International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 5, Issue 05, May-2019

A SYSTEMATIC APPROACH TO STUDY THE SECURITY BEHAVIOR FOR DATA MINING

Mohammad Iquebal Akhter [1]

Research scholar [1]
Prof. Mohammed Gulam Ahmad [2]
Nawab Shah Alam khan College of Engineering and Technology, Hydrabad [1,2]

Abstract: While enabling individuals to elicit hidden knowledge on the one hand, data mining techniques pose many privacy threats on the other hand. Data mining technology has many uses in malware detection. Classification and clustering methods are one of the most common data mining techniques. In this paper, we propose a data mining taxonomy for detecting malware behavior. This article examined some of these issues and described in detail the application of various data mining techniques to provide security. The most effective use of data mining is intrusion detection. Various data mining techniques can be used to effectively detect and report intrusions in real time, in order to take the necessary precautions to prevent intruder attempts. This article discusses privacy protection, anomaly detection, and classification.

Keywords: Data mining, Classification, Clustering, Privacy Preservation, Outlier Detection, Anomaly Detection

I. INTRODUCTION

Securities in data mining is a very important research area. In this article, we will further investigate the behavior of data mining tools and securities. Data mining is one of four identification technologies used today to identify malware. The other three are filtering, mobile observation, and credibility checking. When building security applications, engineers use information mining techniques to increase the speed and nature of malware detection while increasing the number of zero day attacks detected. Data mining (DM) is the process of extracting useful and useful information from large amounts of data, analyzing the information, and discovering useful patterns using various technologies. It is applied to a variety of applications such as healthcare, healthcare, marketing, finance, privacy, security and more. Security applications can be used for national security to fight against terrorist attacks and cyber security, protect computers and networks from corruption (worms and viruses), intrusions, botnet attacks, malware, denial of service (DoS). Application of non-real-time technologies such as classification, prediction, link analysis etc. to identify the possibility of future attacks by tracking the virus. Real-time technology is suitable for widespread use of data [1]. Communication and Sharing on the Network Data on the Internet increases the risk of cyber-attacks such as data corruption, network degradation, and unauthorized access to sensitive information. Because 3G / 4G technology exposes the IP (Open Protocol), these networks can allow network attackers to break into services and cause problems for end users and mobile operators. Network attackers can steal user data such as IMSI numbers, billing information, contact information, etc., downgrade networks via DoS, or interrupt or interrupt the service of a host connected to the Internet You [2]. Data mining has evolved into a great technology for addressing these security threats. Much research has been done in the literature to detect security issues, vulnerabilities, intrusions, malware etc. This article provides a comprehensive overview of existing security-related data mining techniques by using data mining techniques to make security decisions.

II. DATA MINING TECHNIQUES

Several techniques are used in data mining, such as classification, clustering, link analysis, and association rules. In general, data mining techniques fall into two categories: descriptive and predictive. Descriptive categories provide information (e.g., classification) from the data itself, and predictive categories extract information found based on previous data (e.g., clustering). Next, we will mainly discuss classification and clustering methods and algorithms, and briefly introduce link analysis and association rules processes for network attack detection.

A. Classification

Classification is the process of dividing data into different classes. These classes are predetermined and managed. In other words, the set of possible classes can be known in advance. There are different techniques consisting of different algorithms for classification. [3].

B. Decision tree (DT)

The decision tree has a tree structure including nodes. Nodes without input edges are called roots, and all other nodes are called leaves. Each node in the tree is a decision that independently defines the output of the decision result, and each leaf is marked with a class based on the goal value. In addition, there is a probability vector that indicates the target probability. [5, 6].

C. Support vector machines (SVM)

This method divides data samples into two groups, positive and negative. Support vector machines find hyper planes in dimensional space separating two classes with maximum margins. The new input data belongs to one of two classes.

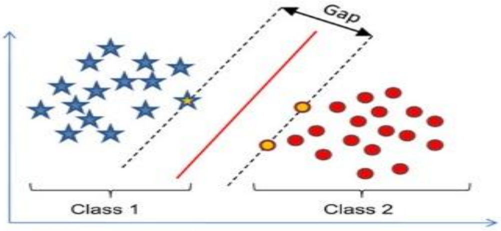


Fig. 1: *Support vector machines (SVM)*

d. SVM

If (xi, yi) are training data from two classes which $yi = \pm 1$ then SVM finds a hyper-plane that can separate the two classes the best: $f(x) = w^t\Phi(x) + b = 0$; Where f(x) represents the discriminant function associated with the hyper-plane and WTϕ(x) is a nonlinear kernel function that maps the input xi into a higher-dimensional space.

e. K-Nearest neighbor (KNN)

Classifiers are based on the similarity of new entries to sample data. The query examples are grouped into different groups in the multidimensional feature space, and the similarity is based on the distance (nearest neighbor) between the two samples. Figure 5 shows an error situation where the black sample is marked as positive because the nearest neighbor is misclassified. [10].

