Data Understanding Analysis for Analytical Mining IDS

Anurag Bhardwaj
Computer Sciences Corporation, India

Divydeep Agarwal
Computer Sciences Corporation, India

ABSTRACT

With the ephemeral time every information stands a greater risk of being exposed than ever before. System's security is endangered in a blink and intrusion takes place [8]. Keeping this in mind, the effectiveness of various data mining approaches are discussed. Some methods involved in classification and clustering are stated. Analysis of SVM classifier and K-means clustering is also presented. Intrusion Detection System (IDS) maintains the integrity of the system, monitors network traffic detecting potential hostile activities [6]. A case study using Snort has been done. The key idea is to study various data mining techniques and how they can be applied to IDS to maximise the effectiveness in identifying attacks, and henceforth adding to the creation of a more secured system.

Keywords

Data mining, data clustering, intrusion detection system, confusion matrix, classifier.

1. INTRODUCTION

An Intrusion Detection System (IDS) is designed to detect unwanted activities which involve manipulating or accessing computer systems through a network [6] [7]. It has become a challenging task for the network administrators. Intrusion detection systems can be: Network based, which monitors all traffic in a network or coming through an entry point such as an internet connection, and Host based, which monitors activity on a local system as a whole. Such systems are most popular target of intrusions by the attackers. There are security policies used by IDS to define an event. If any unwanted event occurs, IDS issues an alarm. It might include disabling a user account, logging off of user, etc. Hence there are three essential security functions of IDS, namely, monitoring, detecting and responding to unauthorised activity. The traffic is monitored, unwanted activities are detected and effective response is produced. But IDS should be such that it detects as many attacks possible, and number of false alarms should be least [4]. The system should be accurate in detecting attacks and should be able to handle large amount of network traffic. Figure 1 explains the working of IDS. Generally, there are two approaches taken towards network intrusion detection: misuse detection and anomaly detection. In misuse detection, the gathered information is analysed by IDS and compared with attack signatures of large databases. It looks for specific attack that has already been documented [11]. In anomaly detection, system administrator defines the baseline, network segments are monitored to compare their state to the normal baseline and look for anomalies.

Extracting various patterns from a large database where integrity of system is maintained and the data is consolidated is called Data Mining. A large flood of data is manipulated and generalised by having enough instances to validate it. There are a number of algorithms available for data mining, from the fields of statistics, machine learning, pattern recognition and databases [5]. Detection, classification, clustering and deployment are critical parts of the entire process.

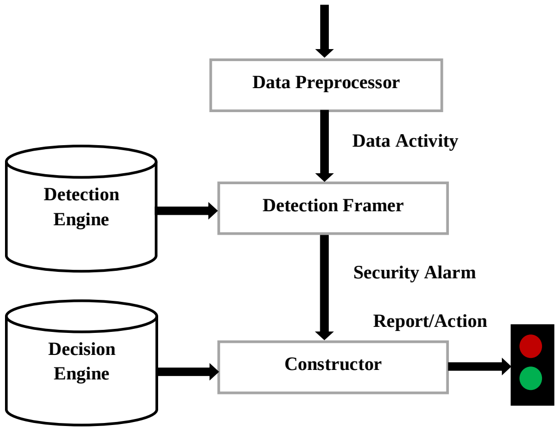


Figure 1: Working of Intrusion Detection System

2. DATA MINING APPROACHES TO IDS

2.1 Classification

Distribution of data to their respective categories as per the predefined norms set by the system administrator is known as classification. In this approach, the algorithm gives a detailed analysed output in the form of decision trees or rules. Classifier predicts unseen audit data which belongs to either normal class or abnormal class and hence decides how new records should be classified [5]. Classification is similar to clustering (section 2.2) as it also partitions records into classes. But unlike clustering, in classification it is essential for the end user to know before time how classes are defined. Each record has a value for the attribute which is used to define classes. The end user decides which attribute to use, so classification is less exploratory than clustering. Some of the most widely used methods involved are:

2.1.1 K-nearest Neighbour (KNN)

KNN is one of the most simple and traditional non-parametric technique to classify samples [14] [15]. Given an input vector, KNN calculates the approximate distances between the vectors and then assign the points which are not yet labelled to the class of its K-nearest neighbours. K is an important parameter to be considered and the performance depends upon the values of K. Larger the K, the neighbours involved in classification will take large amount of time and thus accuracy of the result will be affected.

