Visualization of Data for Host-Based Anomalous Behavior Detection in Computer Forensics Analysis Using Self Organizing Map

[1]Sushil Kumar Chavhan, [2]Smita M. Nirkhi, [3]Dr. R.V. Dharaskar
[1,2]Dept. of CSE, G.H.Raisoni College of Engineering Nagpur, India
[3]Director, M.P.G.I Nanded, India

Abstract
With the rapidly increasing complexity of computer systems and various media devices and the insufficient attack analysis techniques there is need of improvement of computer forensics analysis techniques. Although many of forensics tools and techniques help in analysis process till forensics analysis process become difficult problem. Here we present an anomalous behavior detection system using self organizing map. Using that system we handle the large volume of data efficiently. Self organizing map has high potential to map high dimensional data also it preserves the topology of data. This technique involves assigning particular values to same data and analyzed visualized pattern with the help of self organizing map. We present result based on implemented system which help to improve investigation.

Keywords
Computer Forensics, Data Mining, Self Organizing Map, Visualization

I. Introduction

Digital forensics is the use of the scientific method to identify, extract and preserve digital evidence that can be used in criminal defense or prosecutions. Further it is divided in computer forensics and network forensics. Computer forensics is the process that applies computer science and technology to collect and analyze evidence which is crucial and admissible to criminal investigations [1]. With the rapidly increasing complexity and interconnectedness of emerging information systems, the number of cyber crimes is increasing day by day. Therefore, the protection of systems as well as the establishment of evidence-based accountability for malicious actions poses significant challenges. Currently there are number of difference types of devices listed available which is capable of providing digital evidence for law enforcement and intelligence purposes [2]. In computer forensics process, a large number of suspect data will be gathered, the amount of these suspicious data is very large, and acquisition process of the useful evidence in forensics is inefficient.

In general, logs and metadata are precisely used by forensic investigators for prediction of crimes [3], also signature-based technique only able to detect the use of well-known malicious hacking tools. These techniques allow us to check whether normal system programs have been modified or not [3].

Data mining is a very appropriate tool for forensics investigation; there are many data mining methods in the computer forensics along with its unsupervised learning. 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. This is performed by representing the data as characteristics in a lower dimensional representation space. The groups are identified according to a similarity measure and are clustered accordingly. Appropriate data mining techniques include support vector machine learning algorithm, behavior based anomaly detection, and heuristic-based anomaly detection [4]. 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 [5]. The main challenges faced by computer investigator are an accurate and efficient representation of knowledge for investigation [1]. Daily use of large media devices generates large volume data. Visualization is useful tool for investigation which represents a powerful link between the most dominant information-processing systems, the human brain and the modern computer (ex. SOM able to visualize large volume of data).

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 [3].

This paper presents a technique which compromised systems by analyzing the hard drives of computer system. This technique extract all content from hard drive, after cleaning useful content we are able to detect anomalous system information.

II. Related Works

Now a day, numbers of researchers had addressed to the security issues of the computer forensics and network forensics, and developed various technologies for the investigative features. In this section, we have analyzed the definitions of digital evidence, computer forensics and anomaly detection and also introduced some studies that had down in visualization.

A. Digital Evidence

In general digital evidence is a series of binary digit numbers which are on transmission [9], or stored information files on the electronic device. These digital evidence file formats includes audio, video, images, and digital, etc. There are some features to look for, the digital evidence can be can be modified easily, can be copied with unlimited differences, hard to be identified the original resource, can be integrated data verification, and cannot be understood directly without technical process.

B. Computer Forensics

Computer forensics is the process that applies computer science and technology to collect and analyze evidence which is crucial and admissible to criminal investigations [1]. It is the practice of lawfully establishing evidence and facts. With Increased

emphasis on social security issue, crime issue is considerable when it comes to the utilization of various media devices and computer system. Digital forensics provides the technical skills to collect evidences for the court to review cases. Due to the tremendous use of computers, Internet, mobile phones, digital cameras, hardware, storage devices etc, digital equipment has changed daily. Network forensics, mobile forensics, computer forensics, and memory forensics, etc. are the widely used areas for digital forensics

In computer forensics number of forensics tool widely used to help the investigation. Hardware tools as well as software tool which acquire data from disk, various storage devices such as CD-ROM,DVD, Compact Flash, Micro Drives, Smart Media, Memory Stick, Memory Stick Pro, xD Cards, Secure Digital Media and Multimedia Cards [8] . EnCase Forensic, Forensic ToolKit, SafeBack, Storage Media Archival Recovery Toolki, FRED System, NTl Secure ToolKit etc are the commonly used forensics tools.

