International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Data Mining Technique to Data Collection and Analysis for Cyber Forensic

Prashant Khobragade, Pranay Saraf, Priya maidamwar, Preeti Thakre

Abstract: ***In the world of Digital forensic the uncovered digital may contain vital information for digital data investigation for investigator. Digital data collected from the crime scene leads to find out the clue after performing analysis by the examiner. This process of data examination data collection and analysis plays important role in cyber world for the forensic investigator. The cybercrime is a part of computer forensics where the digital evidences are analyze by the investigator and to perform analysis special measurements and techniques are required in order to use this details that has to be accepted in court of law for law enforcement. The data collection of evidence is a key aspect for the investigator, such kind of digital data has to be collected from different sources at the crime scene and this process involves to collect each and every evidence of digital crime scene and later this gather data will be analyze by the experts to reach to the conclusion. In this paper the proposed method collected the data from the crime scene efficiently which includes log data, transactional data, physical drive data, and network data; later this collected data analyzed to find out the theft node in the network. In this paper FTK 4.0 digital forensic tool used to reduce plenty of time for data processing and later report will be produce that will be accepted tin the court of law. This paper also focuses the data collection method with in the network and reach to the faulty node and later this faulty node analyzed with all collected data for forensic analysis. For this standard algorithm used to analyze the performance of distinct features used for network attacks. Kmeans clustering methodology is used to create cluster of victim node and represent victim data in systematic manner for the ease of law enforcement.***

Keywords- Data Collection, SVM, KMeans Clustering, Network Analysis,File System Analysis.

Manuscript received on January 02, 2020.
Revised Manuscript received on January 15, 2020.
Manuscript published on January 30, 2020.

* Correspondence Author

Prashant Khobragade*, Department of Computer Science and Engineering, G H Raisoni College of Engineering, Nagpur, India. Email: prashant.khobragade@raisoni.net.

Pranay Saraf, Department of Computer Science and Engineering, G H Raisoni College of Engineering, Nagpur, India. Email: pranay.saraf@raisoni.net

Preeti Thakre, Department of Computer Science and Engineering, G H Raisoni College of Engineering, Nagpur, India. Email: preeti.p.thakre@raisoni.net

Priya maidamwar, Department of Computer Science and Engineering, G H Raisoni College of Engineering, Nagpur, India. Email: priya.maidamwar@raisoni.net

I. INTRODUCTION

In Data mining the different technique used for knowledge representation from the huge data. Data mining methods have huge potential in area of forensic investigation, where data mining models, techniques and tools can be developed for the betterment of digital forensic investigators, the digital forensics experts and law enforcement officers are work on same data to find important data or clues from the collected data also they are searching for much more efficiently and faster if the data is smaller to analyzed. In cyber forensic data gathering is an important issue as on every single click in network data is generated at the background. As the technology increases the cyber-crime also be increases as a result hugh data has been generated. Cybercrime is any criminal act in a computer network or illegal activity performed by the hackers. In cybercrime the crime is supported by data clue using some methods or involves the use of digital forensic application/devices. In digital cybercrime the use of digital data that are involved only in the computer storage memory such as hard drive or digital storage medium, and to examine such data computer forensic techniques were used. The Computer forensic is a part of forensic science process and the objective of forensic investigation is to explain the present scenario of the digital evidence or artifact [2]. There are plenty of digital devices in the market which are used by individual's person. For example it includes personal computer, digital camera, mobiles, laptops, wearable devices, wireless devices, telephone, hubs, modems, routers, and iPods etc. Cyber terrorism [8] is a crime where the online running website, routers, e-mail server, or computer system were used for various kinds of cyber-attacks such as use of distributed denial of service or other attacks by the hackers.

As an each individual today maintains more than one email account, is a member of many communities, virtual groups, takes active part in chat rooms and other networking sites with his/her identity or under an alias, juggles multiple flash drives and other digital storage media[1][2]. The concept of the crowd sourced forensic investigation via the construction of a simple process model presents a simple model for crowd sourced digital forensics, and discussed various technique utilized in such forensic investigations [6].

In cyber forensics Network forensics is used in the case of network security. It further extended in the field of network security as the collected data for analysis of intruder from the raw data for security purpose from different resources like network firewalls, internet or intranet server database, hubs, routers using Intruders detection System (IDS). The results IDS data analysis are utilized for forensic investigation and identify what kind of attacks is to be done on which system [9]. However, there are some cybercrime as it not intended in the network and does not breach standard network security policies. The idea behind network forensics analysis define as the data is to be collected across the network i.e.,

Mostly ingress traffic and egress traffic over the network form one hos to another host. Network forensics attempts to analyze traffic data over the network logged through firewalls or intrusion detection systems or at network devices such as hub, switches, and routers [5].