2.1.2 Support Vector Machines (SVM)

SVM was proposed by Vapnik in 1998 [13]. SVM works as follows: (i) Mapping of input vector into high dimensional feature space, (ii) Optimal hyper-plane is obtained, (iii)

Decision boundary is determined by support vectors. SVM basically aims at separating training data sets which belong to different classes [3].

2.1.3 Decision Tree (DT)

It is a well-known machine learning technique which is composed of 3 basic elements [1]: a decision node specifying a test attribute; an edge corresponding to one of the possible attribute values; and a leaf known as answer node, containing the class to which the object belongs.

Analysing the previous works done in the classification of data, we found out the statistical data as mentioned in figure 2. It can be observed that SVM is the most commonly used classification technique.

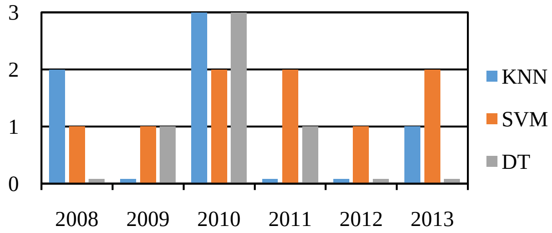


Figure 2: Total number of articles using algorithms

2.2 Clustering

The process of labelling data and then assigning it into groups is called as clustering. Since the fraction of network audit data instances available is huge, if humans do the labelling, it will be time consuming and expensive. Hence, idea of clustering comes into picture. These groupings increase the performance of existing classifiers [7] [10]. If a cluster contains instances of data from one category only, then it is said to be 100% pure. There are two types of clustering techniques: pairwise clustering and central clustering. Pairwise clustering combines data instances which are similar based on distances between them in pairwise fashion whereas in central clustering models the focus is on each of the cluster centroid. Centroid clustering is more efficient than pairwise clustering. A single variant of single linkage clustering can be used to create clusters, but it is not so efficient, only advantageous as it works in linear time. The system is ready to perform intrusion detection once the clusters are created.

2.2.1 K-means Clustering

K-means is the most famous and fundamental in clustering algorithms. Because the performance is quite stable, many research experiments treat the results from K-means as baseline. Other clustering algorithms, such as Gaussian Mixture Models (GMMs), usually employ K-means' product as initialized value.

Let X = $\{x_1,...,x_N\}$ be a d-dimensional observed dataset of N vectors, S = $\{s_1,s_2,...,s_K\}$ be the set of K clusters and $\mu = \{\mu_1,...,\mu_K\}$ be a set of d-dimensional vectors which represents the mean of every cluster, where $1 < K \leq N$.

Furthermore, let R = $\{r_{11},...,r_{NK}\}$ be a N × K matrix, where $r_{nk} = 1$ when x_n is a member of s_k, otherwise $r_{nk} = 0$. K-means is based on a simple criterion x_n belongs to the cluster with nearest mean. Therefore, K-means aims to partition N observations into K clusters minimizing the function given by (1).

$$J(R, \mu, X) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\| \qquad (1)$$

Or more formally,

$$arg\, min_k\ J(R, \mu, X) \qquad (2)$$

3. ANALYSES OF ALGORITHMS USED

3.1 Analysis of SVM Classifier

Support Vector Machine was employed for classification of 3 different categories of data and a confusion matrix was obtained as shown in figure 3. It can be observed that with 85.33% accuracy, the method is able to categorise different data elements. Also, class to class relationship can be obtained using confusion matrix. From figure 3, data 1 is related to data 2 and data 3 by membership degree 0.15 and 0.15 respectively. Similarly, one can observe membership degrees between other data elements.

