# DATA MINING FOR INTRUSION DETECTION

**R. J. Jadhav[1], U. T. Pawar [2]**

*[1] Department of computer Application Bharati Vidyapeeth University, Pune Yashwantrao Mohite Institute of Management, Karad. INDIA*

*[2] Department of computer science Shivaji University Kolhapur S.G.M college Karad INDIA*

*[1] rjjmail@rediffmail.com,*

*[2] usharanipawar@rediffmail.com*

***ABSTRACT***

***Increasing network intrusion becoming crucial problem in security infrastructures. Data mining techniques have been successfully applied in many different fields including insolvency prediction, churn prediction, marketing, process control, fraud detection, and network management. Today number of research projects is using data mining for intrusion detection system (IDS) and prevention. The goal of intrusion detection is to identify entities attempting to subvert in-place security controls. In this paper, we are mostly focused on data mining techniques that are being used for such purposes.***

***Keywords : Intrusion detection system (IDS), Data mining, Intrusion prevention, Insolvency prediction, churn prediction..***

## I. INTRODUCTION

Intrusion detection is the process of monitoring and analysing the events occurring in a computer system in order to detect signs of security problems (Bace, 2000). As the cost of the information processing and Internet accessibility falls, more and more organizations are be-coming vulnerable to a wide variety of cyber threats. Ac-cording to a recent survey [11] by CERT/CC (Computer Emergency Response Team/Coordination), the rate of cyber attacks has been more than doubling every year in recent times (Figure 1).

Intrusion detection (ID) is a type of security management system for computers and networks. An ID system gathers and analyses information from various areas within a computer or a network to identify possible security breaches.

- Intrusion detection functions include:
- Monitoring and analysing both user and system activities
- Analysing system configurations and vulnerabilities
- Assessing system and file integrity
- Ability to recognize patterns typical of attacks
- Analysis of abnormal activity patterns
- Tracking user policy violations

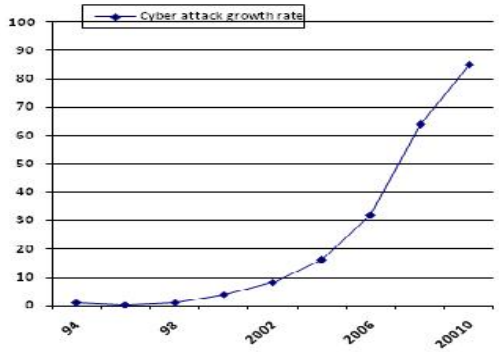


**Figure 1 Increasing rate of cyber crime**

ID systems are being developed in response to the increasing number of attacks on major sites and networks. The safeguarding of security is becoming increasingly difficult, because the possible technologies of attack are becoming ever more sophisticated; at the same time, less technical ability is required for the novice attacker, because proven past methods are easily accessed through the Web.

In 1998, ICSA.net, a leading security assurance organization, formed the Intrusion Detection Systems Consortium (IDSC) as an open forum for ID product developers with the aim of disseminating information to the end user and developing industry standards.

## II. INTRUSION DETECTION: A CAPSULE DESCRIPTION

An intrusion detection system (IDS) monitors network traffic and monitors for suspicious activity and alerts the system or network administrator. In some cases the IDS may also respond to anomalous or

malicious traffic by taking action such as blocking the user or source IP address from accessing the network.

IDS come in a variety of "flavours" and approach the goal of detecting suspicious traffic in different ways. There are network based (NIDS) and host based (HIDS) intrusion detection systems.

*NIDS*- Network Intrusion Detection Systems are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network. Ideally you would scan all inbound and outbound traffic, however doing so might create a bottleneck that would impair the overall speed of the network.

*HIDS*- Host Intrusion Detection Systems are run on individual hosts or devices on the network. A HIDS monitors the inbound and outbound packets from the device only and will alert the user or administrator of suspicious activity is detected.

***Signature Based***

A signature based IDS will monitor packets on the network and compare them against a database of signatures or attributes from known malicious threats. This is similar to the way most antivirus software detects malware.

***Anomaly Based***

An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is "normal" for that network- what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and alert the administrator or user when traffic is detected which is anomalous, or significantly different, than the baseline.

***Top five intrusion detection systems***

*Snort :* This lightweight network intrusion detection and prevention system excels at traffic analysis and packet logging on IP networks. Through protocol analysis, content searching, and various pre-processors, Snort detects thousands of worms, vulnerability exploit attempts, port scans, and other suspicious behaviour.

*OSSEC HIDS:* OSSEC HIDS performs log analysis, integrity checking, root kit detection, time-based alerting and active response. In addition to its IDS functionality.

*Fragroute/Fragrouter :* Fragrouter is a one-way fragmenting router - IP packets get sent from the attacker to the Fragrouter, which transforms them into a fragmented data stream to forward to the victim.

*BASE*: BASE is a PHP-based analysis engine to search and process a database of security events generated by various IDS, firewalls, and network monitoring tools. Its features include a query-builder and search interface for finding alerts matching different patterns, a packet viewer/decoder, and charts and statistics based on time, sensor, signature, protocol, IP address, etc.