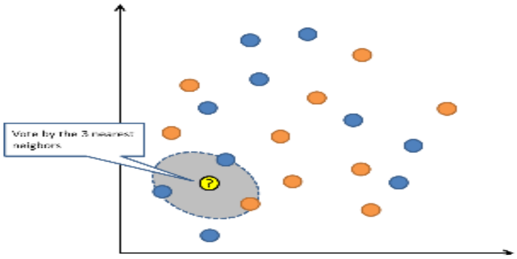


Fig. 2. Error case in k-nearest

f. Clustering

A cluster distributes data to multiple groups, and objects with similar profiles or attributes are assigned to the same group. Data mining helps to analyze correlations between attributes. Objects belonging to one cluster are similar to and different from objects belonging to another cluster. Clustering is an unmonitored process. That is, the capabilities of the cluster are unknown in advance and need to be discovered during the clustering process. Clustering methods fall into five categories: partitioning, layering, density-based, grid-based, and model-based [11]. The following is a brief introduction to some of the main clustering methods.

g. Fuzzy clustering (grid based)

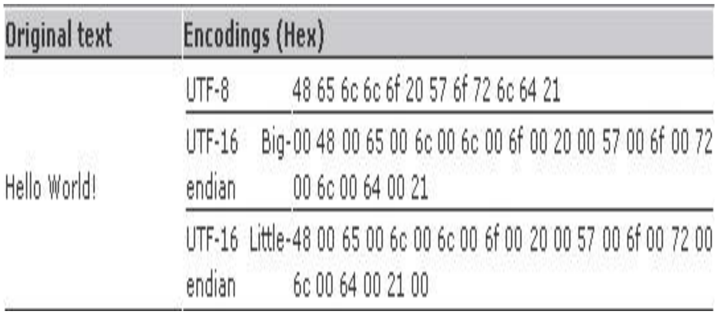
This algorithm belongs to soft clustering. That is, each object is part of a particular level cluster, up to a likelihood factor. This concept is the opposite of hard clustering, where each object belongs to only one cluster. Fuzzy clustering algorithms calculate these likelihood factors and assign objects to one or more clusters [12].

```
If (data >a) && (data <b)
prob = data – a/(b-a)
if ( data >=b and data <=c) prob = 1;
if (data >c and data <d)
prob = d – data/(d-c)
else: prob =0
```

There are various algorithms for this method, with fuzzy c-means (and improved versions of fuzzy c-means) being the most common algorithm. Fuzzy c-means and c-means algorithms have very similar functions [12].

h. Genetic Clustering Algorithm (GCA)

GCA is an adaptation algorithm based on biological evolution, natural selection and genetics. GCA mimics the concept of "strongest or most appropriate for gene survival". In nature, each generation consists of DNA, and each DNA consists of a chromosome that contains a gene. In this algorithm, strings act as DNA as chromosomal attributes (ie, rules) and parameters similar to genes (ie, individuals). Consider the example of network attack detection, the input data is divided into several strings. The goal is to mark each string as traffic or attack based on different attributes. In this example, attributes such as source address, stream size, and session time are like chromosomes. Finally, there are parameters for each property. Each of these parameters has a probability of binary format (0, 1), and the combination of the probabilities of all parameters is the attribute probability. For source address attributes, the parameters can be different packet sizes for source IP address, source port, stream size [19, 20]. In FIG. 7, the calculation parameters of a, b, c and d in the fuzzy rules of FIG. 7 are used in the genetic algorithm. Each string of GCA is 12 blocks and there is a class at the end (displays the type of string). Each block is a rule or feature that includes parameters a, b, c, d (e.g., the number of tcp packets) (e.g., a = 2, b = 3, c = 4, and d = 5). Each parameter can be attacked, and all block probability combinations indicate whether a string (record) has been attacked. α / A-β / B where A is the total number of attack records, B is the total number of regular records, and α is the total number of correctly identified attack records. The total number of regular records (false positives) that are classified as attack (true positives) and beta attacks [13].



Original text	Encodings (Hex)	
Hello World!	UTF-8	48 65 6c 6c 6f 20 57 6f 72 6c 64 21
	UTF-16 Big-endian	00 48 00 65 00 6c 00 6c 00 6f 00 20 00 57 00 6f 00 72 00 6c 00 64 00 21
	UTF-16 Little-endian	48 00 65 00 6c 00 6c 00 6f 00 20 00 57 00 6f 00 72 00 6c 00 64 00 21 00

Fig. 3. String encoding

GCA is based on the concept of an intragenic evolutionary process to find the strongest gene. Fitness scores are assigned to parameters that can be competed and won. In other words, genetic algorithms are simulated to look for optimistic cluster centers in feature space. Genetic clustering algorithms have different stages in which fitness calculation, crossover and mutation are the most important stages [16, 17].

