Host-based Anomaly Detection in Digital Forensics using Self Organizing Maps

Sushilkumar Chavhan
Dept. of I.T,
Y.C.C.E., Nagpur

Smita M. Nirkhi
Dept of C.S.E.
R.C.O.E.M., Nagpur

R. V. Dharaskar, Ph.D
Director,
M.P.G.I. Nanded, MS, India

ABSTRACT

Anomaly detection techniques are widely used in a number of applications, such as, computer networks, security systems, etc. This paper describes and analyzes an approach to anomaly detection using self organizing map classification. We deal with the massive data volumes with the dynamic nature of day to day information networks. So it's difficult to identify the behavior of system. Visualization of data has ability to take into a massive volume of data. In digital forensics self organizing map has high potential handle large data and observe the behavior of computer. This paper provides an overview of anomaly detection system which able to handle massive real data.

Keywords

Digital forensic, self organizing map (SOM), anomaly detection, visualization

1. INTRODUCTION

The detection of unusual behavior patterns is an important problem in computer security as most security breaches exhibit anomalous system behavior. Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior [3].Anomaly detection techniques are applied in a variety of domains, including credit card fraud prevention, financial turbulence detection, virus or system intrusion discovery, and network monitoring, to name a few. By observing various data sets and activities, the anomaly detection systems can classify the behavior and determine if it is either normal or anomalous. Unlike signature-based cyber security systems, which can only detect attacks for which a signature has previously been created, anomaly detection is based on behavioral patterns [19].

Computer forensics deals with the identification, extraction, preservation and documentation of digital evidence [12]. Digital evidence may be sought in a wide range of computer-related crimes. What is unique about digital evidence is the fact that it is fragile by nature and can easily be altered or destroyed. [17]. Computer forensic tools have been developed to assist computer forensic investigators in conducting a proper investigation into digital crimes. Computers are being attacked and compromised on a daily basis. These attacks are made to steal personal identities, to bring down an entire network segment, to disable the online presence of businesses, or to completely obliterate sensitive information that is critical for personal or business purposes. The data analysis can be enhanced through visualization, which can be used in recognizing anomalous trends or patterns. Human observers can quickly detect anomalies in huge volumes of data when the data is represented graphically in three dimensions or in a user selectable, multidimensional space. Visualization can be used not only to present the final results of analysis to decision-makers, but in the knowledge discovery process itself [10].

Data mining techniques have unlimited potential in the field of forensic science where models and tools can be developed to help investigators, digital forensics professionals and law enforcement officers to find the data or clues they are searching for much more efficiently and faster. Appropriate data mining techniques include support vector machine learning algorithm, behavior based anomaly detection, and heuristic-based anomaly detection [19].

Visualization represents a powerful link between the most dominant information-processing systems, the human brain and the modern computer [20]. The process of identifying specific subsets of information from a large mass is one of data-mining and unsupervised learning is a particularly relevant technique for achieving this. The techniques of unsupervised learning allow sets of information to be extracted and grouped together, potentially from disparate sources, without the need to identify the key characteristics of the different groups beforehand. [18]. One method of creating these clusters is by using self organizing maps (SOM). SOMs have been used by many researchers in the field of forensics before. It has been identified as a potential tool for use by computer forensic investigators as well as being used to detect unusual behavior of systems or networks [21].

The rest of the paper is organized as follows: Section 2 focuses on literature review on anomaly detection. Section 3 discusses about the working of Self Organizing Map. Section 4 gives the detailed for the working of anomaly detection techniques. Finally, section 5 concludes the paper.