C. Anomaly Detection

Anomaly detection refers to the problem of finding patterns in data that do not conform to expect behavior [3]. Initial research on anomaly based intrusion detection by Denning was based monitoring the system's audit data [9], Li Yao et.al [10], 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. Jinquan Zeng et.al [11], 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. Ning Chen et.al [12], 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. 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. All this approaches now use in forensics.

D. Visualization

Visualization is the graphical presentation of information, with the goal of providing the viewer with a qualitative understanding of the information contents. Several tools have been proposed that are capable of providing the visualization and statistical analysis of logs, MieLog, NVisionIP etc. Most of the visualization was done in forensics which uses self organizing map. Fei in his master thesis data visualization in digital forensics; "Self Organizing Map Forensic Analysis" (SOMFA) which is unsupervised neural network model developed by him for visualization of computer anomalies [13]. Lerche et al.give the overview of visualization of forensic data where basic and fundamental visualization was explained. And how different techniques could be used in forensic process also discussed. Also they focus how visualization helps to detect anomalies and attack in network forensics [14]. Web history also important in digital forensics. Sarah lowman studied the problem of web history and make developed a tool which is able to visualization web history. Also he explained various tools related with web history visualization. Intruder also visualizes using graph base visualization. Static and dynamic instances are able to visualized using this method [15], Visualization of time related data able to find connections and correlations between different data types. For this purpose time related data visualization is done by Willassen. Further he discusses how to use timestamp and how it is improved [16], .Fanlin Meng, et al. [17] develop framework for email visualization. Which provide easy analysis of network relented data and better understandability of data. Visualization improve efficiency of forensics analysis

III. Self Organizing Map

The SOM is an unsupervised neural network algorithm that uses competitive learning used for analyzing and visualizing high dimensional data [6]. The Self Organizing Map is one of the most widely used. 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 [7].

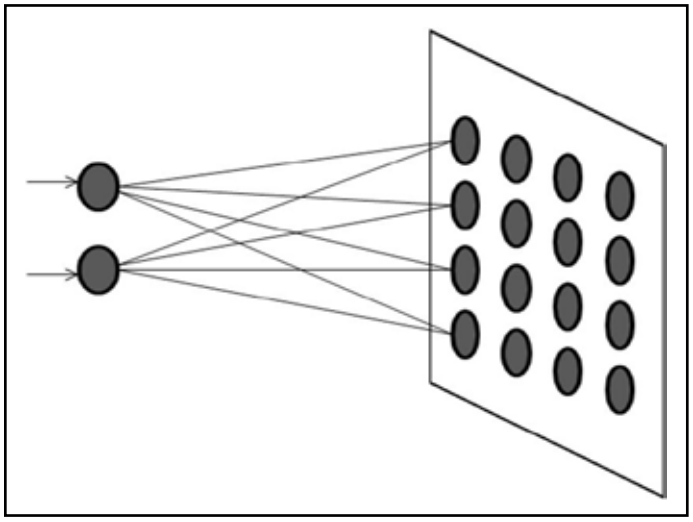


Fig. 1: Architecture of Self Organizing Map (Kohonen)

The SOM consists of a number of neurons arranged according to some topological order, typically a 2D rectangular or hexagonal grid. Fig shows the architecture of self organizing map. Usually the SOM consists of a one or two-dimensional array of identical neurons. The input vector is broadcast in parallel to all these neurons. Each neuron of the map is assigned a weight vector wi, with these weight vectors having the same dimensionality as the input vectors xi. For each input vector, the most responsive neuron is located .The weights of this neuron and those within a neighborhood around it are adapted as in the winner-take-all network to reduce the distance between its weight vector and the current input vector. During the training phase, each input vector is provided to the SOM and each neuron determines its activation. Generally Euclidean distance between input vectors and weight vectors is used for the calculation of activation of neuron.