i. Hierarchical CURE clustering

This algorithm is an excellent solution to heterogeneous (formation) clustering and anomalous problems. The algorithm uses a set of randomly chosen samples rather than using only the centroids in each cluster. For each cluster, select those well-distributed points as cluster representatives. Through the layering process, those clusters with closest representative points are merged together and the other clusters are eliminated. Through the fractional factor, the representative point towards the centroid contraction is used as the final data [18].

j. Swarm Intelligence (SWI)

SWI is an algorithm that can track each other and find the best solution based on interactions between social insects. Ant colonies and ant minors are one of many social behaviors surveyed by many researchers. People notice that ants are looking for the shortest food from their nests. Initially, each ant chooses a different path, but the next few rounds will follow the ants to reach the nest faster. In other words, move in the shortest way. On the other hand, ants have been shown to pick up carcasses from their nests and place them densely outside, depending on the size of the carcass. At first, I did not know where to put my body, just did it at random, and then I searched for the closest cluster to my body. In addition to simplicity, this behavior requires a dynamic process and positive feedback to find a quick solution. This is a good practice and can be used as one of the clustering algorithms. However, it is not very convenient to formulate properties for selection. The clustering of this algorithm is based on similarity and learns from other experiences**.**

Based on the similarity between the input data and surrounding objects, decide whether to discard the object (for example, as normal traffic for attack detection) or pick it up (for example, as an attack). There is no rule at first, but it will be developed in the process. In the learning phase, each object that does not match the rule is deleted and the other objects add attributes to the rule until all variables are used. Picking and dropping are probabilistic actions based on the similarity factor for the nearest cluster in the neighborhood. The more ants carry their corpse to the nearest group, the more likely they will be dropped [19].

III. Data Mining Technology

Data mining is defined as the activity of discovering interesting patterns from a large amount of data. Data can be stored in databases, data warehouses, or other information bases [20, 21]. Data mining is interdisciplinary, such as databases and data warehousing techniques, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, spatial or temporal data analysis, etc. Technology integration]]. Data mining processes are very useful for extracting and understanding patterns from a large number of imperfections, noise, instability, fuzzy, and random data. It is science that allows the system to extract useful information from large datasets and databases [23]. As such, data mining techniques extract implicit, unknown and potentially useful information from data [24, 25, and 26]. Intrusion detection systems generate important information from a variety of sources, including host logs, network packets, and system-specific applications. The use of data mining techniques in intrusion detection is very important [27, 28], as data mining techniques such as data and alarm information are very advantageous in data extraction characteristics and rules. The main idea behind data mining tools is to discover abuse detection rules or anomaly detection models by analyzing network data and host call data. The importance of applying data mining techniques to intrusion detection is that it can provide a model that can help audit the data to accurately capture real intrusions and normal behavioral patterns [27].

Therefore, the main advantage of using data mining techniques is that the same data mining tool can be applied to many data streams. Therefore, building a powerful intrusion detection system is an advantage. However, how to distinguish between normal and abnormal behavior from many raw data attributes and efficiently generate automatic intrusion rules immediately after collecting raw network data is the biggest challenge affecting intrusion detection systems. Correlation analysis algorithms can reduce the negative impact on the problem and help you find attribute relationships in the network connection record. A sequence analysis algorithm can discover the time relationship of network connection records, and correlation analysis and sequence analysis algorithms can be used to obtain a successful anomalous intrusion detection. Operating mode being used. In addition, data mining algorithm classifications can be used to generate mining rules from trained data sets. This will help identify normal operation and intrusion [27].

IV. Classification and Detection Using Data mining Techniques

Malware computer programs that copy themselves and spread from one computer to another are called worms. Malware includes malicious code such as worms, computer viruses, trojans, keyloggers, adware and spyware port scan worms, UDP worms, http worms, user to root worms, and remote to local worms. [28]. Attackers create these programs for a variety of reasons, including interruption of computer processes, collection of confidential information, or intrusion into private systems. Detecting worms on the Internet is important. It creates weak points and reduces system performance. Therefore, worms must be detected as they occur and be classified before being damaged using data mining classification algorithms.

Classification algorithms that can be used include random forests, decision trees, and Bayesian [29]. Most worm detection technologies use intrusion detection systems (IDS) as the rationale. Automatic detection is difficult because it is difficult to predict what form the next worm will take. IDS can be divided into two types: network-based IDS and host-based IDS. Network-based intrusion detection systems reflect network packets before they are propagated to end hosts. Host-based intrusion detection systems, on the other hand, reflect network packets propagated to end hosts. In addition, host-based testing encodes network packets so that they can trigger Internet worm attacks. When focusing on unencoded network packets, it is necessary to investigate the traffic performance in the network. Several machine learning techniques have been used in the area of intrusion and worm detection systems. Therefore, data mining, especially machine learning technology, plays an important role and is essential in worm detection systems. Several new techniques have been proposed to build several intrusion detection models using different data mining schemes. Using decision trees and machine learning genetic algorithms, you can learn abnormal and normal patterns from the training set, generate classifiers based on test data, and mark them as normal or abnormal classes You Data marked as an exception may be a pointer that indicates the presence of an intrusion.