2. RELATED WORK

Most of forensics researcher detects anomalies in networks based system. Dipankar Dasgupta et.al [4], provides with the system to detect the anomaly in the data by considering the essence personnel data like natural immune system. The system designed is capable of comparing the single and multidimensional data sets. Further the work is extended to design a new System which uses the negative selection algorithm developed by Forrest. Li Yao et.al [5], to improve the performance of above algorithm and also developed fuzzy anomaly detection model for IPv6, using fuzzy detection anomaly algorithm. Also the system is capable of detecting most of IPv6 attack. Jun Lv et.al [6], suggested a new algorithm to solve the network traffic anomaly detection problem. This system works by combining Generalized Likelihood Ratio algorithm and wavelet method, and capture the failure point in real time. Jinquan Zeng et.al [14], proposed approach which uses the feedback technique, which adjusts the self radius of self elements, the detection radius of detectors and the number of detectors, to adapt the varieties of self/nonself space and build the appropriate profile of the system based on some of self elements. Also the system improves the accuracy in solving the anomaly detection problem. Nan Zhang et.al [7], addressed how to establish the defender's reputation in anomaly detection against insider attacks. With the help of the present scenario, system proposed on two generic reputation-establishment algorithms for systems consisting of only smart insiders and also those with both smart insiders and naive attackers. Ning Chen et.al [8], gives an anomaly detection and analysis method based on correlation coefficient matrix. Further the system designed discovers the anomaly behaviors in the TCP flows and their types by the variety of correlation coefficients between observed packets, consequently implements network health checking and anomaly behavior detection and analysis. Chi-Yuan Chen et.al [11], proposed an anomaly detection scheme for an IP multimedia subsystem. The system is having an additional features which includes SVM vectors from transaction patterns in the IMS core a transaction-pattern-based anomaly detection algorithm for the high-order Markov-based SVM kernel. Chee-Wooi Ten et.al [12] contribution a new substation anomaly detection algorithm that can be used to systematically extract malicious “footprints” of intrusion-based steps across substation networks. Zhe Yao, et.al [13] proposes a framework for anomaly detection using proximity graphs and the Page Rank algorithm which work on an unsupervised, nonparametric, density estimation-free approach. Various parameter selection, time complexity guarantees, and possible extensions are discussed and investigated during the research.

3. SELF ORGANIZING MAP

The Self Organizing Map is one of the most widely used neural network models. SOM is used to map high-dimensional data onto a low-dimensional space, typically two-dimensional, while preserving the topology of the input data i.e. place similar data in the input space are placed on nearby map [1]. The SOM is used to review interesting patterns. SOMs are used to help investigators to get a visual snapshot of a hard drive enabling one to make better decisions on were to focus a digital forensic examination on a large disc. By doing this examiner can conduct the forensics analysis process more efficiently and effectively [19].

The SOM consists of two layers of neurons i.e. the input layer and the output layer. The input layer is fully connected with nodes at the output layer and each neuron in the input layer represents an input signal [17]. The output layer generally forms a two-dimensional grid of neurons where each neuron represents a node of the final structure. The connections between the layers are represented by weights whose values represent the strength of the connection. During the learning process, when an input pattern is presented to the input layer, the neurons in the output layer will compete with one another. The winning neuron will be the one whose weights are the closest to the input pattern in terms of Euclidian distance [1]. Once the winning neuron has been determined, the weights of the winning neuron and its neighborhood will be updated, i.e. shifted in the direction of the input pattern. After the learning process, the SOM configures the output neurons into a topological representation of the original data, by means of a process called self-organization [18].

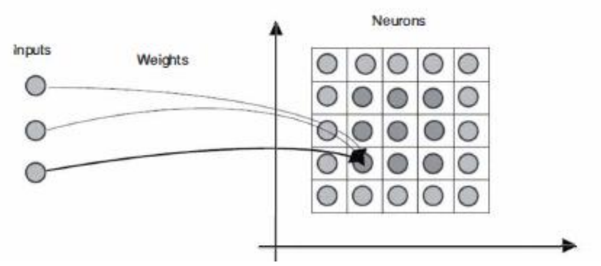


Figure 1: Structure of SOM [10]

4. ANOMALY DETECTION

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior [3].The basic idea is to first collect the data from computer system. Preprocess that in specified format for anomaly detection. Discover various pattern of each preprocessed data using self organizing map. Analyze this file for anomaly detection. Figure2 shows the flow of anomaly detection.

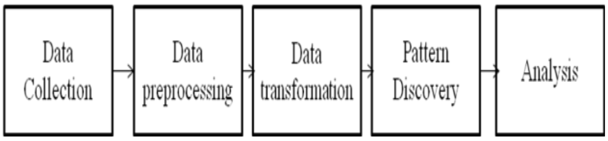
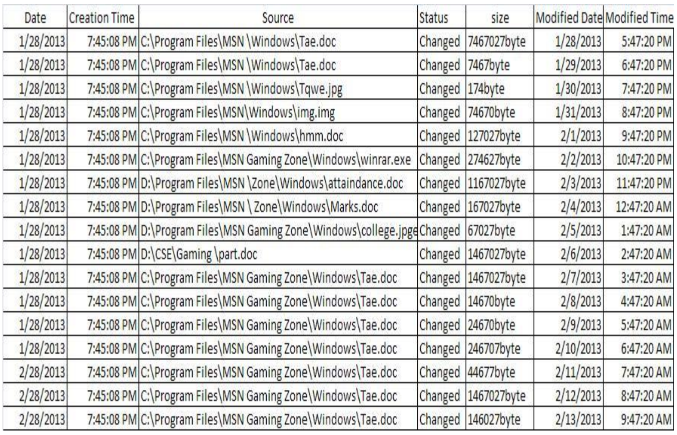


Figure 2: process of anomaly detection

4.1 Data Collection

We have collected the log files of the various host systems. These logs include system logs, application logs and security logs present in control panel of hosts computer system. Application logs gives information about the various application present in computer system. Security logs give the information about is a log that contains records of login/logout activity done in computer system. System logs contain events that are logged by the operating system components. The main challenge about data collection is that collect logs of file related data. But these logs are insufficient for the anomaly detection. Figure 3 shows collected application logs along with the file related details.



