Expert Systems with Applications 41 (2014) 4139–4147

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A novel intrusion detection system based on feature generation with visualization strategy

Bin Luo, Jingbo Xia *

College of Science, Huazhong Agricultural University, Wuhan, Hubei, PR China

ARTICLE INFO

Keywords:
Intrusion detection system
Visualization
Feature generation

ABSTRACT

In this paper, a four-angle-star based visualized feature generation approach, FASVFG, is proposed to evaluate the distance between samples in a 5-class classification problem. Based on the four angle star image, numerical features are generated for network visit data from KDDcup99, and an efficient intrusion detection system with less features is proposed. The FASVFG-based classifier achieves a high generalization accuracy of 94.3555% in validation experiment, and the average Mathews correlation coefficient reaches 0.8858.

1. Introduction

Currently the quick developing Internet has brought great convenience for people globally to share the information. However, owing to the explosive use of networks, network attack becomes an urgent security issues (Liao, Lin, Lin, & Tung, 2013). After a Denial-of-service (DOS) attack upon American Yahoo that caused a halt of server and resulted in inestimable economic losses, investigator reported that the number of illegal code signature increased more than 256% over the previous year (Tjhai, Furnell, Papadaki, & Clarke, 2010).

To supervise user's network connection and prevent malicious attacks, a firewall setting was the initial option. Unfortunately, due to the weakness of connection analysis, firework could hardly discern all of the malicious attacks. Therefore, considerable attentions were paid to intrusion detection system (IDS), which is designed to classify user's activity as either normal or anomalous behavior by examining the dynamic characteristics of network connection records (Panda, Abraham, & Patra, 2012) and has been an essential part of a network security architecture nowadays (Tjhai et al., 2010). In general, architecture of IDS is divided into two categories in terms of analysis method: the anomaly detection and the misuse detection (Mukherjee & Sharma, 2012). While misuse detection aims to detect intrusions through established patterns of well-known attacks, anomaly detection intends to assort visitors into either legal or illegal users by comparing visitors' behaviors in an established profile that contains historical normal behaviors data. However, despite the fact that anomaly detection achieves high performance of detecting new attacks, it leads high misjudgment rate as well. Meanwhile, misuse detection performs less ideal in identifying the unknown attacks even though its misjudgment rate is low.

For the purpose of enhancing detection precision and detection stability, various artificial intelligence algorithms are investigated to improve IDS, such as fuzzy logic (Kumar & Selvakumar, 2013), K-nearest neighbor (KNN) (Tsai and Lin, 2010; Su, 2011), support vector machine (SVM) (Joseph, Das, Lee, & Seet, 2010; Mohammed & Sulaiman, 2012; Seongjun, Seungmin, Hyunwoo, & Sehun, 2013), artificial neural networks (ANN) (Wang, Hao, Ma, & Huang, 2010), Naïve Bayes networks (Koc, Mazzuchi, & Sarkani, 2012; Mukherjee & Sharma, 2012; Sanjai & Gao, 2014), decision tree (Gisung, Seungmin, & Sehun, 2014; Sindhu, Geetha, & Kannan, 2012), genetic algorithm (GA) (Abadeh, Mohamadi, & Habibi, 2011; Amin & Radu, 2013), self-organizing maps (SOM) (Tjhai et al., 2010), Markov chains (Seongjun et al., 2013), Cost matrix (Aikaterini & Christos, 2013), rough set (Nandita, Jaydeep, Jaya, & Moumita, 2013), ant colony algorithm (Feng, Zhang, Hu, & Huang, 2013) and principle component analysis (PCA) (Arunna, Tan, He, Priyadarsi, & Liu, 2013). These artificial intelligent algorithms usually offer an automatic mechanism and enhance the performance of IDSs. Among most mechanisms, feature selection is always a essential strategy which aims to decrease the training- and predicting-time, deal with data redundancy and irrelevancy, and finally enhance the IDS system.

For classification problem with high dimensional data, the purpose of feature selection is to decrease the computational time of the classification model and enhance the classification performance through removing redundant and irrelevant attributes. Generally, its significance could be shown in two respects: (a) Removing irrelevant and redundant features as well as filtering out noise. (b) Optimizing the procedure of finding a subset of features to a proposed desirable approach. Specifically, methods for feature selection can be divided into two categories: filter method

* Corresponding author. Tel.: +86 1339 7115 029.
E-mail addresses: xjb@mail.hzau.edu.cn, xiajingbo.math@gmail.com (J. Xia).

0957-4174/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.12.048

and wrapper method (Li et al., 2012; Mukherjee & Sharma, 2012). Among the two methods, the former is developed to determine which features should be retained through analyzing the contribution of sole features to the classification performance, and the latter aims to select critical features by estimating the resultant probability of error after removing certain features (Li et al., 2012).

Moreover, both wrapper and filter method and hybrid feature selection method along with various artificial intelligence approaches are proposed to achieve better performance. For example, in 2012, Lin, Ying, Lee, and Lee (2012) selected 23 critical features in KDDcup99 data set through support vector machine (SVM) and simulated annealing (SA), and obtained 99.96% classification accuracy. In the same year, by applying gradually feature removal method (GFR) to IDS, Li et al. (2012) extracted 19 important features and achieved 98.6429% classification accuracy in 10-fold cross validation. Furthermore, by adopting triangle area based nearest neighbors (TANN) to IDS, Tsai and Lin (2010) succeeds in generating 10 triangle areas formed by the data and the 5-class cluster centers to replace the 41 original features and outperforming the other three algorithms, e.g., KNN, SVM and the combination of K-means and KNN.