V. Simulative Result

Detection Process Using Data Mining Technology: Detection Process
When using information mining, malware detection consists of two phases.

- Take out features
- Classifying/clustering

The first step is to separate the various highlights (API calls, n-grams, double strings, program execution, etc.) statically and incrementally to understand the characteristics of the document test. Highlight extraction can be performed by performing static or dynamic investigations (with or without actually performing destructive programming that can be imagined).

Hybridization methods that integrate static and dynamic testing can also be used. During the characterization and collection process, the recording tests are organized into parties in the order in which they are highlighted. You can use grouping or aggregation methods to schedule tests. In order to group document tests, checks are made to create alignment models (classifiers) such as decision trees (DT), artificial neural networks (ANN), naive Bayes (NB) or support vector machines (SVM) You need to use the sequential calculation of.) etc. Bunching is used to collect malware tests with similar characteristics. Use machine learning techniques to develop models that meet both good and bad categories. If you use this document to test your cumulative preparation classifier, you will be able to identify even recently released malware. Note that the feasibility of data mining information systems for malware identification is fundamentally dependent on the highlights of your separation and the classification techniques used. This article contains two standard data sets for investigating data mining security behavior, as shown below.

- Spam filtering Data sets
- Intrusion detection Data Sets
- Span filtering data sets

This article introduced the application of machine learning methods in automatic filtering of spam. One of the main challenges in building a system-based stop recognition framework is collecting information about the preparation framework. Nevertheless, some data sets, such as KDD Cup 99, are collected and published, and these data sets are old and inconsistent with the general framework. This paper presents an online structural planning interrupt identification data set. To detect host-based attacks, you need to examine the highlights released by the project and break down system traffic to identify organization-based attacks. More importantly, this is almost the same as the detection of malware, and you can look for cases of abnormal behavior or abuse. The following figure shows a graphical user interface built using MATLAB.

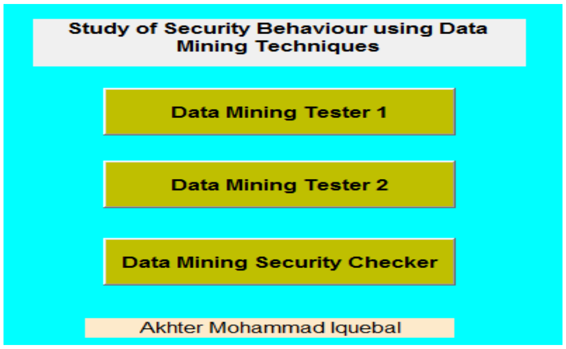
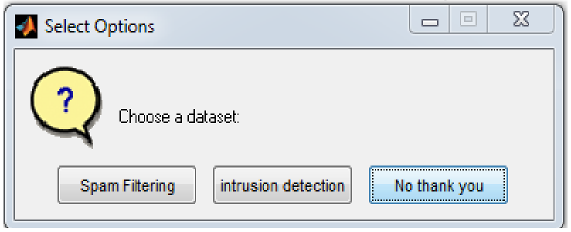

Study of Security Behaviour using Data Mining Techniques
Data Mining Tester 1
Data Mining Tester 2
Data Mining Security Checker
Akhter Mohammad Iquebal

Fig 4: Project Basic GUI of MATLAB

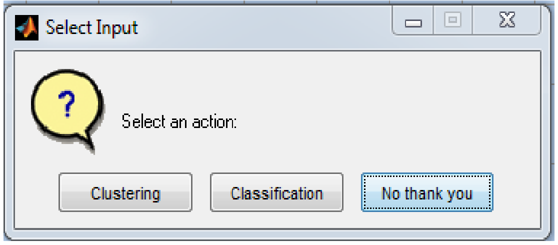
The figure above shows three buttons. Each has its own data mining meaning. The first two are simple tools to calculate data and estimate cluster groups. This requires two databases to represent clustering and separation of data sets.



Choose a dataset:

Fig 5: Choose Dataset Spam Filtering

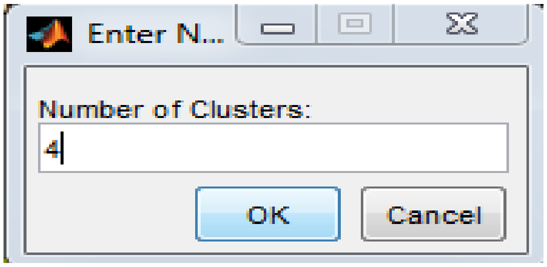
Users can select span filtering and intrusion detection databases. Therefore, there is no other button to select a database. Both datasets have different properties. However, there are some very common data sets that can be used to research securities.



Select an action:

Fig 6: Apply Clustering on Spam Filtering

The user has to select clustering or classification again. This is a more important task for data mining research. Click the appropriate button to expect your choice.