The SOM Learning algorithm as follows:

Assume an output array of two dimensions with k x k neurons, that the input samples, x, have dimensionality N, and that index n represents the nth presentation of an input sample.

1. Set to small random values to all weight vectors are set to small random values. All values must be different
2. Select a sample input vector x and locate the most responsive neuron call winning using some distance metric usually the Euclidean distance

i.e. | x(n)-w_j| is a minimum (and j=1,2,……. M, where M=K.K)
3. Adapt all weight vectors, including those of the winning neuron, within the current neighborhood region. Those outside this neighborhood are left unchanged.

$$w_i(n+1) = \begin{cases} W_i(n) + \alpha(n)[x(n) - w_i(n)] & i \in \Omega(n) \\ W_i(n) & Otherwise \end{cases}$$

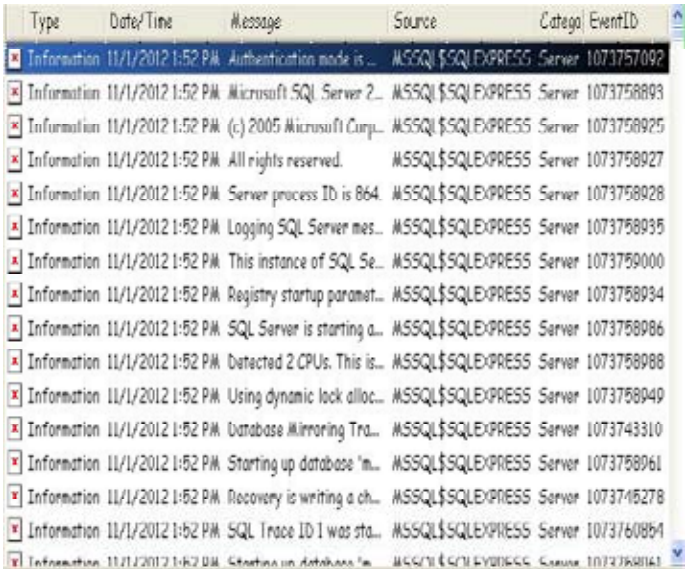
where α(n) is the current adaptation constant.
$\Omega(n)$ is the current neighborhood size centered on the winning neuron.
4. Modify, as necessary and, until no further change in the output feature map is observable (or some other termination condition) otherwise go to step two.

IV. Experimental Results

In this section, we present some experimental results based on the SOM for the detection of expected behavior of computer system. We tested this system where work is equally distributed in any origination. Step of anomalous system detection includes data collection, data cleaning, pattern discovery, and pattern analysis.

A. Data Collection

Initially, we collected CPU usage, disk usage and login-time data using Interface, which gives logs of all activities done on computer.



Type	Date/Time	Message	Source	Catego	EventID
Information	11/1/2012 1:52 PM	Authentication mode is ...	MSSQL$SQLEXPRESS	Server	1073757092
Information	11/1/2012 1:52 PM	Microsoft SQL Server 2...	MSSQL$SQLEXPRESS	Server	1073758893
Information	11/1/2012 1:52 PM	(c) 2005 Microsoft Corp...	MSSQL$SQLEXPRESS	Server	1073758925
Information	11/1/2012 1:52 PM	All rights reserved.	MSSQL$SQLEXPRESS	Server	1073758927
Information	11/1/2012 1:52 PM	Server process ID is 864.	MSSQL$SQLEXPRESS	Server	1073758928
Information	11/1/2012 1:52 PM	Logging SQL Server mes...	MSSQL$SQLEXPRESS	Server	1073758935
Information	11/1/2012 1:52 PM	This instance of SQL Se...	MSSQL$SQLEXPRESS	Server	1073759000
Information	11/1/2012 1:52 PM	Registry startup paramet...	MSSQL$SQLEXPRESS	Server	1073758934
Information	11/1/2012 1:52 PM	SQL Server is starting a...	MSSQL$SQLEXPRESS	Server	1073758986
Information	11/1/2012 1:52 PM	Detected 2 CPUs. This is...	MSSQL$SQLEXPRESS	Server	1073758988
Information	11/1/2012 1:52 PM	Using dynamic lock alloc...	MSSQL$SQLEXPRESS	Server	1073758949
Information	11/1/2012 1:52 PM	Database Mirroring Tra...	MSSQL$SQLEXPRESS	Server	1073743310
Information	11/1/2012 1:52 PM	Starting up database 'm...	MSSQL$SQLEXPRESS	Server	1073758961
Information	11/1/2012 1:52 PM	Recovery is writing a ch...	MSSQL$SQLEXPRESS	Server	1073745278
Information	11/1/2012 1:52 PM	SQL Trace ID 1 was sta...	MSSQL$SQLEXPRESS	Server	1073760864

