THE USE OF DATA MINING IN THE IMPLEMENTATION OF A NETWORK INTRUSION DETECTION SYSTEM

John Sheppard, Joe Carthy, John Dunnion
University College Dublin
Ireland

Keywords: Data Mining, Network Intrusion Detection Systems, anomolies, network attack, decision trees.

Abstract: This paper focuses on the domain of Network Intrusion Detection Systems, an area where the goal is to detect security violations by passively monitoring network traffic and raising an alarm when an attack occurs. But the problem is that new attacks are being deployed all the time. This particular system has been developed using a range of data mining techniques so as to automatically be able to classify network tracffic as normal or intrusive. Here we evaluate decision trees and their performance based on a large data set used in the 1999 KDD cup contest.

1 INTRODUCTION

Over the past fifteen years, intrusion detection and other security technologies such as cryptography, authentication and firewalls have increasingly gained in importance (Allen et al., 2000). This is reflected by the amount of so called CERTs, Computer Emergency Response Teams that have arisen since the early nineties. Before these, reactions to attacks were isolated and uncordinated resulting in much duplicated effort and in conflicting solutions (First,). These CERTs worked together through organisation such as FIRST,the Forum of Incident Response and Security Teams, to prevent, detect and recover from computer security incidents by sharing alert and advisory information on potential threats and emerging security situations.

The impact of these attacks can be extreme. In some cases corporations have lost millions in loss of revenue or knowledge they were hiding from competitors. On other occasions universities or individuals have been hacked and used as middlemen for causing damage to others. The victim is then held responsible for any damage caused by not ensuring the security of their machine. Some experts believe that one day insurance for computer networks will be as common as fire and theft poliies. Detecting such violations is a necessary step in taking corrective action such as blocking the offender, reporting them or taking legeal action against them. Alternatively it allows administrators to identify those areas whose defences need improving such as the identification of a previously unknown vulnerability, a system that wasn't properly patched or a user that needs further education against social engineering attacks.

A good security model should be built upon several layers including:

- A good security policy for the organization
- Host system security
- Auditing
- Router Security
- Firewalls
- Intrusion Detection Systems
- Incident response Plan

(SANS,)

Using multiple layers strength ens deterrence of unauthorized use of computer systems and network services. As each layer provides some form of protection against an attack, the security breach of one layer does not leave the whole system open to attack. The layer focused on in this paper is Intrusion Detection Systems.

1.1 Intrusion Detection

An intrusion can be defined as any set of actions that attempt to compromise the integrity, confidentiality or

Sheppard J., Carthy J. and Dunnion J. (2005).
THE USE OF DATA MINING IN THE IMPLEMENTATION OF A NETWORK INTRUSION DETECTION SYSTEM.
In *Proceedings of the First International Conference on Web Information Systems and Technologies*, pages 399-404
DOI: 10.5220/0001231403990404

availability of a resource (Heady et al., 1990). Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems.

Traditionally, both in a commercial and research sense, these systems have been based on the idea of attack signatures provided by humans. Signatures were stored in a database and manually revised with the emergence of new attacks. Systems were unable to detect these new attacks, and it often took a substantial amount of time before the latest signatures were created and deployed. Attempts to detect attacks beyond this limited realm typically results in an unacceptable level of false positives.

In recent years the focus has turned to Data Mining as an alternative approach. Data mining techniques are used in two ways, Misuse Detection and Anomaly Detection. In misuse detection each instance is labeled as 'Normal' or 'Intrusive'. Then learning algorithms are trained on this data. Models created automatically are more precise and sophisticated than those that are manually created leading to high accuracy and low false alarm rates, but are unable to detect attacks whose instances have not yet been observed.

Anomaly detection algorithms build models of normal behavior and automatically detect any deviation from it which may result in unforeseen attacks. This is also known as the 'Paranoid Approach' as anything not seen before may be dangerous. Bearing in mind that "just because you're paranoid, doesn't mean they're not out to get you", this outlook can be seen as advantagous especially in a securitycontext. This gives the ability to detect attempts to exploit novel or unforeseen vulnerabilities. The high false alarm rate is generally seen as the major drawback.

The architecture of these systems can be host based or network based. Host based systems reside on an individual computer monitoring that particular machines log files. Network based systems monitor the network traffic analysing packets. In the Intrusion Detection domain data is so volumous and the analysis process so time consuming that administrators dont have the resources to examine and extract all the relevant knowledge except in rare circumstances such as a legal investigation following an attack. Some of the challenges being faced are

- large data sizes,
- high dimensionality,
- temporal nature of the data,
- skewed class distribution,
- data preprocessing and high performance computing.