Number of Clusters: 4

Fig 7: Choose 4 Clusters for Spam Filtering

After pressing the cluster button just before layout, the user is selected as 4 clusters. Based on the data, the results are displayed in four similar data separation modes.

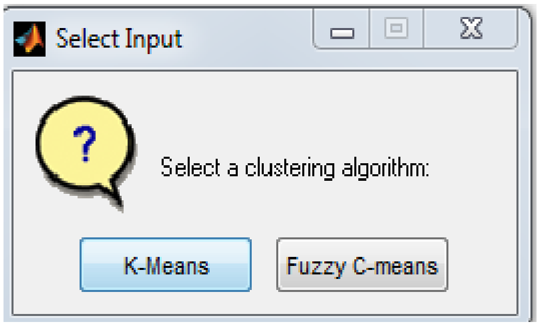
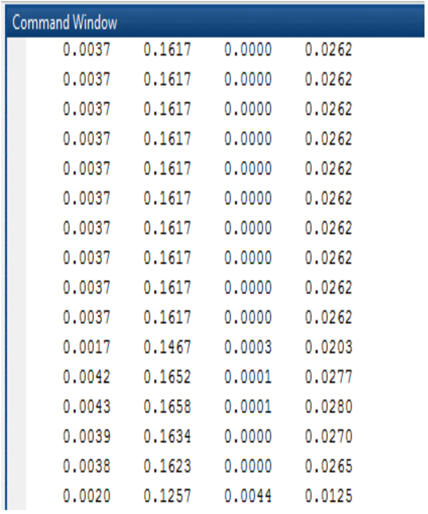


Fig 8: Apply K-means Algorithms over Cluster of Spam Filtering

Furthermore, the user can choose the type of clustering as K-mean or fuzzy C-mean, which is the selection criterion of the clustering algorithm. Basically very small difference between K mean and fuzzy C mean. According to the required data or experiments, the application of these buttons was applicable.



```
    0.0037    0.1617    0.0000    0.0262
    0.0037    0.1617    0.0000    0.0262
    0.0037    0.1617    0.0000    0.0262
    0.0037    0.1617    0.0000    0.0262
    0.0037    0.1617    0.0000    0.0262
    0.0037    0.1617    0.0000    0.0262
    0.0037    0.1617    0.0000    0.0262
    0.0037    0.1617    0.0000    0.0262
    0.0037    0.1617    0.0000    0.0262
    0.0037    0.1617    0.0000    0.0262
    0.0017    0.1467    0.0003    0.0203
    0.0042    0.1652    0.0001    0.0277
    0.0043    0.1658    0.0001    0.0280
    0.0039    0.1634    0.0000    0.0270
    0.0038    0.1623    0.0000    0.0265
    0.0020    0.1257    0.0044    0.0125
```

Fig 9: Results of 4 Cluster of Spam Filtering

As mentioned above, MATLAB shows four columns very clearly. These four columns show many data sets. In the actual implementation, the analysis should focus on a specific data trending group, as the data needs to be examined.

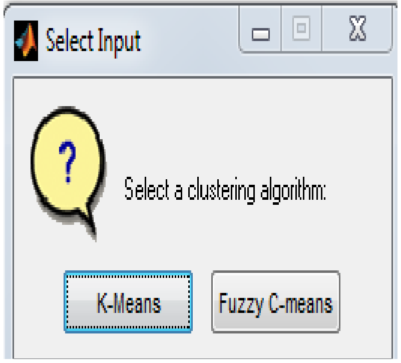
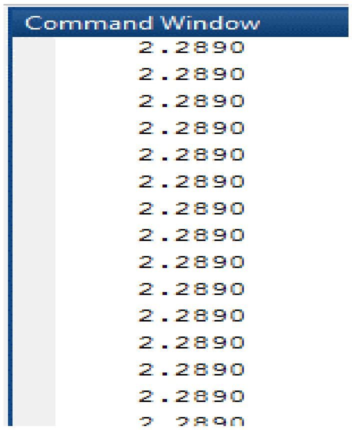


Fig 10: Apply Fuzzy c-means Algorithms Spam Filtering data

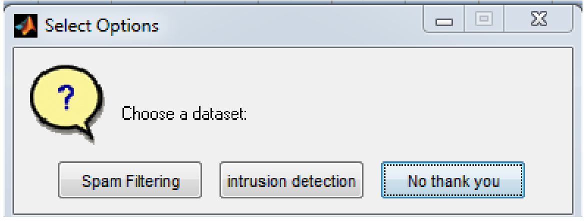
Similarly, C-means is also applied to the span filtering data set. The results were output from MATLAB, as shown below as a screenshot.