Date	Creation Time	Source	Status	size	Modified Date	Modified Time
1/28/2013	7:45:08 PM	C:\Program Files\MSN \Windows\Tae.doc	Changed	7467027byte	1/28/2013	5:47:20 PM
1/28/2013	7:45:08 PM	C:\Program Files\MSN \Windows\Tae.doc	Changed	7467byte	1/29/2013	6:47:20 PM
1/28/2013	7:45:08 PM	C:\Program Files\MSN \Windows\Tqwe.jpg	Changed	174byte	1/30/2013	7:47:20 PM
1/28/2013	7:45:08 PM	C:\Program Files\MSN\Windows\img.img	Changed	74670byte	1/31/2013	8:47:20 PM
1/28/2013	7:45:08 PM	C:\Program Files\MSN \Windows\hmm.doc	Changed	127027byte	2/1/2013	9:47:20 PM
1/28/2013	7:45:08 PM	C:\Program Files\MSN Gaming Zone\Windows\winrar.exe	Changed	274627byte	2/2/2013	10:47:20 PM
1/28/2013	7:45:08 PM	D:\Program Files\MSN \Zone\Windows\attaindance.doc	Changed	1167027byte	2/3/2013	11:47:20 PM
1/28/2013	7:45:08 PM	D:\Program Files\MSN \ Zone\Windows\Marks.doc	Changed	167027byte	2/4/2013	12:47:20 AM
1/28/2013	7:45:08 PM	D:\Program Files\MSN Gaming Zone\Windows\college.jpge	Changed	67027byte	2/5/2013	1:47:20 AM
1/28/2013	7:45:08 PM	D:\CSE\Gaming \part.doc	Changed	1467027byte	2/6/2013	2:47:20 AM
1/28/2013	7:45:08 PM	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	Changed	1467027byte	2/7/2013	3:47:20 AM
1/28/2013	7:45:08 PM	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	Changed	14670byte	2/8/2013	4:47:20 AM
1/28/2013	7:45:08 PM	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	Changed	24670byte	2/9/2013	5:47:20 AM
1/28/2013	7:45:08 PM	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	Changed	246707byte	2/10/2013	6:47:20 AM
2/28/2013	7:45:08 PM	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	Changed	44677byte	2/11/2013	7:47:20 AM
2/28/2013	7:45:08 PM	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	Changed	1467027byte	2/12/2013	8:47:20 AM
2/28/2013	7:45:08 PM	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	Changed	146027byte	2/13/2013	9:47:20 AM

Figure 3: Collection of Data

For anomaly detection with visualization using anomaly detection require data which gives information about application ,size ,date, time, creation time.

- Application gives information about which application is trying to modify, change or create.
- Date and time is the current accessing time and date at which that file or application being to be used.
- Creation time is time when that file is created.

Collected data may contain several unnecessary fields. In order to detection of anomaly, few fields that are prominent would be considered. Hence, we have to preprocess the data. Figure 3, shows the preprocessed data.

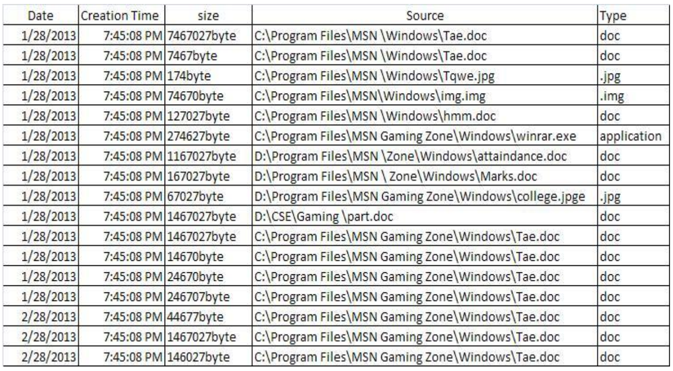
4.2 Data Preprocessing

Data preprocessing is the important step in data mining. This includes data cleaning, Data integration, Data transformation, Data reduction, Data discretization [20]. The work of Data cleaning is to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

