ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering
(IJERCSE)
Vol 4, Issue 3, March 2017

Host-Based Intrusion Detection System Using Analytics

[1] G.Yedukondalu, [2] Dr.J.Anand Chandulal, [3] Dr.M. Srinivasa Rao
[1] Vignan Institute of Technology& Science,Vignan Hills,Deshmukhi, Hyderabad, INDIA,
[2] K.L University,Vijayawada, Andhra Pradesh, INDIA
[3] School of Information Technology, Jawaharlal Technological University Hyderabad , INDIA

***Abstract* - File signatures are computed to improve the efficiency and effectiveness of the Intrusion Detection System. File Signatures are generated using Hashing Method and Superimposed Coding technique. This paper discusses the techniques that works fast and efficiently in detecting the malicious users. DARPA data set is used to apply these techniques to find out the intruders through IDS. The performance of the similarity search algorithm is efficient since all the signatures are in the binary format and computations are done by low level logical operations[1]. Clustering and Similarity search techniques are applied to increase the efficiency of the Host-Based Intrusion Detection System.**

Index Terms:— **File Signatures , Intrusion Detection System, Hashing Method, Superimposed Coding Technique, Similarity Search.**

I. INTRODUCTION

Intrusion Detection System(IDS) is the process of observing the events taking place in a computer system or network and analyzing them for signs of intrusion. It is useful for not only detecting successful intrusions, but also in monitoring attempts to break security, which provides important information for timely counter-measures. Intrusion Detection System can be categorised into two types: Misuse Intrusion Detection and Anomaly Intrusion Detection. Ancient protection techniques such as authentication, data encryption, avoiding programming errors, and firewalls etc., used as first line of protection for computer security. If a weak password is compromised, user authentication cannot prevent unauthorized use. Also, firewalls are very vulnerable to errors in system configuration and susceptible to ambiguous or undefined security policies. In such networks, even if an intrusion is detected, the system cannot be shut down to check it fully since it may be serving users who are making deals or completing one transaction or the other [1]. An IDS inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. Generally, an IDS detects unwanted manipulations of computer systems, mainly through the Internet. The manipulations may take the form of attacks by crackers. Section2 gives a brief survey on anomaly-based schemes to understand different approaches to IDS. Section 3 clustering and Section 4 presents the proposed system. We conclude our work in section 5 with experimental results.

Misuse Detection vs. Anomaly Detection

The Misuse Detection part analyses the information it gathers, and compares it to large databases of attack signatures by looking for a specific attack that has already been documented while the Anomaly Detection part monitors network segments to compare their state to the normal baseline defined by the systems administrator and look for anomalies [4].

Network-Based vs. Host-Based Systems

In the Network-based system, the network analyses individual packets of information flowing through it and detects those that are malicious but designed to be overlooked by a firewall's simplistic filtering rules. In a Host-based system, the IDS examines the activity on each individual computer or host[4].

II RELATED WORK:

Anomaly-based Intrusion Detection has the capability to identify new attacks, as any attack will differ from normal activity. However, such systems have a very high rate of false positives[2]. Hence, a lot of research being done in area of anomaly-based intrusion detection. Sanjay Rawat et al[3] proposes an approach that captures users' behavior Singular Value Decomposition Technique for fast intrusion detection system. Sanjay Rawat et al[4] proposes another approach BWC metric for anomaly-based intrusion detection scheme that rely on using sequence of system calls. This approach improves the capability of the k-nearest neighbor classifier significantly. Sanjay Rawat et al proposes a fast Host-Based Intrusion Detection System using Rough Set theory published in Springer – Verlog(2005). Dash and G. Vijaya Kumari et-al proposes well-known framework IA network is employed for the detection purpose[5]. In their paper they purposed a novel masquerade detection method based on constraint satisfaction problem.

III. CLUSTERING

Clustering is a process of grouping data of similar objects. Each group, called a cluster. The features of each object in a cluster is mostly similar. The object patterns within a cluster are very similar to each other than a pattern belong to a other cluster. We use K-means algorithm to cluster the DARPA data set.

K-means is a well-known grouping technique here Objects are classified as belonging to one of k groups, k chosen a priori.

Suppose, here k is considered as 3. So three clusters are created.

Pseudo code of the k-means algorithm is to explain:

X Considered K as the number of clusters.

Y Initialize the codebook vectors of the K clusters randomly, for example.

Z For every new sample vector:

Z1 Calculate the distance between the new vector and every cluster's codebook vector.

Z2 Recalculate the nearest codebook vector with the new vector; using a learning rate that decreases in time.

The factors behind choosing the k-means algorithm because of its popularity for the following reasons:

Its time complexity is O(nkl), where n is the number of patterns, k is the number of clusters, and l is the number of iterations taken by the algorithm to converge.

Its space complexity is O(k+n). It requires additional space to store the data matrix.