Command Window

```
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
    2.2890
```

Fig 11: getting One Cluster after Apply Fuzzy c-means Spam Filtering data

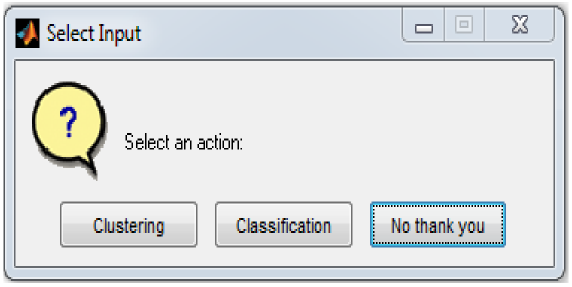


Select Options

Choose a dataset:

Fig 12: Select Intrusion detection dataset

Similarly, intrusion detection databases are expected to have very different attributes. Also, press the button and then instruction follow.

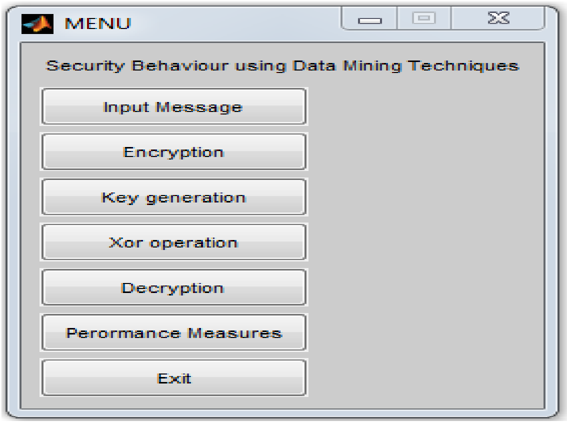


Select Input

Select an action:

Fig 13: Select Clustering for Intrusion detection

In addition, a cluster for intrusion detection was run and similar results were found. Now will go to the third button. This is called a security checker. When you execute this button, the result is as follows:

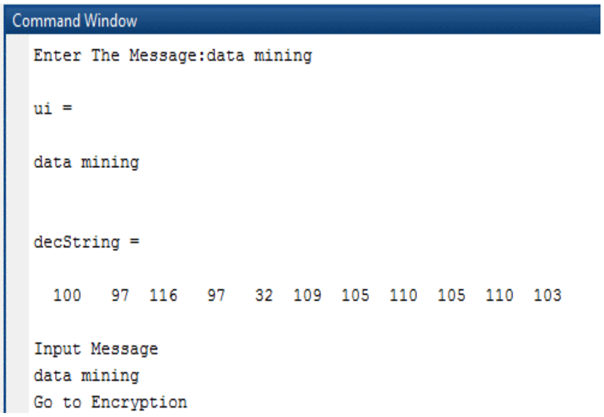


MENU

Security Behaviour using Data Mining Techniques

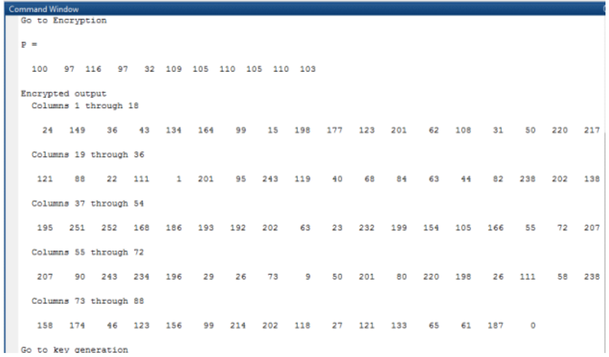
Fig 14. Security behavior using data mining techniques

The screen above shows seven different buttons for security inspector data. When you press the Enter Message button, you need to enter a test message at the MATLAB command prompt. Then proceed to "Encryption" and change the data input to encrypted form. It then goes to key generation, which is essential for the recipient of the user message as a test message recipient.



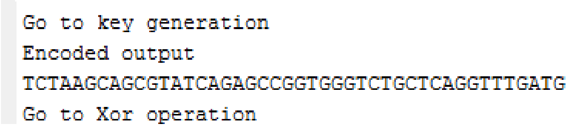
```
Enter The Message:data mining

ui =

data mining


decString =

  100   97  116   97   32  109  105  110  105  110  103

Input Message
data mining
Go to Encryption
```

Fig 15: Enter message in command window



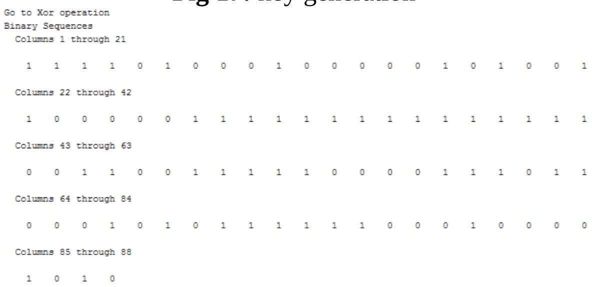
```
Go to Encryption

p =

  100   97  116   97   32  109  105  110  105  110  103

Encrypted output
  Columns 1 through 18

    24   149    36    43   134   164    99    15   198   177   123   201    62   108    31    50   220   217

  Columns 19 through 36

   121    88    22   111     1   201    95   243   119    40    68    84    63    44    82   238   202   138

  Columns 37 through 54

   195   251   252   168   186   193   192   202    63    23   232   199   154   105   166    55    72   207

  Columns 55 through 72

   207    90   243   234   196    29    26    73     9    50   201    80   220   198    26   111    58   238

  Columns 73 through 88

   158   174    46   123   156    99   214   202   118    27   121   133    65    61   187     0

Go to key generation
```