In computer forensic analysis, hundreds of thousands of files are usually examined. Such data files may be consists of unstructured kind of data, and it's difficult for the forensic investigator to analyze the data and find the clue. To investigate data present in computer drive and analyze such unstructured data require lots of time. Also this process involves headache to the forensic investigator to find out clue for further investigation. To solve this issue data mining clustering algorithm is typically used for exploratory data analysis and getting similar kind of data evidence for the examiner for further investigation, as clustering algorithm were used and such algorithm does not require prior knowledge about the data [10]. In the field of network forensics [4] [10] it includes monitoring various network traffic and determining whether there area is an anomaly is present in network traffic and determining whether it indicate any kind of attack. If an attack is detected over the network, which is based on parameter or pattern of attack used by the attackers. Network forensic [6] methods used to track back the attackers to its original node and location using system address and IP address. The ultimate goal of network forensic is to provide valid and sufficient evidence to the forensic instigator to reach to final conclusion and able to generate report which will further used in the court of law.

The proposed system [13] presents a fundamental model for data collection and analysis in network as well as file system analysis. There were different attacks has been happens over the network and resulting to malicious activities, the data is related with those attacks has been retrieved for forensic evidence examination using proposed method. Data mining technique is used to identify and analyses attacker's node and victims' system using clustering technique. Digital forensic toolkit FTK 4.0 were used to find out malicious pattern and to analyze the victim system for collecting the data and generate report in the court of law[12].

This research paper further organized are as follows: In section 2 literature surveys associated to put this paper in the right context of evidence. Section 3 describes the proposed method for data collection and analysis in network. In section 4 clustering technique and different methods of data mining is describe. Section 5 provides extensive experimental results and the lastly section 6 draws conclusions and future work.

II. LITERATURE SURVEY

In the field of cyber forensic there are studies related to data generation or collection is discussed, like data collection from hard disk, portable storage disk and flash drive. The J.A. Hamilton et, al., in 2011 focuses the cyber forensic analysis is also be done with the help of some data mining tool like WEKA, this data mining tool used to represent the data in systematic manner. In network forensic the data analysis, it highlighting and understand using visualization method that related to various network devices that used to build more efficient forensic system [5]. The cyber forensic is the method of computer forensics to collect the data from crime scene and analyze all related evidence to the crime which is useful and important for creation of strong evidence against cybercrime investigation. The Lianfi Y in 2012 discussed the data collection [2] approach from the physical storage media. In physical storage devices the data collection give method to collect, stored and analyzing data. The data collected from the hard drive or physical storage devices that includes facts will further analyze by forensic analysis experts [12] [13].

The theory of network forensics [7] deal as the found across a network connection i.e., incoming packets traffic and traffic for out from one host to another. In network forensic includes identify the network traces and improve IDS detection system from various attacks. This network traces include various parameters from routers, log data from firewall, sniffers log data and data generated at network service providers database. There are huge data on each single event generated in the network so it is very difficult for forensic investigators experts to find out important clue and analyze those data with minimum time frame [13].

III. METHODOLOGY USED

The Dataset present KDD Cup 1999 for intrusion detection system used to train the system, proposed method implemented and process was used to acquire raw TCP/IP data from the local network by creating working prototype in the LAN environment. Departmental LAN. The local area network were created to capture real time data in real environment. The In this process total connection of TCP/IP checked and used in experimental result, total 24 different qualitative and quantitative features were used from the given dataset. In the subset of dataset around 494,021 data were used in which 32% data represent normal network patterns data.

The attack type's fall into four main categories were used:

1. SQL Injection: SQL based query
2. U2L: unauthorized access to local admin user.
3. XSS: Cross site Scripting.
4. DDOS: Distributed denial of service.
5. R2L: unauthorized access from a remote machine

1. SVM Based Classification and Ranking:

SVM ranked based method with 21 distinct feature used for IDS, for this we have used different parameter to check the overall accuracy of classification method [11].

The parameters are training time; testing time; relevant; irrelevant. According to SVM based feature comparison the following rules for analysis were used.

Rule 1. If decreases in testing time and increasing in training time and decrease in accuracy, then the selected feature is accepted.

Rule 2: If increasing in testing time and increasing in training time and decreasing in accuracy, then the selected feature is accepted.