To obtain the desired filed, we must clean data in order to remove the unwanted field from the collected logs file. The data cleaning algorithm used is detailed as follow:

```
    Field1={‘application’,’size’,’date’,’time’,’creationdate’}
    Field2= remaining fields from logs
Begin
1. Read records in logs
2. For each record in logs
3. Read fields
4. If Field1=’TRUE’
Then
5. Extract Field1 and ‘SAVE’
6. else
Remove Field2 and ‘SAVE’
7. Display Logs
8. Next record
End if
End
```

After applying this algorithm on collected log files the output is shown as in Figure 4 which gives only selected fields which mention in Field1.



Date	Creation Time	size	Source	Type
1/28/2013	7:45:08 PM	7467027byte	C:\Program Files\MSN \Windows\Tae.doc	doc
1/28/2013	7:45:08 PM	7467byte	C:\Program Files\MSN \Windows\Tae.doc	doc
1/28/2013	7:45:08 PM	174byte	C:\Program Files\MSN \Windows\Tqwe.jpg	.jpg
1/28/2013	7:45:08 PM	74670byte	C:\Program Files\MSN\Windows\img.img	.img
1/28/2013	7:45:08 PM	127027byte	C:\Program Files\MSN \Windows\hmm.doc	doc
1/28/2013	7:45:08 PM	274627byte	C:\Program Files\MSN Gaming Zone\Windows\winrar.exe	application
1/28/2013	7:45:08 PM	1167027byte	D:\Program Files\MSN \Zone\Windows\attaindance.doc	doc
1/28/2013	7:45:08 PM	167027byte	D:\Program Files\MSN \ Zone\Windows\Marks.doc	doc
1/28/2013	7:45:08 PM	67027byte	D:\Program Files\MSN Gaming Zone\Windows\college.jpge	.jpg
1/28/2013	7:45:08 PM	1467027byte	D:\CSE\Gaming \part.doc	doc
1/28/2013	7:45:08 PM	1467027byte	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	doc
1/28/2013	7:45:08 PM	14670byte	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	doc
1/28/2013	7:45:08 PM	24670byte	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	doc
1/28/2013	7:45:08 PM	246707byte	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	doc
2/28/2013	7:45:08 PM	44677byte	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	doc
2/28/2013	7:45:08 PM	1467027byte	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	doc
2/28/2013	7:45:08 PM	146027byte	C:\Program Files\MSN Gaming Zone\Windows\Tae.doc	doc

Figure 4: Data cleaning

4.3 Pattern Discovery

Pattern discovery using the SOM occurs after the data pre-processing phase. After cleaning a data, extract the pattern for each field. This process is referred as pattern discovery [19]. Pattern Discovery involves the data mining features like clustering and classification [20].This process involves the analyzing of depth of data. The characteristics of the SOM make it ideal for association, classification, clustering and forecasting, therefore SOM is mostly used in analyzing forensics data. SOM is able to map high dimensional data into low dimensional data. Using the proposed system, we are able to visualize pattern of data after preprocessing. By observing the visualized data in the form of graphs, we are able to understand of pattern of host’s computer system.

4.4 Pattern Analysis

It is easy to understand pattern if they are represent in visual representations. After pattern discovery till we need to analyze the patterns in order to determine which ones are interesting and relevant. Visualization technique supports the identification [19]. The visualization of results obtained from the pattern discovery phase is simply the presentation of the results in visual forms, but can help to identify complex relationships within multi-dimensional data [11]. More importantly, it is used as a visualization platform offering a powerful framework for visualizing and analyzing the data that was provided to the pattern discovery process.

The Pattern discovery process analyses each and every pattern on different computers and changes in pattern represents anomalous behavior of the host’s system.

5. CONCLUSION

This paper describes process of anomaly detection using self organizing map. We are able to detect anomaly from computer system on massive data efficiently with visualization pattern. Here anomalies are considered as a unscheduled behavior of computer system which is recognized by analyzing different pattern of logs of computer system. These logs files are first filtered and then cleaned for the attributes which helps to detect the anomaly. Using self organizing maps we visualize the data set for each and every attributes to detect anomaly. The future scope may focuses on use of 3D visual mapping of attributes of data set.

6. REFERENCES

[1] Kohonen, T. 1990, “The self-organizing map”, Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480.

[2] B.K.L. Fei, J.H.P. Eloff, H.S. Venter and M.S. Olivier, 2005, “Exploring Data Generated by Computer Forensic Tools with SelfOrganising Maps” Advances in digital forensics, pp. 113-123. Springer.