In addition, in the time of explosive development of Big Data and storage capacity, the data analysis tasks are becoming increasingly difficult and challenging. Meanwhile, the amount and the complexity of the data available are also increased significantly. Due to the limitations in human's cognitive and perceptual ability, it is necessary to adopt new ideas in data analysis so as to better cope with the massive knowledge discovery in Big Data. Under such circumstance, visualization is developed, which is a crucial component of research presentation and possesses two principal advantages: (a) Merging huge amounts of data into simple and effective graphics. (b) Providing efficient ways to analyze the information exist in the data sets in direct, easy-to-understand formats (Kelleher & Wagener, 2011). Besides, visualization serves two primary purposes: data analysis (Kelleher & Wagener, 2011; Shieh & Liao, 2012; Xu et al., 2010) and data presentation (Derick, Pedro, Abelardo, & Carlos, 2013; Kelleher & Wagener, 2011).

Over the past decades, visualization has been developing at a startling speed. With the emergence of computers vision, visualization has been incorporated in lots of fields and promoted the understanding of complicated concepts and ideas. An illuminative example was given in Bioinformatics, when Parkinson and Blaxter (2003) applied visualization to model a 4-class classification and resolved a gene classification problem. By mapping a certain gene into a triangular phase space, whose three vertexes represent three selected distinct gene categories, the position of the gene within the phase space indicates its relationship to each of the three selected data sets and helps determine the class the data belongs to. In geological science, Xie and Seng (2012) adopted three-dimensional (3D) visualization to synthesize and process geological engineering data and proved that the application of 3D visualization technology is helpful in improving the utilization and management of the geological engineering data.

Inspired by the applications of visualization strategy, especially the one in Parkinson and Blaxter's work (2003), visualization technology is considered to be applied in a new and novel intrusion detection system in this paper. We decrease the complexity of the data used in IDS and make the process of classification more intuitive by mapping the experimental data to a certain graph and generating new features to replace the original ones.

The purpose of this paper is to develop a novel intrusion detection system that combines the idea of feature generation and visualization technology. Basically, this work is a meaningful attempt that aims to adopt visualization strategy to achieve data presentation, feature selection, feature reduction, and classifier enhancement. We use a four star graph to give a intuitive simulation of high dimension data classification for IDS. Finally, generation of new visualized numerical features decrease the dimensionality of the data 41 to 16 or 4, and increase the computation speed of the new IDS. Visualization is an intuitive way for feature selection and feature reduction. For better understanding the enhanced performance, the proposed novel IDS is also compared with three other intrusion detection systems, and results show the proposed IDS outperforms the other three IDSs.

2. Materials and methods

2.1. Data set and data preprocessing

2.1.1. Data set

The KDDcup99 data set (downloaded from http://www.sigkdd.org/kddcup/index.php?section=1999&method=data) is a benchmark data set for intrusion detection. This data set consists of 9 weeks intrusion simulation in the US Air Force environment, and includes two versions, namely, the full data set (4,898,431 recordings, 18 M, 743 M Uncompressed), the 10% subset (494,307 recordings, 2.1 M, 75 M Uncompressed). For the sake of simplicity, the 10% subset is chosen as the experimental data set.

The whole recordings in KDDcup99 data set are divided into five classes, namely, Normal, denial of service (DOS), unauthorized access to local supervisor privileges (U2R), unauthorized access from a remote machine (R2L), probing, surveillance and other probing (Probe). And each record contains 41 features, of which 9 features are categorical data and 32 features are continuous data. Details of the 41 features are shown in Table 1.

2.1.2. Data preprocessing

The 10% KDDcup99 data set is highly redundant that 70.5% of the recordings are repetitions, which certainly affect the utilization of the data set. For these repetitions, we simply delete them. After the deletion, the size of the data set is decreased from 494,307 to 145,831. The particular size change of each class before and after the deletion is shown in Table 2.

Since the experiment is time consuming, while 10% of class Normal and Dos are randomly selected as the final experimental Normal and DOS data, data of the rest three classes are fully remained. In this way, the size of experimental data set is reduced from 145,831 to 17,480.

However, among the 41 features the second, third and fourth feature are characters. For each character, the sum of the difference between the value of American Standard Code for Information Interchange (ASCII) of each letter and upper case A or lower case a is considered as the value of the character. In this way, all the characters are transformed into numerical values. Finally, the data are normalized into [0, 1] by using the following formula:

$$x'_{ij} = \frac{x_{ij} - \min_{1\leqslant i\leqslant n}(x_{ij})}{\max_{1\leqslant i\leqslant n}(x_{ij}) - \min_{1\leqslant i\leqslant n}(x_{ij})},$$
$$i = 1, 2, 3, \ldots, n; \quad j = 1, 2, 3, \ldots, 41,$$

where $\min_{1\leqslant i\leqslant n}(x_{ij})$ and $\max_{1\leqslant i\leqslant n}(x_{ij})$ are the minimum and maximum value of the jth feature, respectively.

2.2. The four angle star based visualized feature generalization approach (FASVFG)

2.2.1. Idea of the FASVFG approach

The four angle star based visualized feature generalization approach (FASVFG) is a visualized method that enables people to identify the class to which the data belongs. The idea of visualization of FASVFG stems from SimTri (Parkinson & Blaxter, 2003), which amounts to a 3-class classification problem. In SimTri, each