Rule 3. If increasing in testing time and decreasing in training time and decreasing in accuracy, then the selected feature is accepted.

Rule 4. If increasing in testing time and increasing in training time and accuracy unchanged, then the selected feature is accepted.

Rule 5. If increasing in testing time and decreasing in training time and accuracy is not changed, then the selected feature is accepted.

Rule 6. If decreasing in testing time and increasing in training time and accuracy remain unchanged, then the selected feature is accepted.

Rule 7. If decreasing in testing time and decreasing in training time and accuracy remain unchanged, then the selected feature is rejected.

Rule 8. If decreasing in testing time and increasing in training time and increasing in accuracy, then the selected feature is rejected.

Rule 9. If increasing in testing time and increasing in training time and increase in accuracy, then the selected feature is rejected.

Rule 10. If decreasing in testing time and decreasing in training time and increasing in accuracy, then the selected feature is non-relevant.

2. KMeans Clustering

The data mining technique include data clustering, it is used to build group of similar kind of data together using unsupervised data models. The uncategorized data instances are grouped together, defined by clustering method, multidimensional data and based on similar kind pattern matching schemes. The clustering method used to create a similar and significant interest together and able to detect outliers from the given dataset. In this proposed method simple k-means method were used to identify the victim node from the given KDD Cup dataset. Simple k-means algorithm start to create k cluster, in this proposed method given unorganized data categorized in cluster and calculate the mean from the given cluster data, as an input parameter and partitions the given set of n objects into k clusters so that the consequential intra-cluster relationship is high while the inter cluster similarity is low [11].

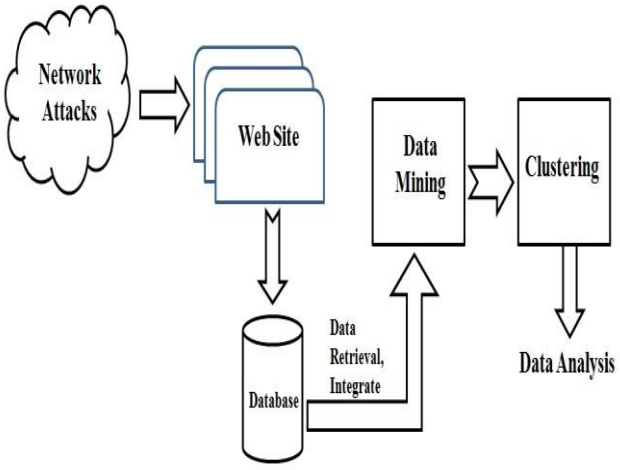

Figure 1: Proposed Method for data collection analysis

In proposed method clustering technique is used as a step in the analysis of collected data as part of the forensic investigation. Our significance is to examine that data instance which doesn't group naturally into cluster to form groups, for forensic evidence. The mean is calculated using Euclidean distance assign to each instances of a clusters.

IV. RESULT AND DISCUSSION

1. Network Data Analysis

Network forensics is capturing, recording and analysis of network events in order to discover the source of cyber-attacks. In network forensic analysis the data gathered from source system are collected for examination. The network data contains parameter which are gives sufficient information about network malicious activity and the source system, where actually thread happened. There is real time an attack has been made on website in network. The each attacks has their own pattern and standard for hackers, based on this the data has been reflected on the server database.

There are other reasons that we use SVMs for intrusion detection. The first is speed; because real-time performance is of primary importance to intrusion detection systems, any classifier that can potentially run "fast" is worth considering.

Table 1: Selected 21 features using SVM

Categories of Attack	Training Time (sec)	Testing Time (sec)	Accuracy (%)
SQL Injection	11.42	1.09	99.78
XSS	19.23	1.23	99.12
DDOS	24.77	2.92	99.25
U2L	4.56	1.12	99.45
R2L	11.78	1.29	99.20
Normal	6.76	1.22	99.55

Table 2: Feature Analysis using SVM

Categories of Attack	Number of Features	Training Time (sec)	Testing Time (sec)	Accuracy (%)
SQL Injection	7	12.89	1.07	99.59
XSS	8	8.36	1.87	99.38
DDOS	19	22.89	1.70	99.49
U2L	8	2.86	1.08	99.33
R2L	6	3.98	0.85	99.87
Normal	25	9.24	2.10	99.78

Table 3: Selected 21 Features using K-Means Algorithm

Item	Data (Training)	Data (testing)	Accuracy (%)
Cluster 1 (Attacks)	5483	2378	99.12
Cluster 2 (Normal)	1209	953	98.24

K-Means clustering method is used to collect the all data such as attacker's data and normal user data to form a cluster, so from this we come know the malicious activity and normal user activity.

