Analyzing the Extracted File Metadata Evidences from Suspicious Nodes in DFXML format using Clustering Techniques

Ms. Shruti B. Yagnik[1], Dr. Binod C. Agrawal[2]

[1]Calorx Teachers' University, Computer Science and Engineering Department, Ahmedabad, Gujarat, India.

[2]Calorx Teachers' University, Computer Science and Engineering Department, Ahmedabad, Gujarat, India.

Abstract

Cyber-Crime by running malicious files on workstations are becoming rampant in information industries. Malicious files are made to run which tend to be dangerous to sensitive information and must be detected as and when they are triggered. Digital Forensics, a sub branch of Cyber Forensics includes the analysis and investigation of these malicious files by investigators. Various file properties (metadata) of each file running/ran on the workstation should be studied in depth in order to reveal the files that were/are not supposed to be run or not. In order to prove that a file that was run was malicious or not, investigators had to manually observe each and every file in the log and study its metadata. This was a tedious and time consuming task. Also, it became very confusing to note down all the metadata and analyze it manually leading to errors and flaws. Henceforth, the research goal here is to provide a methodology wherein the file metadata evidence from a DFXML file is automatically analyzed and helps predict the malicious files that were run on the computer system. This is done to reduce the burden of manually analyzing all the files of the computer system and making the analysis process faster and with minimal errors for Forensic Investigations.

Keywords: Cyber-Crime, Cyber-Forensics, Evidences, Forensic Investigations, File Meta Data, Digital Forensic, Digital Forensic XML, Malicious files, automatic analysis, predict, Clustering, K-Means, Clusters

I. INTRODUCTION

Many files with malicious intentions can reside in computer systems, that can be triggered by any external activity done with a malicious intension or done un-intentionally. Each file may have its own malicious intentions that could be performed in background. To accomplish its goal, these files may use variety of ways to hide themselves This paper highlights the way of detecting such files which appear to be genuine at a glance but are actually black hatted files with malicious code. There are lots and lots of files running on the workstation. Checking each and every file, manually, whether it is a malicious or white hat file will become a tedious task and kill a lot of time thereby completing the work of the malware and fulfilling the attackers desire. Also there are other important places where manual intervention of the investigator is needed than spending it here on the process. That is why we are automating the process and saving the time for the Investigators to thwart attack attempts and safeguard the node as fast as possible [1]. There are clusters formed for all the files and we can detect outliers from those clusters and point out the malicious files. Clustering is an unsupervised learning technique. It is the task of grouping together a set of objects in a way that objects in the same cluster are more similar to each other than to objects in other clusters [7][2][5].

II. DIGITAL FORENSIC XTENSIBLE MARKUP LANGUAGE (DFXML)

Digital Forensic Xtensible Markup Language or the DFXML is a standard structured way to exchange forensic information between forensic tools. It stores information of various digital forensics objects. It provides structure to store information of persistent and volatile data. It includes storage structure for Windows Registry, Running Processes, Event Information and File and directory information. These structures are specially designed to store digital forensics information. Hence, Digital forensic investigators can use that information to predict the nature of evidences and use to prove committed crimes. The paper focuses on pointing out malicious/ suspected files in form of clusters using the K-Means algorithm. Here, the evidences are extracted from malicious/ suspected nodes into a DFXML format. There is a file namely "FileInformation.dfxml" which has all the details for the files running/ ran on the computer system [3][4]. This file is analyzed and investigated by the investigators to gain further knowledge.

III. PROPOSED METHODOLOGY AND IMPLEMENTATION

In order extract knowledge from persistent data, FileInformation.dfxml file must be analyzed. It contains various attributes of files and directories from a computer system. Various attributes that represent file information includes filename, filesize, Modified DateTime, Accessed Date Time, Created Date Time etc. The research focuses on analysing from extracted digital forensic information in order to gain knowledge.