Fig. 2: Collection of Data

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.

B. Data Preprocessing

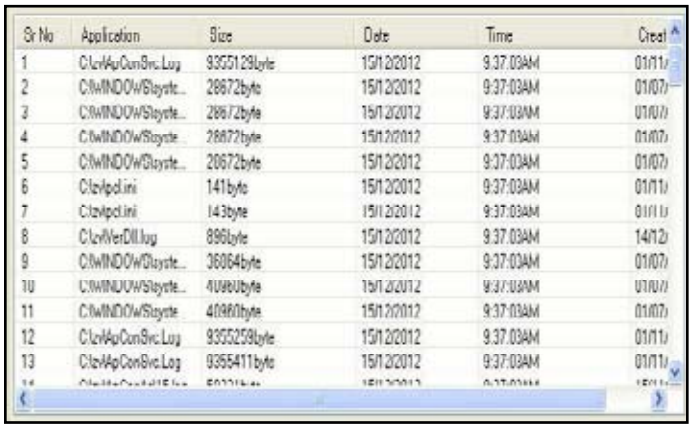
Data preprocessing is the important step in data mining. This includes data cleaning, Data integration, Data transformation, Data reduction, Data discretization [18]. The work of Data cleaning is to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. To obtain the desired filed, we must clean data in order to remove the unwanted field from the collected logs file. Generally data cleaning involves several stages such as Data analysis, Definition of transformation workflow and mapping rules, Verification, Transformation, Backflow of cleaned data [19].

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 the field which is size, time, and type of all logs are remain output is as follows



Sr No	Application	Size	Date	Time	Creat
1	C:\zvApConSvc.Log	9355129byte	15/1/2012	9:37:03AM	01/11/
2	C:\WINDOWS\byte...	28672byte	15/1/2012	9:37:03AM	01/07/
3	C:\WINDOWS\byte...	28672byte	15/1/2012	9:37:03AM	01/07/
4	C:\WINDOWS\byte	28672byte	15/1/2012	9:37:03AM	01/07/
5	C:\WINDOWS\byte...	20672byte	15/1/2012	9:37:03AM	01/07/
6	C:\zvApd.ini	141byte	15/1/2012	9:37:03AM	01/11/
7	C:\zvApd.ini	143byte	15/1/2012	9:37:03AM	01/11/
8	C:\zvVerDll.log	896byte	15/1/2012	9:37:03AM	14/12/
9	C:\WINDOWS\byte...	36864byte	15/1/2012	9:37:03AM	01/07/
10	C:\WINDOWS\byte...	40960byte	15/1/2012	9:37:03AM	01/07/
11	C:\WINDOWS\byte	40960byte	15/1/2012	9:37:03AM	01/07/
12	C:\zvApConSvc.Log	9355259byte	15/1/2012	9:37:03AM	01/11/
13	C:\zvApConSvc.Log	9355411byte	15/1/2012	9:37:03AM	01/11/

