Expert Systems with Applications 36 (2009) 5605–5612

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Data mining-based intrusion detectors

Su-Yun Wu [a], Ester Yen [b,*]

[a] Department of Information Management, Vanaung University, Taiwan
[b] Mathematical Sciences Research Institute, Berkeley, CA 94720-5070, USA

ARTICLE INFO

Keywords:
Classification tree
Support vector machine
Internet attack
Intrusion detection system (IDS)
.

ABSTRACT

With popularization of internet, internet attack cases are increasing, and attack methods differs each day, thus information safety problem has became a significant issue all over the world. Nowadays, it is an urgent need to detect, identify and hold up such attacks effectively. The research intends to compare efficiency of machine learning methods in intrusion detection system, including classification tree and support vector machine, with the hope of providing reference for establishing intrusion detection system in future.

Compared with other related works in data mining-based intrusion detectors, we proposed to calculate the mean value via sampling different ratios of normal data for each measurement, which lead us to reach a better accuracy rate for observation data in real world. We compared the accuracy, detection rate, false alarm rate for four attack types. More over, it shows better performance than KDD Winner, especially for U2R type and R2L type attacks.

1. Introduction

In recent years, as internet and personal computers are populated, utilization rate of internet keeps increasing. It is changing people's lives gradually, and the majorities of people study, recreate, communicate and buy through internet. Besides common people, enterprise structure and business mode also undergoes transformation due to internet, and large enterprise or government organizations, in order to achieve operation purpose and efficiency, develop many application and service items resting on internet; these are an irresistible tendency in the new era.

However, though internet brings about convenience and real-timeliness, consequently comes information safety problem; for example: servers are attacked and paralyzed, inner data and information are stolen, and so on. In the event of such cases, big losses may be caused in money and business credit. For example, in 2000, American Yahoo was subject to DDos attack, the servers were paralyzed for 3 hours approximately, 1 million users were affected, and the losses involved were too large to calculate. Other famous business internets, such as CNN, eBay, Amazon.com, Buy.com, and so on, also suffered such internet attacks.

Because of convenience of internet, it is easy to get access to attack knowledge and methods. At present, hackers are unnecessary to have a wide knowledge of specialized knowledge, and annual internet attack cases are increasing to a great extent. According to the statistics of American Computer Emergency Response Team/Coordination Center (CERT/CC) (http://www.cert.org/), annual network attack cases showed index growth, in recent years; according to the report of Information Security (http://www.isecutech.com.tw/), internet attacks have became new weapon of world war, and the report said that Chinese Military Hacker had drew up plan, with the view of attacking American Aircraft Carrier Battle Group to make it lose fighting capacity through internet. Such information reveals that it is an urgent need to effectively identify and hold up internet attacks nowadays.

Common enterprises adopt firewall as the first line of defense for internet safety, but the main function of firewall is to supervise accessing behaviors of internet, and it owns limited detection capacity for internet attacks. Therefore, Intrusion Detection System, IDS is always applied to detect internet *encapsulation*, to improve protective capacity of internet safety.

IDS appears like internet supervision and alarm device, to observe and analyze whether the internet attacks may occur, timely send alarm before risks are caused by attacks, execute corresponding response measures, and reduce occurrence of bigger losses. Moreover, some technologies are based on pattern check, with low mis-judgment rate, but the pattern-based should be upgraded on a regular basis, such technologies do not possess enough detection capacity for unknown and renewed attack manners. Recently, many researches applied the technology of data mining and machine learning, which can analysis bulk data, and such technologies own better detection capacity for unknown attacks. Though some research achievements have been scored, there is a lot of development potential.

[*] Corresponding author.
E-mail address: ester_yen@yahoo.com (E. Yen).

0957-4174/$ - see front matter Crown Copyright © 2008 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.138

Under such circumstance with most same conditions, how is the efficiency of different machine learning methods applied in intrusion detection. Besides the said manners, what methods are there? Therefore, the research intends to compare the efficiency of different machine learning methods applied in intrusion detection, include classification tree, support vector machine, and so on, with the hope of providing possible suggestion for improvement, as the reference for building intrusion detection system.

The research process is shown in Fig. 1.