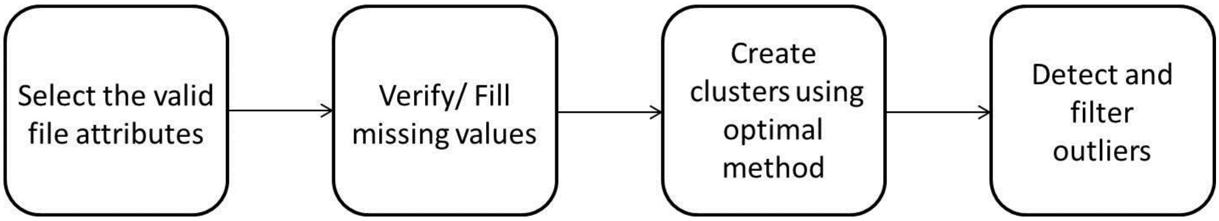
Select the valid file attributes
Verify/ Fill missing values
Create clusters using optimal method
Detect and filter outliers

Figure 1: File Meta Data Analysis after extracting evidence in DFXML

Below are the steps to create knowledge visualization from available digital forensic information evidences of file/directory metadata as shown in Figure 1.

1. Select the valid attributes: Name of file may not impact much to prove committed crime. Hence, it can be removed during analysis. Created Datetime, Modified Datetime and CreatedDatetime attributes of various files may be used to prove the crime. Retention of those attributes may be very useful.

2. Verify/ fill missing values: Verify whether any missing values in available forensic information as per shown in Figure 2.

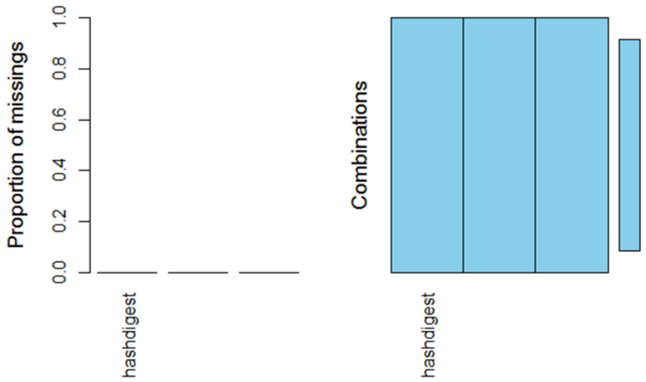

Figure 2: Missing Value Validation

3. Create clusters using optimal methods: Create clusters based on attributes Created Datetime and Accessed DateTime that shows usual file creation and Accessed time during working hours of users. Here, We have chosen K-means clustering algorithm to cluster homogeneous data points.

K-Means: The idea is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance. The algorithm consists of two separate phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids. Once we find k new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop. As a result of this loop, the k centroids may change their position in a step by step manner. Eventually, a situation will be reached where the centroids do not move anymore. This signifies the convergence criterion for clustering [8][9][10].

4. It would be helpful to get insight about outlier files. There are various methods to choose the optimal number of clusters from which the research focuses on the Elbow Method. The “elbow” method helps data scientists select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point [6].

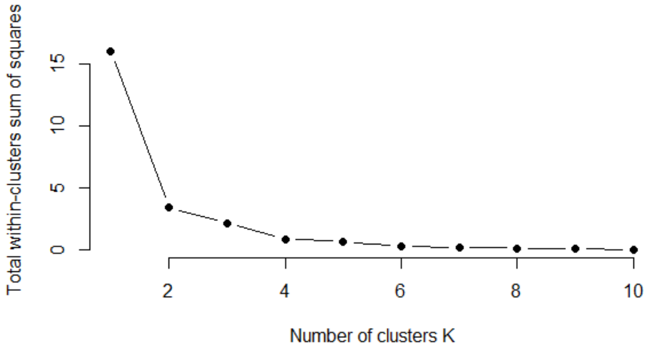

Figure 3: Optimal number of Cluster finding using the Elbow Method

As shown in Figure 3, sum of squares for each data point has been calculated for each value of k, ranges between 2 to 10. The graph plots sum of squared error against number of clusters. For the value k=2, the sum of squared error would be higher compared to the value k=3. Hence, We should increase number of clusters till the rate of change of SSE(Sum of Squared Error) rate would be less as compared to other values of k. Therefore, value k=4 would be the key point, thereafter which the fluctuation in SSE is very negligible. To avoid the problem of overfitting, we choose k=5 instead of k=4. For chosen optimal value k=5, We have plotted clusters based on their attributes access time and modified time.