Fig. 3: Data Cleaning

Apply SOM algorithm to cleaned data of following graph is generated. Graph is the various component maps of cleaned fields. The first component map represented the file type; the second component map represented the time when the files were created or modified or changed. The third component maps represented the day of the week on files were created. In each graph different color indicates different values. In file type component map yellow, green red and black color indicates document, audio, videos and application values respectively. In time component map red, green, blue and yellow color indicates morning hours, midday hours, evening and night hours respectively. In third day of weak component map green for early days (Monday, Tuesday), red of mid days (Wednesday. Thursday) and yellow for weekend days (Friday, Saturday).
The specific signature of any computer in organization is not specific so to achieve the objective was to study the anomalous system of each computer user. For that time and days component map is mostly important to locate the anomalous system.
We apply this on three systems whose component graph is as follows.

ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

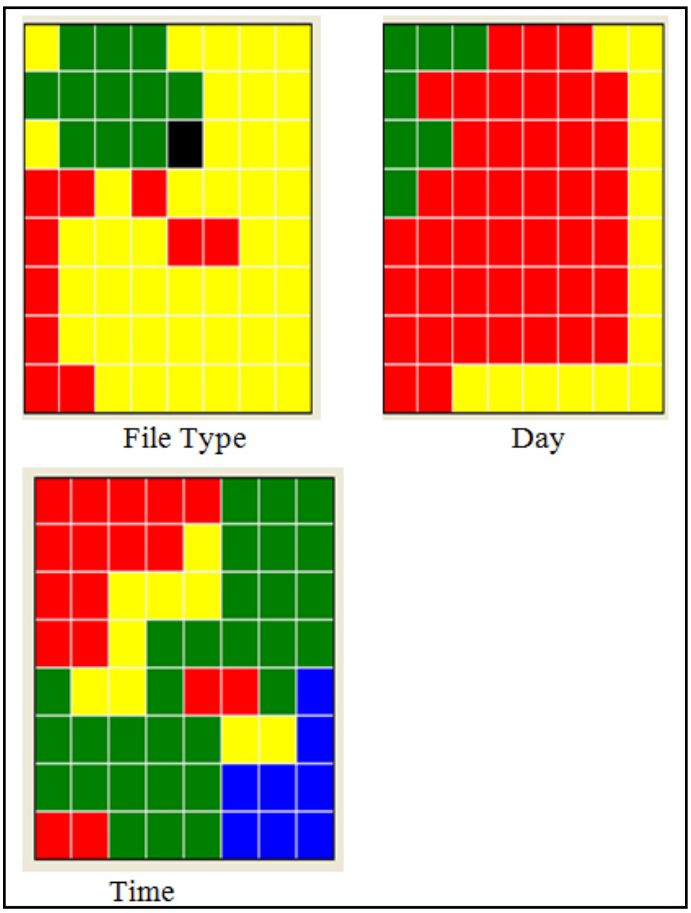


Fig. 4: Component Maps of user 1

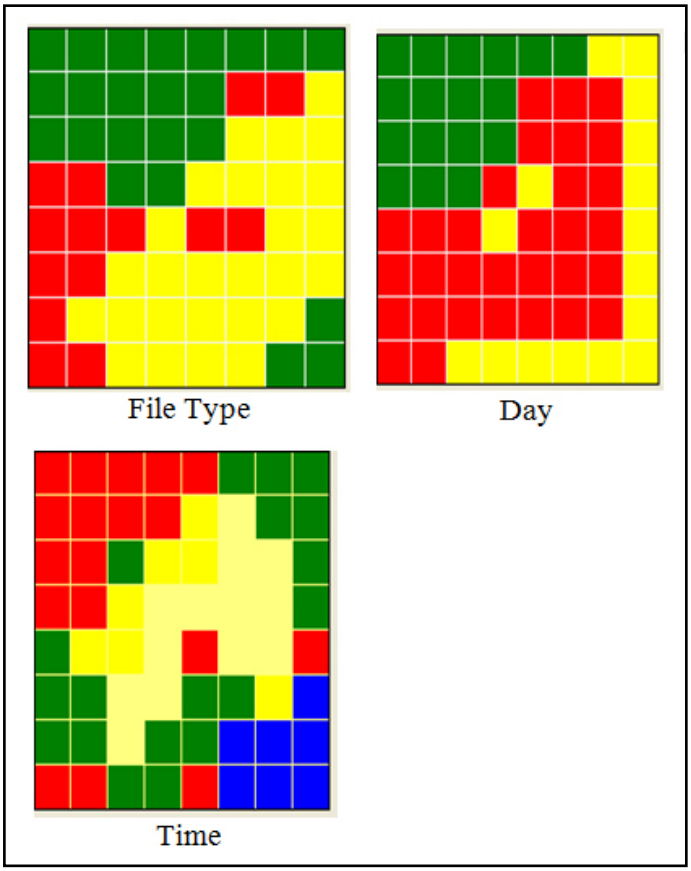


Fig. 5: Component Maps of user 2

In fig. 4 user use host computer mostly for documented work with audio video, and application. This work is mostly on mid days some on weekends and few at starting days of week. User works in midday mostly, few work on morning, evening also view on component map.