IV PROPOSED SCHEME: SIGNATURE GENERATION AND INTRUSION DETECTION

The Data Mining techniques are best suited for intrusion detection to trigger alaram when any intrusion takes place in a host system.The proposed IDS can thought of decision making using DARPA data set. Clustering and file signature techniques are applied on DARPA to identify the anomalus actions. The proposed scheme is shown in Fig. 1.

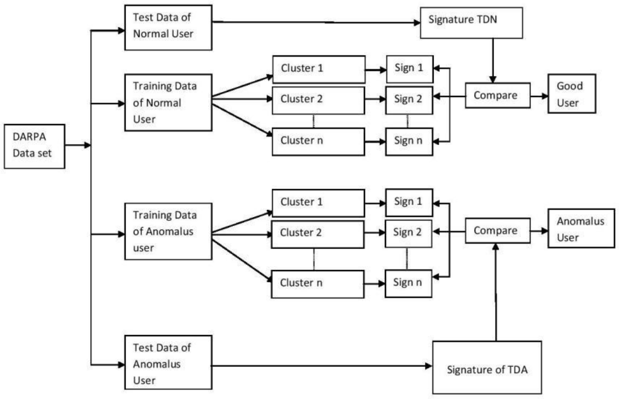


Fig. 1 Framework for host-based anomaly detection

The DARPA Data set consists of test data, training data and attacks. This data set used in the detection process of intruders. Broadly the DARPA data set consists of test data of normal users, test data of anomalous users, training data of normal and anomalous users. Each data set is clustered using K-Means algorithm. For each cluster the signature is calculated. Test data signature will be matched with signatures of the shortest distance cluster. Remaining clusters will be omitted for searching. Because of this approach the processing time will be effectively reduced. So the efficiency of ID system will be increased. The signature of test data of normal users will be compared with signatures of the training data of normal users . That user will be allowed to access the Host system. Likewise if the match found in the clusters of training data of anomalous users with the signature of test data of anomalous user. Then the ID system will give anomalous user as outcome.

4.1 FILE SIGNATURE

File signature method is an efficient technique for text retrieval. File signatures are computed using Hashing and superimposed coding technique.

Hashing Algorithm:

Input: n size of signature, r number of bits set to 1, Word[] word whose signature is to be computed.

Output: s signature of Word[]. procedure Hash (n,r,word[]: in; s:out)

```
Step1: H(word)=0; l=length of the word[]; p=nCr;
Step2: for i=I to l do
Step3:H(word)=H(word)*2+ASCII(word[l]);
Step4:End do
Step5:s=H(word) mod p
Step6:End
```

The following example illustrates the above algorithm. Let us suppose that n = 4 and r = 2. Then all the possible (4 C2

) combinations for the word DATA and it's signatures are shown in table I

Table 1.

Hash Value	Signature
0	0011
1	0101
2	0110
3	1001
4	1010
5	1100

The Hash value of word "DATA" is 1037 as per the above algorithm and the signature of word DATA is 1037 mod 6. The resultent signature is integer 5 and its corrensponding binary value is the signature of the word DATA. So the signature of the word "DATA" = 1100. DARPA data set consists of total 606 text documents.Each document consists of sequence of system calls. Superimposed coding technique is used to compute the signature of the file. The computational steps involved are shown below:

```
Input: Doc document consists of k words w1, w2 ..., wk.
Output: S signature of the document.
procedure superimposed-coding(.Doc : in; S : out}
1.for i = 1 to k do
2.si — Hash(n,r, wordi[]);
3.end do
4.S= sl V s2 V ... V sk
5.End.
```

For Example if a book by title "Computer Applications " consists of three keywords,
computers 2. applications 3. mathematics. If n = 12 and r = 4 the signatures of the keywords are as shown in the table. The signature of the book is obtained by superimposing the signatures of the keywords with OR operation was shown in below TableII

Table II

computer	1100	1000	0100
applications	0001	0101	0100
mathematics	0011	0001	1001
signature of the book	1111	1101	1101

After file signature generation of DARPA data set over then it is devided into 3 groups using k-means algorithm. The distance between the test data sample to the random sample of cluster is computed using Jaccard Distance.The distance of the test sample is near to whatever the cluster The test data sample is matched with random samples from a particular cluster based on the distance.

Jaccard Similarty Measure:

The distance between two binary strings computed using Jaccard Similarity Measure using the below Equation

$$d = \frac{r+s}{q+r+s+t}$$

Here q=no.of variables that equal 1s for both signature and test string

r=no.of variables that equal 1 for signature and 0 for test signature

s=no of variables 0 for train signature and 1 for test signature

t=no.of variables that are equal 0 for both strings

For example:

Original String = 11010

Test String = 01110

Here

q=2, r=1, s=1, t=1

so, d=1+1/(2+1+1+1)2/5=0.4

The distance between above two strings is 0.4.