5. Detect and filter outliers: It mainly focuses on unusual values of attributes. It gives direction to digital forensic

investigators about unusual timing of file accessed, modified or created with malicious intentions. It may be done by manual or automated by running script.

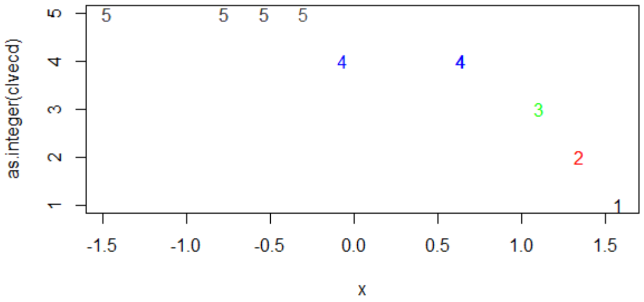

Figure 4: Clustering based on AccessedDatetime attribute after cleaning data

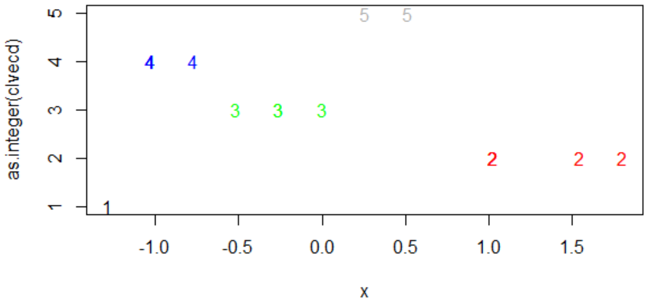

Figure 5: Clustering based on ModifiedDatetime attribute after cleaning data

Figure 4 concludes that digital forensic investigators can focus more on cluster number 1,2 and 3 rather than 4 and 5 whereas Figure 5 shows that cluster number 1 and 5 to be focused for digital investigation process.

IV. ANALYSIS

Traditional approach perfoms sequential steps to compare each data point and then give output based on provided conditions. Using traditional approach the time complexity of algorithm would be O(n) where n is the number of data points in the dataset.

Table 1: Performance Comparison

Attribute	Traditional Approach (Minimum No. of files to be analyzed)	Proposed Approach (Minimum no. of files to be analyzed)	Efforts reduced
AccessedDateTime	102	3	97.05%
ModifiedDateTime	102	6	94.11%

V. CONCLUSION

When you apply machine learning Un-supervised clustering methods K-means on the evidences extracted from the suspicious nodes, the automated clustering that happens, better predicts the files which need to be studied and pondered into depth as they are differentiated by being outliers. The remaining files can be assumed to be white hatted and ignored currently from being taking into analysis.

REFERENCES

[1] Shruti B. Yagnik, "Requirements to Build a System that Uses Machine Learning Based Approach for Analysis of Forensic Data"*International Journal of Computer Trends and Technology (IJCTT)*,V4(4):927-932 April Issue 2013 .ISSN 2231-2803.www.ijcttjournal.org. Published by Seventh Sense Research Group.

[2] Clustering-Kmeans" https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html(Accessed on 15/04/17)

[3] Esan P. Panchal "Extraction of Persistence and Volatile Forensics Evidences from Computer System"*International Journal of Computer Trends and Technology (IJCTT)*,V4(5):964-968 May Issue 2013 .ISSN 2231-2803.www.ijcttjournal.org. Published by Seventh Sense Research Group.

[4] Premal C. Patel "Aggregation of Digital Forensics Evidences"*International Journal of Computer Trends and Technology (IJCTT)*,V4(4):881-884 April Issue 2013 .ISSN 2231-2803.www.ijcttjournal.org. Published by Seventh Sense Research Group.

[5] Machine Learning, T.M. Mitchell, McGraw Hill, 1997.

[6] "Elbow Method", http://www.scikit-yb.org/en/latest/api/cluster/elbow.html (Accesses on 26-Nov-2017)

[7] Sejal Jaiswal, "K-Means Clustering in R Tutorial" Online: https://www.datacamp.com/community/tutorials/k-means-clustering-r , March 14th 2018.

[8] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006.

[9] Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006.

[10] K. A. Abdul Nazeer, M. P. Sebastian "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm" Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K., ISBN: 978-988-17012-5-1