SVM algorithm implementation is based on the LIBSVM which is a MATLAB toolbox written by Chih-Chung Chang and Chih-Jen Lin. In addition, all experiments were conducted using MATLAB 2010B for 64-bit.

	Data 1	**Data 2**	**Data 3**
Data 1	0.85	0.15	0.00
Data 2	0.15	0.80	0.05
Data 3	0.10	0.00	0.90

Figure 3: Confusion matrix with accuracy 85.33%

3.2 Analysis of K-means Clustering

K-means is used to find natural groupings of alarm records which are similar in one way or the other. The records which are far off from any of these clusters indicate that some unwanted event has occurred, which may be a part of network attack. Following steps are involved [2] in detecting intrusion:

1. Find the largest cluster which has most number of data instances and label it as normal.

2. Remaining clusters are to be sorted in increasing order of their distances to the largest cluster.

3. Select the first cluster K1 such that number of data instances in these clusters sum up to ¼ `N

4. Label them as normal, where ` is the percentage or normal instances.

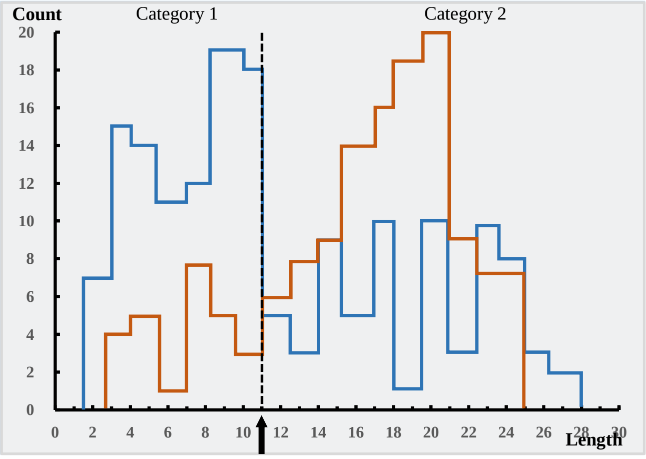
5. Label all the other remaining clusters as attacks.

Figure 4(a) and (b) are the histograms showing the difference in resulting cluster groups when different features are considered.

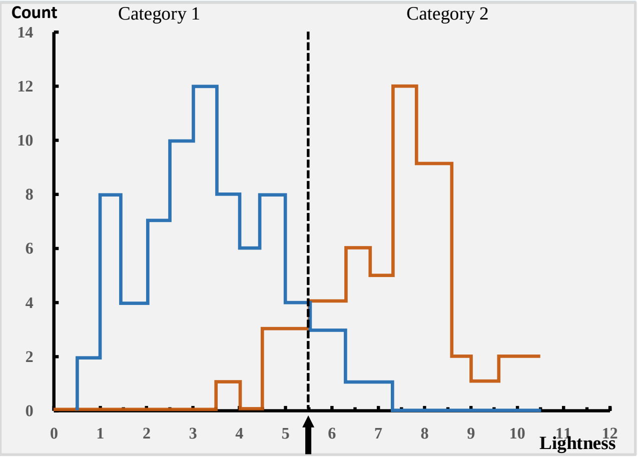
4. CASE STUDY

We used a tool called Snort to detect attacks. The traces describe the event according to its port address as well as source and target host address. Anomaly based methods were deployed.

HISTOGRAMS OF THE LIGHTNESS FEATURE FOR THE TWO CATEGORIES



Leads to the smallest number of errors on average
We cannot reliably separate category 2 from category 1 by length alone.



Leads to the smallest number of errors on average
The two categories are much better separated.

