KICS The Korean Institute of Communications and Information Sciences

CrossMark

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 2 (2016) 103–116

www.elsevier.com/locate/icte

FCAAIS: Anomaly based network intrusion detection through feature correlation analysis and association impact scale☆

V. Jyothsna*, V.V. Rama Prasad

Sree Vidyanikethan Engineering College A. Rangampet, Tirupati, India

Received 2 March 2016; received in revised form 6 August 2016; accepted 8 August 2016

Available online 26 August 2016

Abstract

Due to the sensitivity of the information required to detect network intrusions efficiently, collecting huge amounts of network transactions is inevitable and the volume and details of network transactions available in recent years have been high. The meta-heuristic anomaly based assessment is vital in an exploratory analysis of intrusion related network transaction data. In order to forecast and deliver predictions about intrusion possibility from the available details of the attributes involved in network transaction. In this regard, a meta-heuristic assessment model called the feature correlation analysis and association impact scale is explored to estimate the degree of intrusion scope threshold from the optimal features of network transaction data available for training. With the motivation gained from the model called "network intrusion detection by feature association impact scale" that was explored in our earlier work, a novel and improved meta-heuristic assessment strategy for intrusion prediction is derived. In this strategy, linear canonical correlation for feature optimization is used and feature association impact scale is explored from the selected optimal features. The experimental result indicates that the feature correlation has a significant impact towards minimizing the computational and time complexity of measuring the feature association impact scale.

Keywords: Intrusion detection; Feature reduction; Correlation analysis; Association impact scale

1. Introduction

Intrusion Detection Systems (IDSs) are mostly of two types: misuse IDSs and anomaly IDSs. A misuse IDS identifies intrusions based on parameters of known attacks and system weaknesses. It however does not recognize new or unfamiliar kinds of attacks. An anomaly IDS is based on the parameters of normal behavior and uses them for identifying any action that strays considerably from normal behavior. The mechanism of misuse intrusion detection trains on the existing patterns of intrusion and matches the data considered for examination with previous patterns to identify intrusions whereas, anomaly intrusion detection is based on identifying patterns from the examination data of normal usage.

IDSs that are efficient in nature are usually developed utilizing data mining techniques owing to their excellent performance of detecting intrusions and capability of generalization. However the process of implementing and installing such systems is complicated in nature. The inherent complications of the systems could be organized into separate problem sets based on the parameters of accuracy, competence, and usability. A key problem associated with IDSs built using data mining techniques, and mostly with those techniques based on anomaly detection is that they show a higher percent of false positive occurrences compared to the previous detection techniques based on hand-crafted signature. Hence, processing of audit data and detection of intrusions on-line are difficult for these techniques. Furthermore, compared to existing methodologies these techniques require vast training data and great complexity is associated with the learning process of the system. The key idea of the approach is to apply heuristic technique to anomaly based intrusion detection. The heuristic approach tends to define a scale that helps to assess the significance of the network

* Corresponding author.

E-mail addresses: jyothsna1684@gmail.com (V. Jyothsna), vvramaprasad@rediffmail.com (V.V. Rama Prasad).

Peer review under responsibility of The Korean Institute of Communications Information Sciences.

☆ This paper is part of a special issue entitled ICT Convergence in the Internet of Things (IoT) guest edited by Yacine Ghamri-Doudane, Yeong Min Jang, Daeyoung Kim, Hossam Hassanein and JaeSeung Song.

http://dx.doi.org/10.1016/j.icte.2016.08.003

transaction. The process devised requires feature extraction, dimensionality reduction for reducing the features extracted and feature selection. Feature extraction involves using all the features with transformation which comprises a combination of all the initial features. During feature selection, features are selected according to the classification criteria.