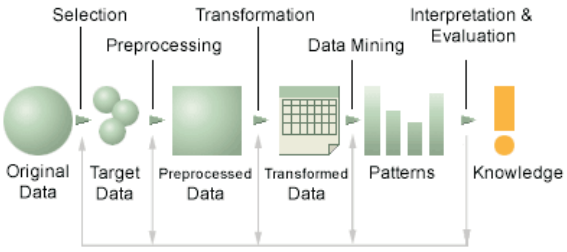

Figure 1: The KDD process

1.2 Data Mining

Data Mining is the process of extracting useful and previously unnoticed models or patterns from large datastores. (Bass, 2000; Manilla, 2002; Fayyad et al., 1996). Data mining is a component of the Knowledge Discovery in Databases (KDD) process (Carbone, 1997; Fayyad et al., 1996). There are several steps to the KDD process. These can be seen in figure 1,

- Selection, obataining the data from a source.
- Preprocessing, correcting any incorrect or missing data.
- Transformation, converting all data into a commmon format.
- Data Mining, using algorithms to extract the information and patterns derived from the kdd process.
- Interpretation/evaluation, how to present the results.

Data mining techniques can be differentiated by their model functions and representation, preference criterion, and algorithms (Fayyad et al., 1996). The function model in question here is classification. Common representations for data mining techniques include rules, decision trees, linear and non-linear functions, instance based examples and probability models.

2 DATA

The data set used here was the same as the data set used for the 1999 KDD intrusion detection contest. It is a version of the 1998 DARPA Intrusion Detection Evaluation Program which was prepared and managed by MIT Lincoln Labs. Lincoln labs set up an environment to acquire nine weeks of raw TCP dump data for a LAN simulating a typical US Air Force LAN. The raw training data was about 4 gigabytes of compressed binary TCP dump data from seven weeks of network traffic. This was processed into about 5

million connection records. Another two weeks of test data yielded approximately two million more connection records.

Each connection record consists of about 100 bytes. A connection is a sequence of TCP packets starting and ending at some well defined times between which data flows to and from a source IP address number under some well defined protocol. Each connection in the training set is labelled either as normal or as an attack.

The training data contains twenty four attacks, while the test data contains an additional fourteen attacks. These thirty eight attack types can be categorized into one of following four categories

DOS, a denial of service attack is an attack where the assailant makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine. An example of a DOS attack would be a Back Attack. This strike attacks the Apache web server by submitting requests with many frontslashes. As the server tries to process these frontslashes it consumes excessive CPU time slowing down and becoming unable to process other network requests.

R2L, a remote to local attack takes place when an attacker manages to send packets across a network targetting a machine the attacker does not have access to. There are many ways to do this varying from a dictionary attack on the users password to exploiting buffer overflow vunerabilities.

U2R, a user to root attack occurs when a normal user, (or possibly an unauthorized user who has already gained access to a normal user account through social engineering or sniffing passwords), is able to exploit some vulnerability in order to gain root access to the system. A common means to carrying out many U2R attacks is buffer overflow. When a programmer writes a program to recieve some input the size of the buffer in which it will be stored needs to be decided. An intruder will profit from this opportunity by filling the buffer and then including some extra commands to be submitted and understood by the operating system.

PROBE, a probe atack is a surveillance procedure where an attacker can quickly scan a network in order to gain information on the network and the machines for a possible attack in the future. (Kendall, 1999)

Each connection record is made up of forty two features in the training data and forty one in the test data. These features are divided into three categories. The first set is basic features of individual TCP connection such as the length of the connection, the type of protocol, the network service and the number of bytes from source to destination and vise versa. The second set are content features within a connection suggested by a domain knowledge including the number of failed login attempts, whether the attacker gained access, whether the attacker attempted to gain root access, whether the attacker was successful in gaining root acces, the number of file creation operations and the number of outbound commands in an ftp session. The final set of features are Traffic features. These were computed using a two second time window and resulted in

- Number of connections to the same host as the current connection
- Percentage of connections that have ”SYN” errors
- Percentage of connections that have ”REJ” errors
- Percentage of connections to the same service
- Percentage of connections to different services
- Number of connections to the same service as the current connection

Several categories of higher level features were defined including same host and same service features. The same host features study the connections in the past two seconds that have the same destination host as the the current connection while the same service features inspect the connections in the past two seconds that have the same service as the current connection. These features together are called time based features and it is from these that statistics relating to service and protocol behaviour are derived.

3 ARCHITECTURE

The architecture of the system can be seen in figure 1. It is a distributed system, a computer architecture consisting of interconnected processors. With distributed systems each processor has its own local memory. Processors communicate by message passing over the network. For any particular processor its own resources are local whereas the other processors and resources are remote. Together a processor and its resources are called a node.

The distributed architecture of this system allows for a shorter response time in analysing incoming network traffic and higher reliability due to various data mining algorithms performing differently depending on the attack.