In fig. 5 user use host computer mostly for documented work with audio, video and application. This work is mostly on mid days some on weekends and few at starting days of week. User works in midday mostly, few work on morning, evening also visualized in component map.

In fig. 6 user use host computer usage is for audio videos and application purpose. Documentation was also done by user. This work is mostly on starting day's equal work on midday and weekend. User works in morning mostly, after that his work progress is less.

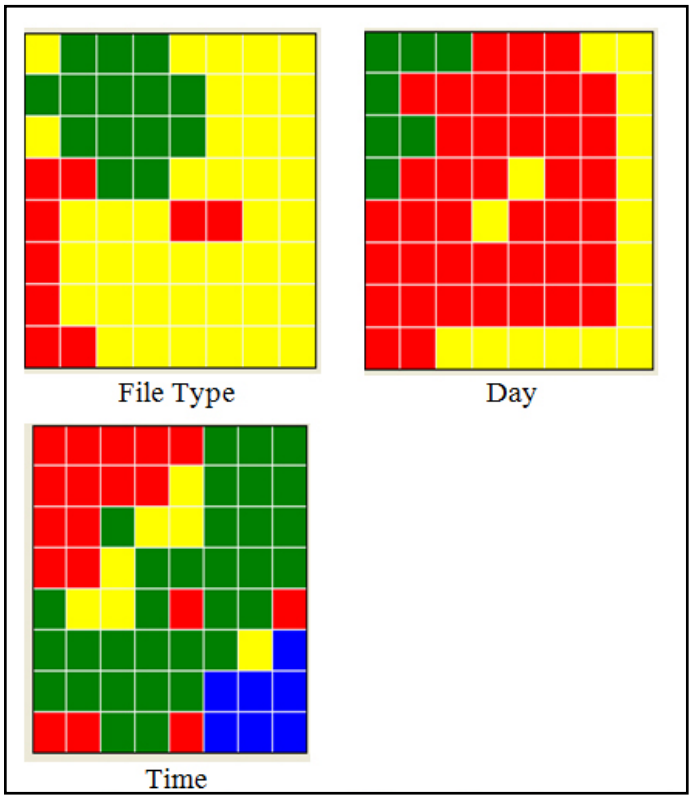


Fig. 6: Component Maps of user3

Analyzing this three component graph individually. All host computers tested who had a similar work task. So expected that all the system have same component map visualization. Using this assumption it is found that third has user is anomalous behavior. That host system computer system is found to be anomalous.

V. Conclusion

We presented concept where use of self organizing map is used for anomaly detection with graphical visualization. The color code map is used for same. We are able to detect anomaly from computer system on massive data efficiently with visualization pattern. Here anomalies are considered as 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 work may includes use the concept described in paper for multimedia and streaming data.

References

[1] Kan, S., Kim, J.,"Network Forensic Analysis Using Visualization Effect", International Conference on Convergence and Hybrid Information Technology 2008, pp. 466-473.

[2] G. Mohay,"Technical challenges and directions for digital forensics", In SADFE '05: Proceedings of the First International Workshop on Systematic Approaches to Digital Forensic Engineering. Washington, DC, USA: IEEE Computer Society, 2005, pp. 155.

