Expert Systems with Applications 38 (2011) 14891–14898

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Self-adaptive and dynamic clustering for online anomaly detection

Seungmin Lee [a], Gisung Kim [b,*], Sehun Kim [c]

[a] Information Security Research Division, ETRI, 161, 138 Gajeongno, Yuseong-gu, Daejeon 305-700, South Korea
[b] Internet Security Lab, Department of Industrial and Systems Engineering, School of Information Technologies, Korea Advanced Institute of Science and Technology, 291, Daehak-ro, Yuseong-Gu, Daejeon 305-701, South Korea
[c] Graduate School of Information Security, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-Gu, Daejeon 305-701, South Korea

ARTICLE INFO

Keywords:
Self-organizing map
K-means clustering
Online anomaly detection

ABSTRACT

As recent Internet threats are evolving more rapidly than ever before, one of the major challenges in designing an intrusion detection system is to provide early and accurate detection of emerging threats. In this study, a novel framework is developed for fully unsupervised training and online anomaly detection. The framework is designed so that an initial model is constructed and then it gradually evolves according to the current state of online data without any human intervention. In the framework, a self-organizing map (SOM) that is seamlessly combined with *K*-means clustering is transformed into an adaptive and dynamic algorithm suitable for real-time processing. The performance of the proposed approach is evaluated through experiments using the well-known KDD Cup 1999 data set and further experiments using the honeypot data recently collected from Kyoto University. It is shown that the proposed approach can significantly increase the detection rate while the false alarm rate remains low. In particular, it is capable of detecting new types of attacks at the earliest possible time.

1. Introduction

There is an increasing need for early prevention against emerging Internet threats by intelligently analyzing the vast quantities of online data continuously generated from network devices (Kim, Shin, Kim, & Han, 2007; Lee, Kim, Kwon, Han, & Kim, 2008). One of the key objectives of deploying an intrusion detection system is to detect various types of attacks quickly and accurately. To achieve this objective, previously observed attack patterns need to be analyzed and profiled so that criteria for what constitutes normal traffic or an attack can be determined and applied to newly captured patterns for intrusion detection.

Conventionally, two different approaches have been used in detecting intrusions (Denning, 1987). The first approach, commonly known as misuse detection, is a rule-based approach that uses stored signatures of known intrusion events to detect known attacks. This approach has been highly successful in detecting occurrences of previously known attacks. However, it fails to detect new attack types and variants of known attacks whose signatures are not stored. When new attacks occur, the signature database has to be manually modified for future use.

The second approach is commonly known as an anomaly detection approach. Traditional anomaly detection, which is known as supervised anomaly detection, builds a model of normal data using labeled data and detects deviation from the normal model in observed data. In this approach, data mining techniques such as a support vector machine (SVM) and a neural network are usually used (Lee et al., 2001; Tand & Maxion, 2003). However, such supervised anomaly detection is impractical since it is very difficult or impossible to obtain either labeled or purely normal data (Eskin, Arnold, Prerau, Portnoy, & Stolfo, 2002; Portnoy, Eskin, & Stolfo, 2001; Zanero & Savaresi, 2004). To overcome this problem, there has been study on unsupervised anomaly detection, also known as anomaly detection over noisy data (Eskin, 2002). This approach basically assumes that the volume of normal data vastly overwhelms that of anomalous data. Then, it takes as input a set of unlabeled data and attempts to find intrusions buried within the data. That is, the unsupervised anomaly detection can be performed using unlabeled data, which is easy to obtain since it can be simply collected from the raw audit data of a system.

However, these conventional approaches often result in high false positive rates (Maxion & Tan, 2000). Such approaches also tend to give less than satisfactory performance in terms of noticing the changes of normal data as well as detecting new types of attacks. The reason is that real-world data is rapidly changing over time; that is, it is non-stationary. To reflect the state of non-stationary data in real time, a computational model constructed in the training process should be continuously updated whenever new data is observed. In particular, it is desirable for the model to be operated without any human intervention.

To address these problems, a novel framework is developed in this paper for fully unsupervised training and online anomaly detection. The framework is designed so that an initial model is constructed and then it gradually evolves according to the current

* Corresponding author. Tel.: +82 42 350 2954; fax: +82 42 350 3110.
E-mail address: kks00@kaist.ac.kr (G. Kim).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.05.058

state of online data. This approach consists of three main phases in which the self-organizing map (SOM) is seamlessly combined with *K*-means clustering. In the first phase, the map structure and size of the SOM is changed by the degree to which new data matches the best matching unit of the SOM. Then, the new data is classified by the boundaries distinguished by the *K*-means clustering. The second and third phases, respectively, update the boundaries of the current clusters and separate an attack cluster from a normal cluster while strictly monitoring the normal data patterns. Consequently, the network structure of the SOM is appropriately adapted to incoming data and then dynamically clustered. It is noted that clusters are automatically reconstructed or split up based on runtime accumulative measures. This enables new attacks to be detected at the earliest possible time. The performance of the proposed approach is evaluated through experiments using the well-known KDD Cup 1999 data set and further experiments using the honeypot data recently collected from Kyoto University.

This paper is organized as follows: Related work is discussed in Section 2. Section 3 describes the main phases of the proposed algorithm in detail. In Section 4, the algorithm is evaluated and discussed. The study concludes in Section 5 with a summary and plans for future research.