The system works as follows,

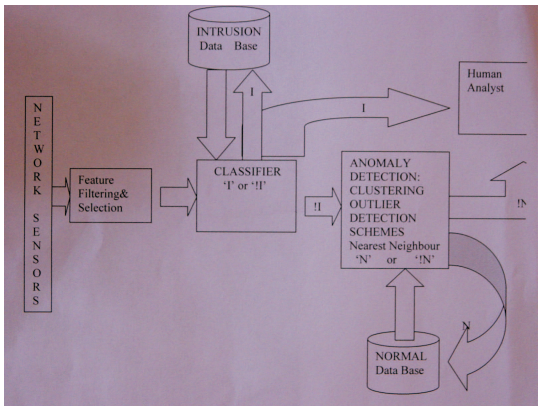

Figure 2: System Architecture.

- 1 TCP dump data is gathered from the network. The data is then filtered and the required features are extracted.
- 2 Each incident is classified as either an 'Intrusion' or 'not definitely an intrusion'.
- 3 If the incident is an intrusion the human analyst is alerted and takes appropriate action and the new incident is added to the data base of known intrusions.
  If the incident is classed as 'not definitely an intrusion' it goes through to be checked for anomalous behaviour.
- 4 In the anomaly detection phase the incidents are compared to the data base of normal incidents using a range of data mining techniques. Incidents that exceed a certain threshold are labeled 'Normal' and added to the Normal data base. Those that aren't similar enough are labeled as 'Not definitely Normal' and the human analyst is alerted to investigate the incident further.

4 METHOD

There are many different data mining techniques available such as clustering, summarization, regression, association rules and classification to name a few. The subsystem looked at in detail here is the classifer. A classification problem is when given a databse D={t1,t2,...,tn} of records and a set of classes C={c1,c2...,cm}, the problem is to define a mapping f:D->C where each ti is assigned to one class. A class cj contains precisely those tuples mapped to it.(Dunham, 2003)

The classification technique used in this instance is a decision tree. Decision trees are a predicitve modelling technique. A decision tree is a tree where the root and each internal node is labelled with a question. The arcs emenating from each node represent each possible answer to that particular question. Each leaf node represents a solution to the classification problem. Hence decision trees adopt a disjunctive normal form representation. The important difference between decision trees is to be found in their splitting criterion, how best to split the sample.

Quinlans C5.0 algorithm was used here. It is an extension from C4.5, which itself was an extension from Quinlan's earlier ID3 algorithm for constructing trees. In these algorithms the splitting criterion used have their origins in information theory.

The basic procedure followed by them is

- 1 See how the attributes distribute the instances.
- 2 Minimize the average entropy; Calculate the average entropy of each test attribute and select the one with the lowest degree of entropy.

Entropy is a measure of the impurity or uncertainity of the data. It is based on the probability of an instance on a branch being positive, where homogenous positive means Pb = 1, and homogenous negative occurs where Pb = 0. The equation for this probability meausre is shown in equation 1.

$$P_b = \frac{n_{bc}}{n_b} \tag{1}$$

where
n b, is the number of instances in branch b
n bc, is the number of instances in branch b of class c
n t, is the total number of instances in all branches

The information gained over an attribute is a measure in the reduction of entropy.

$$AverageEntropy = \sum_b (\frac{n_b}{n_t}) * [\sum_c -(\frac{n_{bc}}{n_b}) log2(\frac{n_{bc}}{n_b})] \tag{2}$$

The entropy value is normalised between 0 and 1. The lower the entropy value the purer the set. C4.5 and C5 are based on ID3 but have added functionality with C5 being much faster and a lot more memory efficient.

5 RESULTS

The system was trained on the learning dataset comprising of over 5 million records. It was then tested with the test data set containing 311029 records. In the test dataset we know what classes should be given by the tree. From this the system is able to produce a decision tree and a confusion matrix which can be seen in figure 3 and table 1 respectively. The decision tree consists of a series of quesitions at points called nodes. Following these decisions should lead to the conclusion as to whether this is normal behaviour.

count $>$ 80
:....dst_bytes $<=$ 1
: :....diff_srv_rate $<=$ 0.23: DOS
: : diff_srv_rate $>$ 0.23: PROBE
: dst_bytes $>$ 1:
: :....diff_srv_rate $<=$ 0.53: normal
: diff_srv_rate $>$ 0.53: PROBE
:count $<=$ 80
.....

Figure 3: Example of branaches on the tree

- The confusion matrix in table 1 shows us that 60302 test incidents were correctly identified by the system to be deemed as normal behaviour. These are known as true positives, the tree outputs the class correctly. These true positives follow the diagonal through the matrix with respect to each attack.
- The righthand column labelled (b) indicates that 232 normal test incidents were thought to be probe attacks. These are known as false negeatives indicating that the tree has output class normal falsely.
- The values diagonally through the matrix are known as true negatives when we are looking to the normal figures as our true positives. This show us that 3572 PROBE attacks, 223687 DOS attacks, 4 user to root and 28 remote to local attacks were correctly identified.
- All the values listed in column (a) other than the number of true positives are called false positives, that is to say the tree incorrectly classified them as not normal.

