Expert Systems with Applications 41 (2014) 5972–5983

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Efficient classification using parallel and scalable compressed model and its application on intrusion detection

Tieming Chen [a,*], Xu Zhang [a], Shichao Jin [b], Okhee Kim [b]

[a] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[b] School of Software Microelectronics, Peking University, Beijing 100190, China

ARTICLE INFO

Keywords:
Compressed model
MapReduce
Parallelization
Classification
Intrusion detection

ABSTRACT

In order to achieve high efficiency of classification in intrusion detection, a compressed model is proposed in this paper which combines horizontal compression with vertical compression. OneR is utilized as horizontal compression for attribute reduction, and affinity propagation is employed as vertical compression to select small representative exemplars from large training data. As to be able to computationally compress the larger volume of training data with scalability, MapReduce based parallelization approach is then implemented and evaluated for each step of the model compression process abovementioned, on which common but efficient classification methods can be directly used. Experimental application study on two publicly available datasets of intrusion detection, KDD99 and CMDC2012, demonstrates that the classification using the compressed model proposed can effectively speed up the detection procedure at up to 184 times, most importantly at the cost of a minimal accuracy difference with less than 1% on average.

1. Introduction

With the larger and larger amount of network communication data generated, the design of Intrusion Detection System (IDS) with high efficiency has become much more challenging. It is very important to discover abnormal behaviors at early stage, therefore, compared to the traditional signature-based detection, research on anomaly detection has been more popular in academia, as it has the potential power to detect unknown attacks by kinds of heuristic learning on the historical training data.

Anomaly detection generally includes two steps, building a model on training data and using the model for detection. However, training data are usually in a large scale, which can severely impede the detection since many detection models may need to scan all of them in certain cases. Intuitionally, an effective and direct way to reduce time cost for detection is to minimize the volume of a model that is used in the detection process, but building a systematic and scalable solution on generating such minimizing training data model for efficient intrusion detection is still in challenge.

To address this problem, our work will pay attention largely to the building of data compression instead of the detection phase, striving to boost the detection efficiency based on a proposed compressed model of training data. Therefore, the solution presented in this paper is applicable for those model-based anomaly detection approaches, especially for the classification based system (Varun, Arindam, & Vipin, 2009) because extracting the classification model directly from the huge volume of training data instances inevitably needs intense computation.

As for how to build compressed model, our proposal is made through inspecting into the following common natures of the training data, that is to say the motivation and inspiration of our works are generated from the following observations:

(a) By analyzing the attributes of the training data, we can easily find that the values of some attributes (features) in the whole training data only range in a small scale, which may have less impact on the detection accuracy.
(b) Some training instances are similar, because they are only different from each other on several attributes and the values of these attributes are slightly different, which may have redundancy for detection model building.
(c) For high dimensional training data, computing the similarity or some akin metric between each pair of instance is time-consuming. That means, for some novel but promising data processing algorithms such like Affinity Propagation (Frey & Dueck, 2007), the general computing memory would likely explode when the dimensionality of training data matrix increases to some large extent.

[*] Corresponding author. Tel.: +86 13857195552; fax: +86 57185290034.
E-mail address: tmchen@zjut.edu.cn (T. Chen).

http://dx.doi.org/10.1016/j.eswa.2014.04.009
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

For the purpose of effectively and efficiently handling these problems, we will propose a new framework of compressed model on training data. The model compression procedure mainly includes horizontal compression and vertical compression. The overall idea is presented in Fig. 1, where the first step is to normalize the original data followed by the horizontal compression and the vertical compression sequentially. Based on the compressed model, efficient classification can be directly built to detect new data without losing accuracy.

Furthermore, to computationally compress as large-scale as the training dataset could be, a cloud-based computing framework will be employed to parallelize the compression procedure, which realizes the scalability on training data compressing.

To summarize, we mainly make the following contributions in this paper:

(a) We propose a compressed detection model, which is a compact version from the original training dataset with regard to reducing both feature dimension and instance volume.
(b) We implement a MapReduce based parallel computing solution for the abovementioned model compression, which can compress larger scale of training data if only scaling up the involving distributed nodes.
(c) We finalize our compressed model-based classification approaches, and demonstrate the performance of our method on detection efficiency and accuracy on two publicly available intrusion detection datasets, KDD99 (Information and Computer Science of University of California, 1999) and CDMC2012 (The 3rd Cybersecurity Data Mining Competition, 2012).