*Sguil:* Sguil (pronounced sgweel) is built by network security analysts for network security analysts. Sguil's main component is an intuitive GUI that provides real time events from Snort.

## III. DATA MINING MEETS INTRUSION DETECTION

### *A. What is Data mining?*

The term data mining is often used to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information about the characteristics of the collected data, using a number of techniques (e.g. association rule mining). Forecasting and predictive modelling provide prediction of future events, and the processes may range from the transparent (e.g., rule based approaches) through the opaque e.g., neural networks). A primary reason for using data mining is to assist in the analysis of collections of observations of behaviour.

As a process, data mining keeps your business profitable. As a software application, it empowers you to uncover previously undetected facts. According to IBM, "Data mining offers firms in many industries the ability to discover hidden patterns in their data patterns that can help them understand market trends.

In order for us to determine how data mining can help advance intrusion detection it is important to understand how current IDS work to identify an intrusion. There are two different approaches to intrusion detection: misuse detection and anomaly detection. Misuse detection is the ability to identify intrusions based on a known pattern for the malicious activity. These known patterns are referred to as signatures. The second approach, anomaly detection, is the attempt to identify malicious traffic based on deviations from established normal network traffic patterns. Most, if not all, IDS which can be purchased today are based on misuse detection. Current IDS products come with a large set of signatures which have been identified as unique to a particular vulnerability or exploit. Most IDS vendors also provide regular signature updates in an attempt to keep pace with the rapid appearance of new vulnerabilities and exploits.

### *B. Why to use Data Mining:*

Current IDS have number of drawbacks.

*Data overload*. This is one of the aspect which does not relate directly to misuse detection but is extremely important is how much data can an analyst effectively an efficiently analyse. That being said the amount of data he/she needs to look at seems to be

growing rapidly. Depending on the intrusion detection tools employed by a company and its size there is the possibility for logs to reach millions of records per day.

*False negatives:* If a signature has not been written for a particular exploit there is an extremely good chance that the IDS will not detect it.

*Variants*. As stated previously signatures are developed in response to new vulnerabilities or exploits which have been posted or released. Integral to the success of a signature, it must be unique enough to only alert on malicious traffic and rarely on valid network traffic. The difficulty here is that exploit code can often be easily changed. It is not uncommon for an exploit tool to be released and then have its defaults changed shortly thereafter by the hacker community.

To overcome above drawbacks data mining can help to improve intrusion detection.

### *C. Data mining in intrusion detection*

The security of a computer system is compromised when an intrusion takes place. An intrusion can be defined [5] as ``any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource". Intrusion prevention techniques, such as user authentication (e.g. using passwords or biometrics), avoiding programming errors, and information protection (e.g., encryption) have been used to protect computer systems as a first line of defense. Intrusion prevention alone is not sufficient because as systems become ever more complex, there are always exploitable weakness in the systems due to design and programming errors, or various ``socially engineered" penetration techniques. For example, after it was first reported many years ago, exploitable ``buffer overflow" still exists in some recent system software due to programming errors. The policies that balance convenience versus strict control of a system and information access also make it impossible for an operational system to be completely secure.

Intrusion detection is therefore needed as another wall to protect computer systems. The elements central to intrusion detection are: *resources* to be protected in a target system, i.e.

User accounts, file systems, system kernels, etc; *models* that characterize the ``normal" or ``legitimate" behavior of these resources; *techniques* that compare the actual system activities with the established models, and identify those that are ``abnormal" or ``intrusive".

Our current architecture for intrusion detection is shown in Figure 2. Network traffic is analysed by a variety of available sensors. This sensor data is pulled periodically to a central server for conditioning and input to a relational database. HOMER filters events from the sensor data before they are passed on to the classifier and clustering analyses. Data mining tools filter false alarms and identify anomalous behaviour in the large amounts of remaining data. A web server is available as a front end to the database if needed, and analysts can launch a number of predefined queries as well as free form SQL queries from this interface.

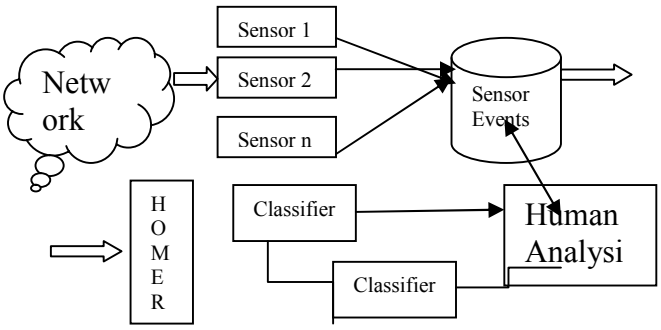


**Figure 2 Architecture for intrusion detection**

Many researchers have proposed and implemented different models which define different measures of system behavior, with an ad hoc presumption that normalcy and anomaly (or illegitimacy) will be accurately manifested in the chosen set of system features that are modeled and measured. Intrusion detection techniques can be categorized into *misuse detection*, which uses patterns of well-known attacks or weak spots of the system to identify intrusions; and *anomaly detection*, which tries to determine whether deviation from the established normal usage patterns can be flagged as intrusions.