[3] A. J. Marcella, R. S. Greenfield,"Cyber Forensics", Auerbach Publications, 2002.

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

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

[6] Samuel Kaski,"Data exploration using self organizing maps", Acta Polytechnica Scandinavica Mathematics Computing and Management in Engineering Series.

[7] Kohonen, T.,"The self-organizing map", Proceedings of the IEEE, Vol. 78, No. 9, pp. 1464-1480, 1990.

[8] Raza Hasan, Akshyadeep Ragha, Salman Mahmood, "Overview on Computer Forensics Tools", In UKACC International Conference on Control 2012.

[9] D. E. Denning,"An Intrusion-Detection Model", IEEE transactions on Software Engineering, Vol. SE-13, No. 2, pp. 222-232, Feb. 1987.

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

[11] Jinquan Zeng, Tao Li, Xiaojie Liu, Caiming Liu, Lingxi Peng, Feixian Sun,"A Feedback Negative Selection Algorithm to Anomaly Detection", Third International Conference on Natural Computation, ICNC 2007.

[12] Ning Chen, Xiao su Chen, Bing Xiong. Hong wei Lu,"An Anomaly Detection And Analysis Based on Corelation Coefficient Matrix", International.

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

[13] B.K.L. Fel,"Data visualisation in digtial forensics", M.S. Thesis, 2007.

[14] Gerald Schrenk, Rainer Poisel,"A Discussion of Visualization Techniques for the Analysis of Digital Evidence", International Conference on Availability, Reliability and Security, pp. 758 -763, 2011.

[15] Sarah Lowman,"Web History Visualisation For Forensic Investigations", M Sc thesis 2010.

[16] Jens Olsson, Martin Boldt,"Computer forensic timeline visualization tool", Science Direct Digital Investigation, 2009.

[17] Junbin Yang Genzhen Yu Fanlin M eng, Shunxiang Wu. "Research of an e-mail forensic and analysis system based on visualization", Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications, pp. 281–284, 2009

[18] Smita.Nirkhi,"Potential use of Artificial Neural Network in Data Mining", International Conference on Computer and Automation Engineering (ICCAE), pp. 339 343, 2010

[19] E. Rahm, Hong Hai Do,"Data cleaning: Problems and current approaches", IEEE Data Eng. Bull., 23(4), pp. 3-13, 2000.

Mr. Sushilkumar Chavhan has received Bachelor of Engineering Degree from B.C.Y.R.CS.Umrer College of Engineering, Umrer, in Computer Engineering, 2011. Currently pursuing M.Tech in computer science and Engineering from G.H.Raisoni college of Engineering, Nagpur. His area of interest include Data mining, Artificial Neural Network, pattern recognition, Digital Forensics.

Asst. Prof. Ms. S. M. Nirkhi has completed M.Tech in Computer Science &Engineering & currently Pursuing PHD in computer science. She has received RPS grant of 8 lakhs from AICTE for her Research. She has attended 6 STTPworkshops along with other training programs. She has Published 15 papers in international conferences & 5 papers in international journals. She had presented paper at International Conference at Singapore. She has 12 years of professional experience. Currently working as Assistant professor in Department of Computer Science & Engineering at GHRCE. Her area of interest include Soft computing, Data mining, web mining, pattern recognition, MANET, Digital Forensics.

Dr. R.V. Dharskar has received Ph.D. Degree (Computer Science & Engineering) from Amravati University, M. Tech. (Computers) from I.S.M. and P.G. Dip. M.Phil., M.Sc. from Nagpur University. He is having more than 29 years of teaching and 23 years of R&D experience in the field of Computers & IT. He is an author of number books on Programming Languages. He has been actively involved in the research on Mobile Computing, Multimedia, Software Engineering, Web Technology, E-Learning and Networking, Digital Forensics. He has authored more than 227 research papers at various International / National Conferences and Journals. His research work has been accepted at IEEE Computer Society of USA, Bristol (UK), Hong Kong (China) etc. He has been invited as a Keynote Speaker, Invited Speaker, and Session Chair for more than 34 International & National Conferences.