Data Mining and Machine Learning Intrusion Detection Techniques in Encrypted Internet Communication

Michal AUGUSTÍN

Dept. of Computer and Informatics, FEI TU of Košice, Slovak Republic

michal.augustin@tuke.sk

Abstract — The emerging number of threats and attacks gives recency to Intrusion Detection and Prevention Systems (IDPS). These threads are even present in encrypted communication (e.g. HTTPS), which have an impact on availability. This paper presents intrusion detection system based on data mining and machine learning supervised algorithm C4.5 on collected and analyzed packet headers. The proposed system can classify normal network activities and main attack types such as DoS and Probe.

Keywords — intrusion detection, packet headers, anomaly intrusion detection, supervised machine learning algorithms

I. INTRODUCTION

Nowadays network security is a very important field of Informatics and with the emergence of Internet a lot of sensitive information, such as financial transactions are sent through this medium. This gave the need to protect this content – the use of encryption technologies and protocols. Some of the most widely used communication protocols that use encryption are SSL/TLS, IPsec, SSH etc [1] [2] [3].

Computer attacks, intruders, hackers and other malicious programs, viruses try to gain control over these systems, crash these systems down (e.g. DoS) or commit other frauds. Stable and secure encryption protocols with correct implementation partially removed these risks, and so it became harder for attackers to break encrypted communication, nevertheless other attacks in encrypted communication are still possible.

Consequently, computer networks have become more prone to different types of network attacks. Intrusion Detection Systems (IDS) offer methods to protect these networks of many attacks, some IDS use signature based rules other use anomaly detection. Signature based detection works well on known attack patterns which are stored in a database or a file. These IDS must be frequently updated in order to keep their signatures up to date for efficient detection of novel attacks or threats. Signature-based detection suffers from the inability to detect new type or zero-day attacks, those attacks that have not signatures included in databases. Another considerable aspect is the continuous growth of the signature database in size.

Anomaly intrusion detection techniques can be applied to identify novel or zero-day attacks against computers and network infrastructure [4] [5]. Anomaly detection algorithms generally work in two phases; a training phase and a detection phase. Generally the training phase contains benign traffic and algorithms learn and create models out of it. In the detection phase the current traffic is compared with the learned profiles/models and any abnormality from the normal behavior is marked as anomaly.

It is also a challenge to detect new worms and viruses in early stages of spreading, an example is a single worm W32/SQLSlam-A which infected 75 000 machines in 30 minutes and caused failures of major network services [7].

Another frequent attack is Denial of Service (DoS) or Distributed Denial of Service (DDoS), which causes significant damages on information availability to intended users. There are two forms of DoS attacks, one form causes services to crash and other floods services [8] Some of these software tools for performing DoS attacks are available online, such as hping, httpping, evil-ping etc [10] [11].

Some of the top intrusion attacks are listed in Table I available from McAfee Labs [9].

<table>
<thead>
<tr>
<th>Some Top Intrusion Detection Attacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
</tr>
<tr>
<td>-------------------------------------</td>
</tr>
<tr>
<td>HTTP Flood</td>
</tr>
<tr>
<td>Jping</td>
</tr>
<tr>
<td>exe.file.b.ident</td>
</tr>
</tbody>
</table>

Encryption communication protocols encrypt content, this content is then sent over the Internet through the TCP/IP protocol, the data is then “broken” into packets for routing purposes (encapsulation) with IP-header and IP-data. IP headers can not be encrypted, because they are used for routing.

Threats and attacks are also a common case against encrypted communication, an example is requests to generate several thousand or million page on HTTPS server – SSL over HTTP.
Collecting, analyzing and storing these IP headers and applying data mining on large amount of IP headers, as well as using machine learning techniques can prevent attacks even on encrypted communication. This paper presents a data mining and machine learning combination approach on detecting intrusive attacks with the use of an open source network intrusion prevention and detection system SNORT and mySQL[6].

II. COLLECTING INFORMATION FROM HEADERS

Information stored in packet header is extracted using SNORT IDS to capture all relevant header information (not packet payload – payload is encrypted and so irrelevant) and then stored locally in a mySQL database. Information from packets such as IP header, TCP header, UDP header and ICMP header are collected. Collected data features are IP Source and IP Destination, IP Version, IP Header length etc. Fig. 1 shows the table data structure of collected information in mySQL database. Table II shows some examples of data obtained from SNORT.