Misuse detection systems, for example [8] and STAT [8], encode and match the sequence of ``signature actions" (e.g., change the ownership of a file) of known intrusion scenarios. The main shortcomings of such systems are: known intrusion patterns have to be hand-coded into the system; they are unable to detect any future (unknown) intrusions that have no matched patterns stored in the system.

Anomaly detection (sub)systems, such as IDES [12], establish normal usage patterns (profiles) using statistical measures on system features, for example, the CPU and I/O activities by a particular user or program. The main difficulties of these systems are: intuition and experience is relied upon in selecting the system features, which can vary greatly among different computing environments; some intrusions can only be detected by studying the sequential interrelation between events because each event alone may fit the profiles.

Our research aims to eliminate, as much as possible, the manual and ad-hoc elements from the process of building an intrusion detection system. We take a data-centric point of view and consider intrusion

detection as a data analysis process. Anomaly detection is about finding the normal usage patterns from the audit data, whereas misuse detection is about encoding and matching the intrusion patterns using the audit data. The central theme of our approach is to apply data mining techniques to intrusion detection. Data mining generally refers to the process of (automatically) extracting models from large stores of data .The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and database. Several types of algorithms are particularly relevant to our research:

***Classification:***

Maps a data item into one of several pre-defined categories. These algorithms normally output ``classifiers", for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient ``normal" and ``abnormal" audit data for a user or a program, then apply a classification algorithm to learn a classifier that will determine (future) audit data as belonging to the normal class or the abnormal class;

***Link analysis:***

Determines relations between fields in the database. Finding out the correlations in audit data will provide insight for selecting the right set of system features for intrusion detection;

***Sequence analysis:***

Models sequential patterns. These algorithms can help us understand what (time-based) sequence of audit events are frequently encountered together. These frequent event patterns are important elements of the behaviour profile of a user or program.

We are developing a systematic framework for designing, developing and evaluating intrusion detection systems. Specifically, the framework consists of a set of environment-independent guidelines and programs that can assist a system administrator or security officer to

- Select appropriate system features from audit data to build models for intrusion detection;
- Architect a hierarchical detector system from component detectors;
- Update and deploy new detection systems as needed.

The key advantage of our approach is that it can automatically generate concise and accurate detection models from large amount of audit data. The methodology itself is general and mechanical, and therefore can be used to build intrusion detection systems for a wide variety of computing environments

## IV. CONCLUSION

Data mining can help improve intrusion detection towards the enhancement of IDS by adding a level of focus to anomaly detection. We have shown the ways in which data mining has been known to aid the process of Intrusion Detection and the ways in which the various techniques have been applied and evaluated. Finally, in the last section, we proposed a data mining approach that we feel can contribute significantly in the attempt to create better and more effective Intrusion Detection Systems.

## REFERENCES

[1]. "Data Mining Approaches for Intrusion Detection“ Lee, Wenke and Stolfo, Salvatore.

[2] ”.Data mining for intrusion detection a critical review” Klaus Julisch IBM Research

[3] “Data mining methods for network intrusionDetection” S Terry Brugger University of California, davisJune 2004

[4]. Data Mining for Intrusion Detection: Techniques, Applications and Systems .Jian Pei Shambhu J. Upadhyaya Faisal Farooq Venugopal Govindaraju Department of Computer Science and Engineering State University .

[5]. “The architecture of a network level intrusion detection system. Technical report, R. Heady, G. Luger, A. Maccabe, and M. Servilla. Computer Science Department, University of New Mexico, August 1990.

[6] Applying Genetic Programming to Intrusion detection by Mark crosbie, Prof. Gene Spafford.

[7]. Rothleder, Neal. "Data Mining for Intrusion Detection." The Edge Newsletter

[8]. K. Ilgun, R. A. Kemmerer, and P. A. Porras. State transition analysis: A rule-based intrusion detection approach. *IEEE Transactions on Software Engineering*, 21(3):181-199, March 1995.

[9]. Kurth thearling Tutorial “An Introduction to Data Mining www.thearling.co

[10]..” A real-time intrusion detection expert system (IDES) - final technical report “ T. Lunt, A. Tamaru, F. Gilham, R. Jagannathan, P. Neumann, H. Javitz, A. Valdes, and T. Garvey.
Technical report, Computer Science Laboratory, SRI International, Menlo Park, California, February 1992.

[11] Successful Real-Time Security Monitoring, *Riptech white paper*, September 2010

[12] Bace, R. (2000). Intrusion Detection. Macmillan Technical Publishing.

[13] Barbara, D., N. Wu, and S. Jajodia [2001]. “Detecting Novel Network Intrusions Using BayesEstimators”, Proceedings Of the *First SIAM Int. Conference on Data Mining*, (SDM 2001),Chicago, IL.

[14] Ramaswarny, S., R. Rastogi, and K. Shim, [2000]. “Efficient Algorithms for Mining Outliersrom Large Data Sets”, Proceedings of the *ACM Sigmod 2000 Int. Conference on Management ofData,* Dallas, TX