2. Related work

There have been several researches on unsupervised anomaly detection using a SOM (Kayacik, Zincir-Heywood, & Heywood, 2003; Litchodzijewski, Zincir-Heywood, & Heywood, 2002; Ramadas, Ostermann, & Tjaden, 2003; Rhodes, Mahaffey, & Cannady, 2000). The SOM is one of the most distinguished artificial neural network algorithms adhering to the unsupervised learning paradigm (Kohonen, 1982). It is a general unsupervised tool for the ordering of high-dimensional data in such a way that similar items are grouped spatially close to one another. This ordering retains the relationship between input data as faithfully as possible, thus describing a topology-preserving representation of input similarities in terms of distances on the feature map. The main advantage of such a mapping is the ease by which a user gains an idea regarding the structure of the data by analyzing the map.

However, the SOM has some limitations on its direct application to real-time intrusion detection (Deng & Kasabov, 2003; Rauber, Merkel, & Dittenbach, 2000; Samarasinghe, 2007). First, the SOM uses a fixed network architecture in terms of the number and arrangement of neural processing elements, and this architecture has to be defined prior to training. It is likely that a predetermined map size will be either too small or too large. It can be difficult to predetermine the map size for unknown input data characteristics and some measure of trial-and-error is necessary. Second, dimension reduction in the SOM can be too drastic and can generate a folded feature map of poor topology reservation, since the original data manifold may be complicated, bearing an inherent dimension larger than that of the feature map. The dimension of the map space usually is set to 2 or 3 for easy visualization. Third, it lacks the interpretability of a trained SOM. That is, a trained SOM can reveal possible cluster regions, but clusters are often not well separated. If the objective is to determine distinct clusters in the data, further clustering of the units on the map is needed to transform the map into unique clusters so that a cluster to which the data belongs can be easily determined. Therefore, it is not desirable to directly apply the SOM to an intrusion detection system that needs to analyze a large volume of data at high speed and automatically.

Several algorithms have been introduced to address these limitations. First, for overcoming the problem of the fixed architecture in the SOM, unsupervised neural learning algorithms were proposed. These algorithms include growing cell structure (GCS), growing SOM (GSOM), hierarchical SOM (HSOM) and growing hierarchical SOM (GHSOM) (Alahakoon, Halgamuge, & Srinivasan, 2000; Fritzke, 1994; Miikkulainen, 1990; Rauber, 2000). These algorithms are useful in reducing the map size since they can eliminate the trial and error required in determining map structure. Among these algorithms, the GHSOM offers a detailed view for a complicated clustering task involving hierarchical relationships, although it may require a substantial amount of time to find some important results. It is generally known that the clustering of hierarchical relationships is more effective in document classification rather than in intrusion detection. However, such algorithms adapt their architectures only during the training process according to the particular requirements of the input data. Thus, it is undesirable to directly apply the algorithms to online data for intrusion detection since adaptation tasks require tools with the ability to learn fast in order to reflect the current state of the non-stationary data. Thus, it may be a shortcoming that the network size is optimally determined during the training process. That is, due to the predetermined map structure, it can fail to detect the possible fluctuation of incoming data unless it adapts to data changes.

Second, the neural gas (NG) algorithm was proposed to remove the topology constraint on the feature maps that have a low dimension (Martinetz, Berkovich, & Schulten, 1993). It has a learning rule similar to that of a SOM, but the weight vectors are organized in the original manifold of the data space. Each time a weight vector is updated, its neighborhood rank needs to be computed. This increases the time complexity for one adapting step of the algorithm to the scale of $N \log N$ in a serial implementation, while searching for the best matching unit in the SOM scales only with N. In Fritzke (1995), the growing neural gas (GNG) algorithm originating from the neural gas algorithm was proposed. In the GNG, a new unit is inserted whenever input data is not closely matched to existing weight vectors. However, the unit with the maximum accumulated error should be searched out when unit insertion occurs. This contributes to extra computational overhead. An algorithm called the evolving SOM (ESOM), differing from GNG slightly in terms of the means of unit insertion and neighborhood updating, was also proposed in Deng and Kasabov (2003).

Third, the U-matrix method was proposed to facilitate the interpretation of the trained SOM by visualizing the distance structures and identifying the clusters on the two-dimensional map space (Kohonen, 1982). It shows the local distance structure of a topology preserving projection of a high-dimensional data set, which can be sufficient to detect cluster boundaries. However, it can be extremely difficult to automatically identify each unit within the boundaries during online processing tasks.

In summary, none of the algorithms mentioned so far has completely solved the three limitations on applying a SOM to online intrusion detection, although a few of them offer some advantages for manipulating the large amount of data encountered online. Here, a novel approach that is superior to the previous anomaly detection algorithms in terms of the above three aspects is presented. In the proposed approach, a SOM that is seamlessly combined with *K*-means clustering is transformed into an adaptive and dynamic algorithm suitable for real-time processing. Thus, it can continuously change its structure and size even during the online process while requiring little additional time overhead. As a result, the proposed approach enables intelligent knowledge interpretation as well as fast learning using online data.

3. Proposed framework

In this section, an overview of the proposed framework is first given and then the training and online anomaly detection procedures are described in greater detail.