III. APPLYING DATA MINING ON COLLECTED HEADERS

Data mining is a common process of extracting patterns from a large data sets by combining different methods of statistics and artificial intelligence with databases. The primary reason for using data mining is to assist in the analysis of collections of a given observation and behavior.

The pre-processing needs to collect a large data set, and then to uncover the patterns presented in the data. The training set was used to apply data mining algorithms and set a normal behavior of the network. The task of data mining was to correctly classify benign and abnormal activity on the network. The domains of interest were **number of packet headers, source of the packet and time**.

This gave the ability to cluster data and discover group structure in the data that is in some way similar.

For data mining purposes RapidMiner, an open-source system, was used with an access to database mySQL [12]. RapidMiner graphical user interface was used to execute a SQL query and plotting features were used to visualise the given results.

```
SELECT DISTINCT(iphdr.ip_src), count(iphdr.ip_src) AS count, sum(iphdr.ip_len) AS length FROM iphdr LEFT JOIN event ON iphdr.cid = event.cid WHERE iphdr.ip_src <> "%ip_src+" AND event.timestamp = "'+fromDateTime+'" GROUP BY iphdr.ip_src;
```

SQL query counts the number of a given IP source and a sum of IP length (of a given IP) with respect to time. Result of data mining were plotted on a graph. With a benign activity the structure of the data had regular patterns as shown in Fig. 2.

To evaluate the ability to detect an attack, a DoS attack on a Local Area Network (LAN) was performed and data collected after this attack was data mined. Fig. 3 shows a significant change in the amount of IP length packets coming from the same IP source address – peaks in the graph.

Fig. 2. Shows sum of packet header length (of a given IP) Y ax in respect to time X ax in network traffic flow

![Fig. 2. Shows sum of packet header length (of a given IP) Y ax in respect to time X ax in network traffic flow. Peaks in the graph indicate malicious activity.](image)

Fig. 3. Shows sum of header length (of a given IP) Y ax in respect to time X ax in network traffic flow. Peaks in the graph indicate malicious activity.
IV. APPLYING MACHINE LEARNING C4.5 ALGORITHM ON HEADERS

Machine Learning, a branch of artificial intelligence, is concerned with the design and development of algorithms that allow computers to evolve behaviors based on data – to learn [13].

Algorithm types in machine learning are organized into taxonomy, based on the desired outcome – Supervised Learning generate a function that maps inputs to desired outputs.

Some well known algorithms of supervised machine learning are those that create decision trees for example ID3 (Iterative Dichotomizer 3) by Quinlan or an extension C4.5 used for classification [14]. Decision tree approach considers with features of packet headers gained from network traffic. Decision tree algorithms have a high performance in classifying unknown attacks [15].

Decision tree approach considers with features of packet headers gained from network traffic. Decision tree consists of non-terminal nodes (roots, internal nodes) and terminal nodes (leaves) which efficiently classify data.

Root node is the first attribute with test conditions to split every record toward each internal node. At first decision tree is trained with known data by a learning type, afterward it can classify new data. This algorithm can predict new data by starting from root to internal nodes and then to a leaf node consisting of class – each non-terminal node has to be tested to meet given conditions.

C4.5 builds its decision from the training data by using the concept of information entropy [13]. Fig. 4 shows the process of creating a C4.5 decision tree and the tested attributes in an open-source machine learning program RapidMiner [12].

![C4.5 decision tree](image)

Fig. 3. C4.5 decision tree

V. BAGGING

Bagging is a method for improving results of machine learning classification algorithms. This method was formulated by Leo Breiman and its name was deduced from the phrase “bootstrap aggregating” [16] [17]. The bagging method creates a sequence of classifiers \( H_m \), \( m=1,\ldots,M \) in respect to modifications of the training set. These classifiers are combined into a compound classifier, this can be interpreted as a voting procedure.

VI. CONCLUSION

This paper proposed various ways of detecting intrusions using data-mining and machine learning supervised algorithms – decision trees C4.5 with an efficient data processing features extracted from packet headers. Packet headers contain important information that should be considered even in encrypted communication protocols, such as SSL and other, for better attack detection on networks. Detection rates can be improved by combining data-mining techniques and decision trees – Algorithm C4.5 and thus creating composite detection. Bagging methods can be also applied to improve results classification algorithms of machine learning.

ACKNOWLEDGMENT

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-0073-07.

REFERENCES