Figure 4(a): Feature oriented histogram comparison (Length feature)

Figure 4(b): Feature oriented histogram comparison (Lightness feature)

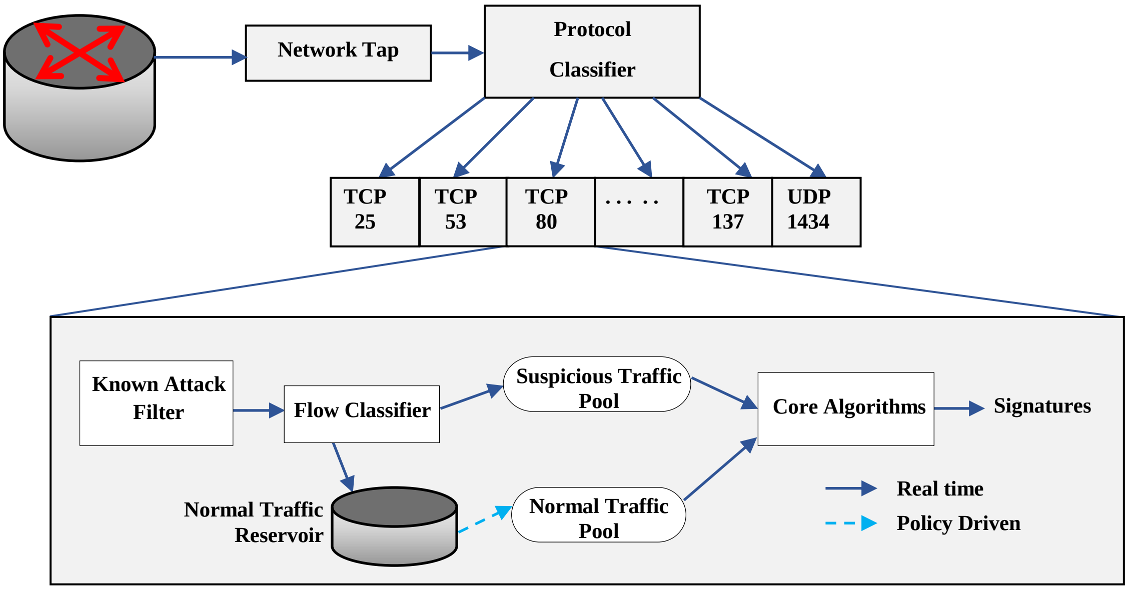


Figure 5: Snort Tracing

4.1 Event Traces

We collected data from multiple sites as well as multiple attacking hosts.
http://www.yahoo.com

June 25 2013 14:20:16 DNS port probe 206.13.28.60

http://www.sans.org/y2k/052300-0802.htm

June 25 15:14:12: 8080 Attempt: 206.13.28.60:65535 -> 192.168.1.103:8080

June 25 15:14:13: 8080 Attempt: 206.13.28.60:65535 -> 192.168.1.105:8080

June 25 15:14:14: 8080 Attempt: 206.13.28.60:65535 -> 192.168.1.110:8080

June 25 15:14:15: 8080 Attempt: 206.13.28.60:65535 -> 192.168.1.109:8080

4.2 Sources of Trace

In order to collect traces for the attacker's addresses, a number of websites were searched. Google’s search engine provided us with the desired information.

4.3 Probability of Spoofing

The possibility of spoofing is high, since from the collected traces a complete TCP handshake cannot be figured out. However, further investigation is required. It is important to have a detailed investigation done on the scanned sites and determine the nature of the target. One can contact the administrator and obtain access to attack tools and then create attack signatures to catch events associated with these scans.

4.4 Attack Description

The attack involves the scanning of various DNS host addresses and web proxy services. First of all, DNS is scanned to identify vulnerabilities present in DNS software. Then web proxies are scanned so that one can gain access to the proxy server. The traffic was TCP-based, port 65535, with same TCP sequence number and IP ID number for all the packets. Under normal conditions, whenever a new packet is sent on the network, the TCP sequence number changes.

4.5 Severity

Severity can be calculated using following formula:

Severity = (Target Criticality + Attack Lethality) - (System Countermeasures + Network Countermeasures)

The severity is measured on a five point scale, 5 being the highest and 1 being the lowest.

5. PERFORMANCE EVALUATION

Data understanding requires the number of records to be large and sparsely populated. Data clustering has the advantage of working in linear time. However it is difficult to find relationships between different features of a single record [11]. Classification, like clustering, partitions records into classes. But unlike clustering, it is essential for the end user to know before time how classes are defined. Each record has a value for the attribute which is used to define classes. The future IDS must be vulnerability and scenario based [12]. Being adaptive will be crucial for zero-day attacks. It should be able to correlate multiple information from multiple sources.