V. CONCLUSION

This paper explains the technique for data collection in network and data mining method shows the analysis of this data using SVM and KMeans methods. Data mining technique used, such as classification and clustering for data analysis, gives analysis of evidence pattern of attackers. On the basis of this network forensic, we developed framework for data generation and analysis. The proposed method used data mining technique to identify useful event in network and select minimum attribute to store the record of each attribute with minimum features. The collected data is sufficient to identify malicious behavior and re-establishment of attack behavior in network. This results in reduction of data storage requirements for forensic investigation.

REFERENCES

1. Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka, "Document Clustering for Forensic Analysis: An Approach for Imp roving Computer Inspection", IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 8, NO. 1, JANUARY 2013.
2. Lianfi Y in, " Research on windows physical memory forensic analysis", Fourth International Symposium on Information Science and Engineering, 2012.
3. Funminiyi Olajide, Nick Savage, Richard Trafford, "Forensic Memory Ev idence of Windows Application", The 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012).
4. Danie l Compton, J.A. Hamilton. "An Examination of the Techniques and Implications of the Crowd-sourced Collection of Forensic Data", IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 2011.
5. Veena H Bhat, Prasanth G Rao, Abhilash R V,P Deepa Shenoy,Venugopal K. R. "A Novel Data Generation Approach for Digital Forensic Application in Data Mining", Second International Conference on Machine Learning and Computing, 2010.
6. Sebastian Schmerl, Michael Vogel, René Rietz, and Hart mut König, " Explo rative Visualization of Log Data to support Forensic Analysis and Signature Development", Fifth International Workshop on Systematic Approaches to Digital Forensic Engineering, 2010.
7. Jooyoung Lee, Sungkyung Un, and Dowon Hong, "Improving Performance in Digita l Forensics", International Conference on Availability, Reliability and Security, 2009.
8. Ka ra Nance, Brian Hay and Matt Bishop, "Digital Forensics: Defining a Research Agenda," Proc. of the Forty Second Hawaii International Conference on System Sciences, pp. 1-6, 2009.
9. Seung-hoon Kang, Juho Kim, "Network Forensic Analysis Using Visualizat ion Effect", International Conference on Convergence and Hybrid Information Technology, 2008.
10. Mohd Taufik Abdullah, Ramlan Mahmod, Abdul A. A. Ghani, Mohd A Zain and Abu Baka r Md S, "Advances in Co mputer Forensics," International Journal Of Computer Science and Network Security, vol. 8, no. 2, February 2008.
11. B. S. Everitt, S. Landau, and M. Leese, Cluster Analysis. London, U.K.: Arnold, 2001. P
12. prashant K. Khobragade, Latesh G. Ma lik, "A Review on Data Generation for Digital Forensic Investigation using Data Mining," International Journal of Computing and Technology, Volume 1, Issue 3, April 2014.
13. Khobragade, P.K., Malik, L.G.,"Data Generation and Analysis for Digital Forensic Application Using Data Mining," in Communication Systems and Network Technologies (CSNT), 2014 Fourth IEEE International Conference on , vol., no., pp.458-462, 7-9 April 2014.

AUTHORS PROFILE

Prashant Khobragade, has obtained his bachelor's degree in Information Technology from Amravati University in 2011. He then completed his Masters of Technology in computer science and engineering from RTM Nagpur University in 2014. His main research interest includes Data mining, Cyber forensics, Cloud Computing and network Security.

Pranay Saraf, faculty of Department of Computer Science and Engineering, GHRCE, Nagpur. He obtained his bachelor's degree in Information Technology from RTM Nagpur University in 2011. He then completed his Masters of Technology in computer science and engineering from RTM Nagpur University in 2013. He has been awarded in National E Goverence conference with hands of Honorable Chief Minister Maharashtra. In to his publication credential total 12 conference /journal papers.

Preeti Thakre, faculty of Department of Computer Science and Engineering, GHRCE, Nagpur.She obtained her bachelor's degree in Computer Engineering from RTM Nagpur University in 2014. She then completed her Masters of Technology in computer science and engineering from RTM Nagpur University in 2017. In to her publication & copyright credential, total 05 conference /journal papers and 1 copyright.

Priya Maidamwar, has obtained her bachelor's degree in Computer Technology from RTMNU in 2010. She then completed her Masters of Technology in Computer science and engineering from RTM Nagpur University in 2013. Her main research interest includes Network Security, Wireless Sensor Networks and Machine Learning.