[3] V. Chandola,A Banerjee and V.Kumar, July 2009, “Anomaly Detection –A Survey”,ACM Computing Survay,vol. 41,no.3, pp. 1-58.

[4] Dipankar Dasgupta and Nivedita Sumi Majumdar ,2002. Anomaly Detection in multimedia data using negative selection algorithm, CEC 02. Proceedings on Evolutionary Computation,

[5] Li Yao , Li ZhiTang, Liu Shuyu 2006. A Fuzzy Anomaly Detection for IPv6 ,SKG '06. Second International Conference on Semantics, Knowledge and Grid,

[6] Lv, Jun; Li, Xing; Ran, Congsen; Li, Tong. 2006., A new algorithm for Network anomaly detection, International Multi-Conference on Computing in the Global Information Technology ICCGI 06

[7] Ning Chen; Xiao-su Chen; Bing Xiong; Hong-wei Lu, 2009, “An Anomaly Detection And Analysis Based on Corelation Coefficient Matrix”,. International Conference on Scalable Computing and

Communications; Eighth International Conference on Embedded Computing, EMBEDDEDCOM'09, SCALCOM-2009.

[8] Mian Zhang; Li Zhang, 2010, "Based On Pattern Discovery Network Anomaly Detection Algorithm" 5th International Conference on Computer Science and Education (ICCSE).

[9] Jinquan Zeng; Tao Li; Xiaojie Liu; Caiming Liu; Lingxi Peng; Feixian Sun, ICNC2007, "A Feedback Negative Selection Algorithm to Anomaly Detection", Third International Conference on Natural Computation.

[10] E.J. Palomo, J. North, D. Elizondo, R.M. Luque and T. atson,2011, "Visualisation Of Network Forensics Traffic Data With A Self-organising Map For Qualitative Features", Proceedings of International Joint Conference on Neural Networks, pp 1740-1247.

[11] Chi-Yuan Chen; Kai-Di Chang; Han-Chieh Chao, 2011, Transaction-Pattern-Based Anomaly Detection Algorithm for IP Multimedia Subsystem", IEEE Transactions on Information Forensics and Security, pp 152-161.

[12] Chee-Wooi Ten; Junho Hong; Chen-Ching Liu, 2011 "Anomaly Detection for Cybersecurity of the Substations", IEEE Transactions on Smart Grid, pp 865-873.

[13] Zhe Yao; Mark, P.; Rabbat, M.,2012, "Anomaly Detection Using Proximity Graph and PageRank Algorithm" IEEE Transactions on Information Forensics and Security, pp 1288-1300.

[14] Aye, T.T., 2011, "Web log cleaning for mining of web usage patterns", 3rd International Conference on Computer Research and Development (ICCRD).

[15] Ying Zhu, 2011, "Attack Pattern Discovery in Forensic Investigation of Network Attacks", IEEE journal on selected areas in communications, pp 1349-1357

[16] H. Günes Kayacık, A. Nur Zincir-Heywood, 2006, "Using Self-Organizing Maps to Build an Attack Map for Forensic Analysis",ACM digital library.

[17] Correa, Renato Fernandes; Ludermir, Teresa Bernarda 2006, "A Hybrid SOM-Based Document Organization System". Ninth Brazilian Symposium on Neural Networks, SBRN '06.

[18] Kohonen, T.; Kaski, S.; Lagus, K.; Salojarvi, J.; Honkela, J.; Paatero, V.; Saarela, A. **"**Self Organization of a Massive Document Collection**"** , IEEE Transactions on Neural Network

[19] B. K. L. Fei , J. H. P. Eloff , M. S. Olivier , H. M. Tillwick , H. S. Venter, 2006, "Using Self Organizing Map for Behaviour Detection in computer forensics investigation" Proceedings of the Fifth Annual Information Security South Africa Conference.

[20] Smita.Nirkhi, 2010, "Potential use of Artificial Neural Network in Data Mining", International conference on Computer and Automation Engineering (ICCAE).

[21] Kevin Phillip Galloway, 2010, "Intrusion Behavior Detection Through Visualization", M.Sc. thesis.

[22] López-Rubio, E. 2010, "Probabilistic Self-Organizing Maps for Continuous Data"Transactions on Neural Networks, IEEE, pp 1543 - 1554

[23] Nan Zhang; Wei Yu; Xinwen Fu; Das, S.K., 2010, "Maintaining Defender's Reutation in Anomaly Detection Aginst Insider Attack", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, pp. 597-611.