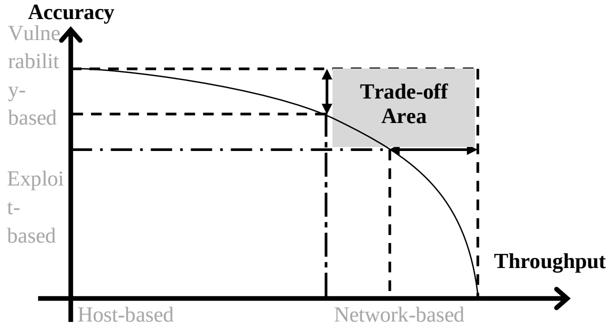


Figure 6: Next Generation IDS

6. CONCLUSION

Data understanding for analytical mining IDS can monitor large volume of data at high speed. In this paper, effectiveness of various data mining approaches, methods of classification and clustering have been stated [9] [10]. Confusion Matrix for SVM classifier and histogram aftermath for K-means clustering has also been conferred along with a Snort study. Intrusion Detection System can be implemented using a number of approaches. Training of data can be done with fewer difficulties if it is efficiently classified and clustered. Mechanisms can be kept distinct from policies so as to have broader detection coverage. Every approach has its own advantages and disadvantages. Therefore the decision to choose a particular technique to implement the next generation IDS is a critical task. There still lies a wide area of research for future work and creation of a fool proof secured system.

7. REFERENCES

[1] Kaushik Sapna and Deshmukh P.R., Comparison of approaches to implement intrusion detection system, International Journal of Computer Science and Communication, vol. 2, no. 1, pp. 45-48, Jan-Jun 2011.

[2] Chang-Tien Lu, Arnold P. Boedihardjo, Prajwal Manalwar, Exploiting efficient data mining techniques to enhance intrusion detection systems, pp. 512-517, IRI 2005.

[3] Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin, Wei-Yang Lin, Intrusion detection by machine learning: A review, Expert Systems with Applications, 36. Jg., Nr. 10, 2009.

[4] T. Abraham, IDDM: Intrusion Detection using Data Mining Techniques, DSTO-GD, 2008.

[5] Rakesh Agrawal and Ramakrishnan Srikant, Privacy-preserving data mining, In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00), ACM, New York, NY, USA, 439-450, 2000.

[6] Min Qin and Kai Hwang, Anomaly Intrusion Detection by Internet Data mining of Traffic Episodes, ACM, TISSec, March 1, 2004.

[7] Alok Ranjan, Dr. Ravindra S. Hegadi, Prasanna Kumara, Emerging Trends in Data Mining for Intrusion Detection, International Journal of Advanced Research in Computer Science, vol. 3, no. 2, March-April 2012.

[8] Bhavani Thuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen, Data Mining for Security Applications, IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, 2008.

[9] Han Jiawei and Kamber Micheline, Data Mining: Concepts and Techniques, 2nd edition, San Francisco, Morgan Kaufmann Publishers, 2006.

[10] Li Bo, Jiang Dong-Dong, The Research of Intrusion Detection Model Based on Clustering Analysis, IEEE International Conference on Computer and Communications Security, 2009.

[11] A.K Maheshwari, Association Rule in Data Mining for Large Transactional Database, IJMIE, vol. 2, issue 2, pp. 358-380, March 2012.

[12] Dr. Sugumar Rajendran, Dr. Rengarajan Alwar, Dr. Saravanakumar Selvaraj, Determining the Existence of Quantitative Association Rule Hiding in Privacy Preserving Data Mining, IJARCCE, vol. 1, issue 2, April 2012.

[13] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[14] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.

[15] Manocha, S. and Girolami, M. A., 2007, An empirical analysis of the probabilistic K-nearest neighbour classifier, Pattern Recognition Letters, vol. 28, 1818-1824.