2. Related work

Eduardo DelaHoz et al. [1] proposed a classification approach that combines statistical techniques and self-organizing maps for detecting the anomalies in the network. Principal component analysis (PCA) and Fisher's discriminant ratio are used for feature selection and noise removal and probabilistic self-organizing maps are used to classify the network transactions as normal or anomalous. Ujwala Ravale et al. [2] proposed a hybrid technique that combines data mining approaches. The K-means clustering algorithm is used to decrease the number of attributes associated with each data point and the Radial basis function (RBF) kernel of support vector machine (SVM) is used for classification. Gaikward et al. [3] proposed a machine learning approach for implementing the IDS. The genetic algorithm is used to reduce the dimensions in the feature set and the partial decision tree is used as a base classifier to implement the IDS. Sunil Nilkanth Pawar et al. [4] proposed a genetic algorithm based network IDS with variable length chromosomes. A chromosome with relevant features is used for rule generation. An effective fitness function is used to define the fitness of each rule. Each chromosome has one or more rules for efficient detection of anomalies. Fangjun Kuang et al. [5] proposed a novel SVM model by combining kernel PCA (KPCA) with improved chaotic particle swarm optimization. KPCA is applied as a preprocessor of SVM to reduce the dimension of feature vectors and shorten training time and improved chaotic particle swarm optimization is proposed to estimate whether the action is normal or intrusion. Iftikhar Ahmad et al. [6] proposed an approach that used PCA for feature subset selection that is based on eigenvalues. Instead of using a traditional approach of selecting features with the highest eigenvalues such as PCA, the authors applied genetic principal components to select the subset of features and SVM for classification. Chun Guo et al. [7] proposed a hybrid learning method, named distance sum-based SVM (DSSVM), for modeling an effective IDS. In DSSVM, the distance sum based on the correlation between each data sample and the cluster centers feature dimensions in the data set is obtained and SVM is used as a classifier. Saurabh Mukherjee et al. [8] proposed a feature vitality based reduction method to identify important features used to detect the anomalies in the selection system, and applied the naive Bayes classifier to detect the anomalies in the IDS.

3. Data set description

The data set developed by Lee and Stolfo et al. [9], KDD-99 is an extensively used data set and is commonly selected for the evaluation of anomaly detection. The data generated from the Intrusion Detection Evaluation DARPA program 1998 was used to build the original KDD-99 data set [10,11] that comprises close to 4,900,000 unique connection vectors, where every connection vector consists of 41 features of which 34 are continuous features and 7 are discrete features. The NSL-KDD [12,13] data set is a polished version of its predecessor KDD-99 data set. As the NSL-KDD data set comprises a huge quantity of data, for experimental purpose sample data from the Kddcup.data_10_percent.gz is taken for the purpose of training. The NSL-KDD data set considered for training is 10% of the main data set equaling 494,020 connection vectors and labeled either as normal or as attack. The activities that show variations with respect to 'normal network behavior' are considered not 'normal' and labeled as attacks [14] and the records corresponding to normal behavior are labeled as normal. The attacks simulated in our experiments belong to any of the four types [15] described below:

1. Denial of service attack (DOS): The DOS attack is a type of attack where an attacker blocks access to valid users by consuming the resources of computer or memory making the system unable to handle valid requests. Examples of DOS attacks are many such as 'teardrop,' 'neptune,' 'ping of death (pod),' 'mail bomb', 'back', 'smurf' and 'land'.
2. Users-to-root attack (U2R): The root attack is a type of attack where the attacker gains access to a valid user account in the system and based on existing system weaknesses acquires access to the systems root component. There are several types of U2R attacks such as 'load-module', 'buffer overflow', 'rootkit', 'perl'.
3. Remote-to-local attack (R2L): The remote-to-local attack is a type of attack where an attacker without an account, accesses locally a legitimate user account based on existing machine vulnerabilities. R2L attacks types are 'phf', 'warezmaster', 'warezclient', 'spy', 'imap', 'ftp_write', 'multihop' and 'guess_passwd'.
4. Probing attack (PROBE): The probing attack is an attack type where an attacker evades the security and collects data on the computers in the network. The PROBE attacks types are 'nmap', 'satan', 'ipsweep' and 'portsweep'.

In the NSL-KDD data set the protocols taken into account are TCP, UDP, and ICMP.

4. Data set preprocessing

The network transactions set contains 42 features with values of type continuous and categorical. To facilitate the optimization process, these values should be numeric and categorical. Henceforth, initially all alphanumeric values have to be converted to numeric values and then the continuous values need to be converted to categorical.

4.1. The procedure to represent the alphanumeric values as numeric values and continuous values as categorical values

- Consider each feature with alphanumeric values and then list all possible unique values and list them with an incremental index that begins at 1.
- Replace the values with their appropriate indexes.