The remainder of this paper is organized as follows. Literature works are firstly studied in Section 2. Section 3 introduces the methodology of our parallel and scalable compressed model, and explains its implementation procedure in detail. Section 4 describes the efficient classification deployments using the compressed model. Application experiments and its performance analysis on intrusion detection are presented in Section 5, while concluding remark and future work are discussed in last section.

2. Related work

During the past two decades, researchers in related fields have paid much attention to intrusion detection. Signature based detection (e.g. Snort Martin, 1999) relies on the knowledge of system vulnerabilities and known attack patterns, which hence is unable to detect unknown attacks. Correspondingly, anomaly detection is more dynamic and be able to detect novel attacks, which has attracted a lot of works worldwide. Generally, an anomaly-based intrusion detection system includes following steps, data gathering, data preprocessing (Davis & Clark, 2011), model building (Lee & Stolfo, 2000), and model-based detection (Liao, Tung, Richard Lin, et al., 2013). Although some common classification methods can be well used for the model-based detection, the way of model building may affect the detection results directly and heavily. Therefore, academic research on the model-based classification approach for intrusion detection is one unceasing focused topic. At beginning, detection accuracy, usually known as detection rate (recall) and false alarm rate (false positive), is widely concerned for the real-world application purpose. Recently, detection efficiency is more considered rather than accuracy to practical significance, especially on the potential abnormal behavior detection for high-speed and real-time network traffics. Nevertheless, both.

2.1. Efficiency concerned intrusion detection

In 1998, Lee and Stolfo (Wenke & Salvatore, 1998) published a data mining approach for the intrusion detection, where they proposed a framework for the agent-based intrusion detection, and deployed data mining methods to extract detection rules. Afterwards many researchers definitely focused on the way of boosting the detection speed. For example, Sung (Srinivas & Andrew, 2003) improved the detection speed by extracting the useful subset of attributes with ANN and SVM, and Srilatha, Ajith, and Johnson (2005) investigated the performance of Bayesian networks (BN) and Classification and Regression Trees (CART) to build lightweight IDS.

Actually, anomaly intrusion detection is a kind of complicated classification problem since there are usually too many attributes or features which may be redundant. So, attribute reduction or feature selection is the most popular method to improve detection efficiency by directly reducing the data attribute dimension (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2011). There are several basic but still in-progressing ways of feature selection for anomaly intrusion detection. PCA is a widely used criteria to select features for intrusion detection, which can be usually incorporated with other soft computing models, such as neural networks (Liu, Yi, & Yang, 2007), genetic algorithms (Ahmad, Abdullah, Alghamdi, et al., 2011), etc. PCA is a statistics-based

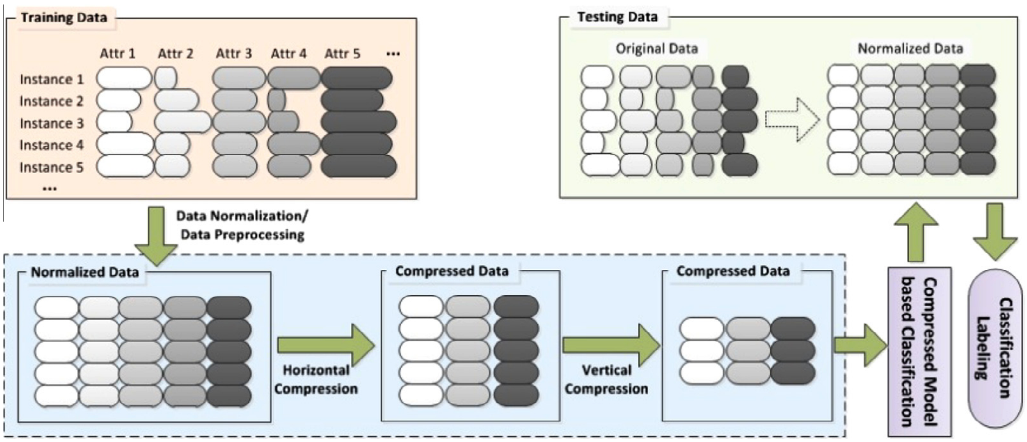

Fig. 1. The main idea of compressed model for data classification.