Fig 16: Encryption of message



```
Go to key generation
Encoded output
TCTAAGCAGCGTATCAGAGCCGGTGGGTCTGCTCAGGTTTGATG
Go to Xor operation
```

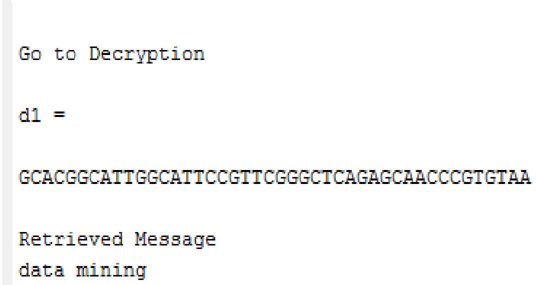
Fig 17: key generation



```
Go to Xor operation
Binary Sequences
  Columns 1 through 21

    1    1    1    1    0    1    0    0    0    1    0    0    0    0    0    1    0    1    0    0    1

  Columns 22 through 42

    1    0    0    0    0    0    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1

  Columns 43 through 63

    0    0    1    1    0    0    1    1    1    1    1    0    0    0    0    1    1    1    0    1    1

  Columns 64 through 84

    0    0    0    1    0    1    0    1    1    1    1    1    1    0    0    0    1    0    0    0    0

  Columns 85 through 88

    1    0    1    0
```

Fig 18: binary sequence

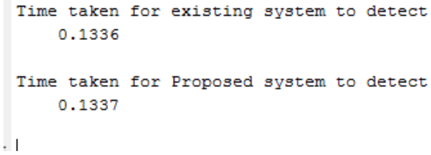
The data for the XOR operation is placed in the final binding of the encrypted text, which can be put in the user's channel to the recipient without testing the original form to prevent massage privacy.



```
Go to Decryption

d1 =

GCACGGCATTGGCATTCCGTTCGGGCTCAGAGCAACCCGTGTAA

Retrieved Message
data mining
```

Fig19: Retrieved message

The text message is then decrypted by the recipient user using the user provided key. Therefore, the data is also safe for the transmission line. This comprehensive study is only to investigate the safe behavior of data mining analysis.



```
Time taken for existing system to detect
    0.1336

Time taken for Proposed system to detect
    0.1337
```

Fig20: detect system

Advances in data mining, used to study safe behavior, have resulted in a huge collection of digital information repositories. The information is not stored but retrieved and the data manager is used to manage the information properly. Data mining technology is effectively used to obtain large amounts of information from outside. Data management is important for decision-making knowledge to support strategic decisions in information technology and information systems within an organization. Data information extraction is a method of freeing learning and sampling a large amount of information, linking the universe with databases, insights and artificial intelligence to obtain big data information.

It uses sophisticated algorithms to analyze large data sets and select relevant information. This article describes how to use two general purpose database spam filtering and intrusion detection datasets to examine the security behavior of data mining. This process is limited in many ways, and it partially fills in the overall analysis of securities behavior. This paper studies a wide field of data mining analysis security behavior research. Various data mining techniques are integrated into the MATLAB GUI as executable buttons to obtain user-selected databases and perform detailed analysis of the databases. Two button contracts for a tool to check the cluster with the appropriate user input required to analyze the entire data set. The last button is used for stock check for final adjustment. The security information mining calculations presented in this work show the elite. In any case, the future scope of the theory is below:

• The models built are semi-legal and obey the rules. In any case, you can adjust the calculations to deal with malicious enemies. Malicious enemies are those who deviate from the tournament.

•An encryption framework such as Paillier, ElGamal, etc. to perform secure calculations. These frameworks can be further studied and designed to incorporate improved attributes to perform fast calculations on encrypted data information.

• The proposed calculations can be stretched out to mine cloud data information. Mining can be performed on the re-appropriated databases, where various information proprietors proficiently share their information without bargaining the security of the data information.

The overall conversion rate is also calculated. The overall time of the proposed system was evaluated for inspection performance (in accordance with a standard real-time system).

VI. Major challenges

The main challenges are to perform safe calculations, plan calculations and improve efficiency. This calculation is considered accurate because it produces high accuracy, low communication costs, low computational costs, and high accuracy. Protection The protection arrangement uses the Naïv eBayesian model to propagate information to multiple meetings in one meeting. Compared to well-known methodologies, the constructed classifiers become increasingly safer and more efficient. Nevertheless, the highlights of the dissemination of information differ among all participants. Then, protection information mining was planned and created to protect vertical propagation information.