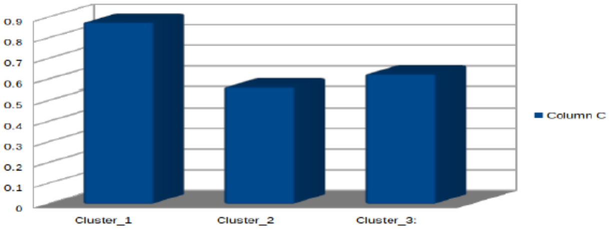
V EXPERIMENTAL RESULTS

DARPA dataset consists of total 606 text files. File Signatures are generated using java .For example some of the file signatures generated using Hash funtion and Superimposed Coding Technique is shown in below Table III

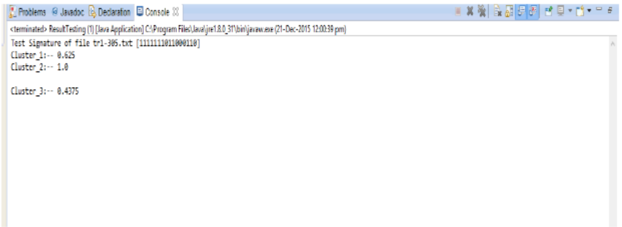
Table III

File Name	File Signature
The sigunature of file tr1-311.txt	1110011001010101
The sigunature of file tr10-320.txt	1111010111101110
The sigunature of file tr100-816.txt	0000101010001111
The sigunature of file tr101-818.txt	1110011000000101

k-means algorithm applied on binary DARPA signatures dataset. The clusteing outputs are shown below:



The distances are measured using Jaccard distance technique to find out the distances between each cluster. Distance between test signature to corresponding clusters are clacutated are shown below. The test signature of the file is very near to cluster 3 because its distance is short when compared with other clusters. So this tecchinque will eleminate the first two clusters, obviously we search in cluster 3 to findout intruder.So the system is more efficient with respect to serch time, the expermental results was shown in Fig2.



```
<terminated> ResultTesting (1) [Java Application] C:\Program Files\Java\jre1.8.0_31\bin\javaw.exe (21-Dec-2015 12:00:39 pm)
Test Signature of file tr1-305.txt [1111111011000110]
Cluster_1:-- 0.625
Cluster_2:-- 1.0

Cluster_3:-- 0.4375
```

The test signature is verified in cluster 3 for pattern matching. This is matched with anamalous users database.So the test signal is an intruder which is shown in the below TableIV

Table IV

File Nmae	Signature	Result
Test Signature of file tr1-305.txt	1111111011000110	Intruder Detected

VI CONCLUSION

In this paper, the file signature is computed. It is a feasible framework and tries to explore a new approach to intrusion detection system. IDS data processing speed will be high because of the suitable clustering technique.

REFERENCES

[1] Conference on Fuzzy Systems, 2004, pp. 691-696. Sanjay Rawat, "On the use of Singular Value Decomposition for Fast Intrusion Detection System" In Proceedings- published in Electronic Note in Theoretical Computer Science URL:www.elsevier.nl/locate/entcs.

[2] Sanjay Rawat, "Intrusion Detection System using text processing with Binary-Weighted Cosine Metric ", In Proceedings: published in Eelectronic Notes in Theoretical Computer URL:www.elsevier.nl/locate/entcs.

[3] Subrat Kumar Dash, Sanjay Rawat, G. Vijaya Kumari and Arun K. Pujari, "Masquerade Detection Using IA Network", First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008

[4] Hind Tribak , Blanca L. Delgado-Marquez, P.Rojas, O.Valenzuela, H. Pomares and I. Rojas, " Statistical Analysis of Different Artificial Intelligent Techniques applied to Intrusion Detection System", IEEE, 2012

[5] S. Revathi and A. Malathi, "Data Preprocessing for Intrusion Detection System using Swarm Intelligence Techniques", International Journal of Computer Applications , Volume 75– No.6, August 2013 [23] Iwan Syarif, Adam Pruge Bennett and Gary Wills, "Unsupervised clustering approach for network anomaly detection.

[6] Faloutsos.C."Access methods for text" , *ACMComputing Surveys.*1985,

[7] Sreenivasa Rao, M., Pujari, A. K., Sreenivasan, B."A new neural network architecture for efficient close proximity match of large databases*". IEEE Computer Society Press, Proceedings of the Eighth International Workshop on DEXA, France,Edited by R. R. Wanger*, 444-449, 1997.

[8] S. B. Needleman and C.D. Wunch."A general method applicable to the search for similarities in the amino acid sequences of two proteins.Journal of Molecular Biology ", 1970.

[9] Shang ,H. ,Merrettal ,T. H.,"Tries for Approximate String Matching knowledge*", IEEE trans on ge and data Engineering* ,1996.

[10] Bethina Schmitt and