang et al., 2015; Wang et al., 2015; Aslahi-Shahri et al., 2015; Sani and Ghasemi, 2015) are used by numerous researchers to detect abnormal network behaviors. For instance, Xiang et al. (2008) introduced a multiple-level hybrid classifiers combining tree classifiers and Bayesian clustering to detect network anomaly. Kuang et al. (2015) presented a hybrid classifier by integrating SVM and principal component analysis. Golmah (2014) proposed an hybrid intrusion detection method integrating both C5.0 and SVM.

To generate a reliable ID system model, feature selection and extraction are considered as critical tasks for saving computational cost as well as for discovering data patterns. The feature selection is used to select a subset of most meaningful features from the original feature. The feature extraction is necessary for converting input data to reduce dimensions. There are various techniques that can be used for the feature extraction and selection such as Genetic Algorithm (GA) (Aslahi-Shahri et al., 2015), entropy of network features (Agarwal and Mittal, 2012), Partial Least Square (PLS) (Gan et al., 2013), Kernel Principal Component Analysis (KPCA) (Kuang et al., 2015), and cuttlefish optimization algorithm (Eesa et al., 2015). When applying the feature extraction, there is an important consideration whether the characteristics of original input data are transmitted to extracted new feature sets. However, it is important to note that the generated new feature set may not maintain the same or similar patterns compared to the original input data (Yang et al., 2011). Sanei et al. (2015) addressed the potential capability of discovering important features from input data by utilizing signal processing techniques. In our previous studies (Ji et al., 2014a,b), we emphasized the importance of detecting network abnormal behaviors. More specifically, in the study (Ji et al., 2014a), two-level ID method was introduced using a publicly available internet traffic data to show its capability in classifying abnormal network traffic. Fractal dimension (FD) was applied to identify the specific attack. Our previous works focused on generating rules to detect network anomalous activities and finding the self-similarity among the attacks. While the generated rules clearly differentiated normal and abnormal behaviors, there was a limitation of providing a detailed information (i.e. reasons) about the detected abnormal behaviors. To address this limitation, the categorical variables are converted to dummy variables. In addition, a visual analytics approach is integrated to identify transparent reasons about detected abnormal activities.

3. Approach

3.1. Data description

In this study, a publicly available intrusion detection dataset (called NSL-KDD dataset NSL-KDD, 2014; Tavallaee et al., 2009) is used. NSL-KDD dataset is the refined version of the KDD cup'99 dataset that redundant data records are removed (Tavallaee et al., 2009; NSL-KDD, 2014). The NSL-KDD dataset includes training set (125,973 records) and testing set (22,544 records). It contains 41 attributes (three nominal, six binary, and thirty-two numeric attributes), and includes normal activity and twenty-four attacks. These attacks are grouped into four major categories. Table 1 represents the four major attacks and intrusion categories. In this study, the training and testing data were combined to make a new input data. A total of 148,517 records were used as an input data.

DoS attack indicates any attempts to disable network access from remote machines (or computing resources). R2L represents that a remote user gains an access to local user accounts by sending packets to a computing machine over the network. Probe indicates that network is scanned to gather information to find known vulnerabilities. U2R denotes that an attacker accesses normal users' accounts by exploring the system as a root-user.

3.2. Methods

In this section, a brief explanation about our proposed multi-level network intrusion detection approach is provided. As shown in Fig. 1, the approach consists of three steps: (1) generating rules to detect outcome (normal/abnormal), (2) building an abnormal network behavior model to detect exact attack categories (i.e. DoS, Probe, R2L), and (3) conducting an interactive visual analysis to provide transparent reasons. First, the input data is divided into two subsets: categorical (i.e. nominal) data and numerical data. The nominal variables are used to generate rules. To determine exact attack categories, an extraction of significant DWT features from the numerical variables is performed. Furthermore, an interactive visual analysis is conducted to find the relationship between the raw and the DWT features and to present transparent reasons about the results.

3.2.1. Detection of abnormal behavior

Pre-Processing: As mentioned above, the NSL-KDD data set contains three nominal variables that include protocol type, service, and flag. However, each nominal variables contains many distinctive attribute values. Protocol type includes three attributes