The metrics of precision and recall were calculated according to equations 3 and 4. Both a high recall and precision rate are desired. However generally as precision increases recall decreases and vice versa. From the graph in figure 4 we see a varied set of results. U2R, user to root attacks, and R2L, remote to local attacks, perform very poorly with only a few being correctly classified. Despite the fact the corpus contained many more instances of R2L attacks than it did PROBE attacks PROBE performed a lot better. Normal cases scored similarly to PROBE attacks but had a higher recall rate. DOS attacks scored excelentally almost with maximum precision and maximum recall.

$$Precision = \frac{relevantnumberretrieved}{totalnumberretrieved} \quad (3)$$

$$Recall = \frac{relevantnumberretrieved}{totalnumberrelevant} \quad (4)$$

Table 1: Precision Vs Recall

	Precision	Recall
normal	0.7347823	0.9951975
PROBE	0.7387797	0.8574172
DOS	0.9989915	0.9664384
U2R	0.0952381	0.0175439
R2L	0.2568807	0.0019195

5.1 Accuracy

The accuracy of the system is the proportion of the total number of predictions that were correct. From equation 5 the system evaluation resulted in 92.5% accuracy based on the confusion matrix in table 1.

$$accuracy = \frac{correctpredictions}{allpredictions} \quad (5)$$

6 CONCLUSION & FUTURE WORK

Various data mining techniques score differently depending upon which type of attack they are trying to detect. Some learning algorithms do well on one part of the dataset where others fail and this clearly indicates the need for hybrid learning.

This is hybrid system and holds true on this fact. The C5 algorithm produced a decision tree that is very valuable in certain attacks but not others. DOS and PROBE attacks have both scored highly. The fact that decision trees perform well with PROBE attacks means that they will play a valuable role in this hybrid system. By uncovering a probe attack early on

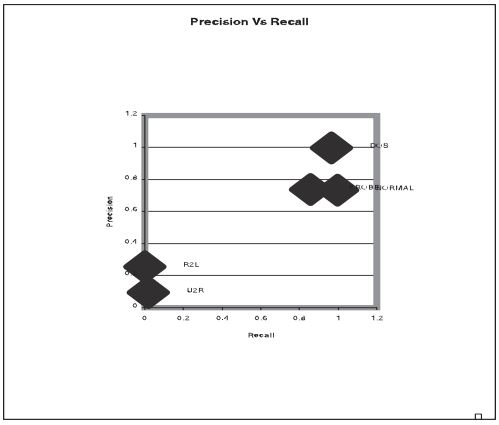

Figure 4: Preecision Vs Recall Graph.

we can prevent it from becoming anythig more than just a PROBE attack. DOS attacks should be discovered with high accuracy and a low false alarm rate.

U2R and R2L attacks however have high miss rates and these instances will need to be evaluated with other algorithms. Work is already underway with the implementaion of alternative data mining techniques such as clustering and outlier detection schemes in other parts of the system. Preliminary results of these alternatives are proving better at detecting U2R and R2L than decision trees.

Table 2: Confusion Matrix

Classified as ->	(a)	(b)	(c)	(d)	(e)	correct
normal (a)	60302	232	57	1	1	73.4
PROBE (b)	410	3572	184			73.9
DOS (c)	7630	28	223687	36	74	99.9
U2R (d)	213	4	1	4	6	9.5
R2L (e)	13558	999	1	1	28	25.7
correct	99.5	85.7	96.6	1.8	0.2	

REFERENCES

Allen, J., Christie, A., Fithen, W., McHugh, J., Pickel, J., and E., S. (2000). State of the practice of intrusion technologies. www.cert.org.

Bass, T. (2000). Intrusion detection systems and multisensor data fusion. Communications of the ACM 43(4) 99-105).

Carbone, P. (1997). Data mining or knowledge discovery in databases, an overview. Auerbach Publications.

Dunham, M. (2003). *Data Mining, Introductory & Advanced Topics*. Prentice Hall.

Fayyad, U. M., Piatetsky-Shapiro, and Symth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. Communications of the ACM 39 (11) 27-34.

First. www.first.org.

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

Kendall, K. (1999). A database of computer attacks for the evaluation of intrusion detection systems. In *ICEIS'99, 1st International Conference on Enterprise Information Systems*. MIT.

Manilla, H. (2002). Local and global methods in data mining. ICALP 2002, The 29th International Colloquim on Automata, Languages, and Programming, Malaga, Spain.

SANS. www.SANS.org.