2. Literature review

2.1. Introduction on intrusion detection system

The concept of intrusion detection system was first suggested in a technical report by Anderson (1980); he considered that computer audit mechanism should be transformed and able to provide internal risks and threats for computer safety technicians, and suggested that statistics method should be applied to analyze users' behavior and detect those masqueraders who accessed system sources illegally. In 1987, Dorothy suggested a prototype of intrusion detection system: IDES (intrusion detection expert system), afterwards, the concept of intrusion detection system was known gradually, and his paper was also regarded as a significant landmark in intrusion detection area. Following this, intrusion detection system with various patterns was put forward, such as: Discovery, Haystack, MIDAS, NADIR, NSM, Wisdom and sense, DIDS, and so on (Bace, 2002).

Intrusion detection system is to supervise and control all cases happening to computer system or network system, analyze any signal arising from related safety problems, send alarms when safety problems occur, and inform related personnel or units to take relevant measures to reduce possible risks (Bace, 2002). Its framework includes three parts (Bace, 2002):

1. Information collection: Data collection: the source of these collected data can be separated into host, network and application, according to the position.
2. Analysis engine: Analysis engine is able to analyze whether or not there are symptom of any intrusion.
3. Response: Take actions after analysis, record analysis results, send real-time alarm, or adjust intrusion detection system, and so on.

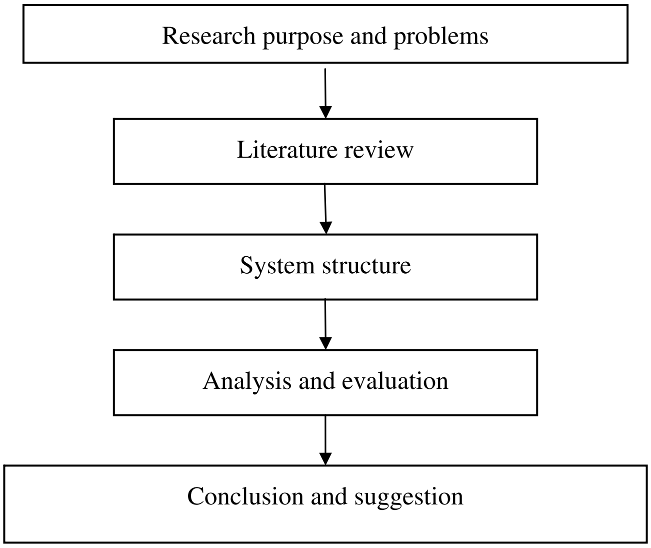

Fig. 1. Research flow.

2.1.1. Classification of intrusion detection system

Generally speaking, there are two kinds of classification methods for intrusion detection system:

1. According to different data sources, intrusion detection system includes host-based IDS and network-based IDS.
2. According to different analysis methods, intrusion detection system includes Misuse Detection and Anomaly Detection.

The following is to give a brief introduction on property, advantage and disadvantage of these intrusion detection systems.

(a) Classification based on different information source:

- (Host-based IDS) (Bace, 2002): Its data comes from the records of various activities of hosts, including audit record of operation system, system logs, application programs information, and so on. Taking Windows NT operation system as an example, its event logs mechanism searches and collects three patterns of system events: Operation system event, safety event and application event; and examples of application program information are as follows: Database system, WWW servers, and so on. Its advantage and disadvantage are stated as follows (Ertoz et al., 2004):
  - Advantage:
    1. It can judge whether or not the host is intruded more accurately: Because its data comes form system audit records and system logs of hosts, comparing with network-based intrusion detection system, it can more accurately judge network attacks or intrusion on hosts.
    2. It can detect attacks under encrypted network environment: Because the data comes from system files and transmitted encrypted data in network which are decrypted in hosts, thus the data is not affected.
    3. It does not need additional hardware: It just needs monitoring system installed in specified hosts, without additional hardware.
  - Disadvantage:
    1. Higher cost: Monitoring systems must be installed in each host; and because of different hosts, the audit files and log pattern are accordingly different, thus different intrusion detection systems are required in each host.
    2. It may affect system efficiency of monitored hosts: Intrusion detection system in monitoring state may occupy system sources of hosts.
- (Network-based IDS) (Bace, 2002): Its data is mainly collected network generic stream going through network segments, such as: Internet packets. And its advantage and disadvantage are stated as follows:
  - Advantage:
    1. Low cost: Only network-based IDS can detect all attacks in a LAN, and the cost is just for the device.
    2. It can detect attacks that cannot be done by host-based IDS, such as: Dos, DDos.
  - Disadvantage:
    1. The flux is large, and some packets may be lost, and it cannot detect all packets in network.
    2. In large-scale network, it requires more rapid CPU and more memory space, to analyze bulk data.
    3. It cannot deal with encrypted packets, and it may not receive attack information in encrypted packets accordingly.