VII. Conclusion and Future Scope

Data information is communicated in a variety of ways. Each of these associations violates protection rules once information is found. Nevertheless, these associations need to mine basic examples or end based on their joint information and data mining. A possible arrangement is to mine information by maintaining the security of sensitive information. The information sent can be brighter highlights or unique highlights. In information mining, order models expect classes to directly depend on highlighting. Bunching is another information mining strategy for gathering comparisons. This data conforms to the data mining data extraction attributes and shows that it needs to be applied to intrusion detection systems. Various detection models require registration data as a training package. This has a major impact on intrusion detection systems. Because of the strength and accuracy of the visitor network, it is difficult to get completely attacked. Furthermore, it is very inconvenient to record attack behavior. The problem of data extraction techniques can be solved when analyzing the public access of the network. Because isolated points are gas behavior, they can reduce the difficulty of acquiring training data. Data mining techniques can be used, for example, as compilations, classifications, feature summaries, and participation rules in intrusion detection systems. Data extraction techniques have been shown to improve intrusion detection, processing speed and reduce the speed of error messages.

References:

1. Thuraisingham, B. (2009, September). Data mining for malicious code detection and security applications. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 02* (pp. 6-7). IEEE Computer Society.
2. Ahmed, F., Rafique, M. Z., & Abulaish, M. (2011, December). A data mining framework for securing 3g core network from GTP fuzzing attacks. In *International Conference on Information Systems Security* (pp. 280-293). Springer, Berlin, Heidelberg.
3. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.
4. Heckerman, D. (1997). Bayesian networks for data mining. *Data mining and knowledge discovery*, *1*(1), 79-119.
5. Fenton, N., & Neil, M. (2012). *Risk assessment and decision analysis with Bayesian networks*. CRC Press.
6. Rokach, L., & Maimon, O. (2005). Decision trees. In *Data mining and knowledge discovery handbook* (pp. 165-192). Springer, Boston, MA.
7. Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627-2636.
8. Beghdad, R. (2008). Critical study of neural networks in detecting intrusions. *Computers & security*, *27*(5-6), 168-175.
9. Singh, Y., & Chauhan, A. S. (2009). Neural Networks in Data Mining. *Journal of Theoretical & Applied Information Technology*, *5*(1).
10. Eschrich, S., Ke, J., Hall, L. O., & Goldgof, D. B. (2003). Fast accurate fuzzy clustering through data reduction. *IEEE transactions on fuzzy systems*, *11*(2), 262-270.
11. L. Rokach and O. Maimon, "Clustering methods," in *Data Mining and Knowledge Discovery Handbook*Anonymous Springer, 2005, pp. 321-352.
12. Havens, T. C., Bezdek, J. C., Leckie, C., Hall, L. O., & Palaniswami, M. (2012). Fuzzy c-means algorithms for very large data. *IEEE Transactions on Fuzzy Systems*, *20*(6), 1130-1146.
13. Jongsuebsuk, P., Wattanapongsakorn, N., & Charnsripinyo, C. (2013, January). Network intrusion detection with Fuzzy Genetic Algorithm for unknown attacks. In *The International Conference on Information Networking 2013 (ICOIN)* (pp. 1-5). IEEE.
14. Wajrock, S., Antille, N., Rytz, A., Pineau, N., & Hager, C. (2008). Partitioning methods outperform hierarchical methods for clustering consumers in preference mapping. *Food quality and preference*, *19*(7), 662-669.
15. P. Tan, M. Steinbach, V. Kumar and Ghosh, "The k-means algorithm," in *Internet Computing, IEEE,* 2002,
16. Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern recognition*, *33*(9), 1455-1465.
17. T. Jea, "Basic concept of data mining, clustering and genetic algorithms,".
18. S. Guha, R. Rastogi and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases," *SIGMOD Rec.*, vol. 27, pp. 73-84, jun, 1998.
19. Handl, J., & Meyer, B. (2007). Ant-based and swarm-based clustering. *Swarm Intelligence*, *1*(2), 95-113.c
20. The architecture of a network level intrusion detection system. Technical Report, R. Heady, G. Luger, A. Maccabe, and M. Servilla. Computer Science Department, University of New Mexico. August 1990.
21. Kovac, S. (2012) Suitability analysis of data mining tools and methods. Bachelor’s Paper, Faculty of Informatics, Masaryk University.
22. http://loremate.com/
23. Hand, D., Smyth, P., Mannila, H. (2001) Principles of Data Mining, MIT Press.
24. Fayyad, Usama; Gregory Piatetsky – Shapiro, and Padhraic Smyth (1996). “From Data Mining to Knowledge Discovery in Databases”.
25. Frawley, W., Piatetsky-Shapiro, G., Matheus, C. (1992) “Knowledge Discovery in Databases: An Overview”, AI Magazine, pp. 213-228.
26. Siddiqui, M.A. (2008), PhD Paper, “Data Mining methods for Malware Detection”, University of Central Florida, Orlando, Florida.
27. LI Min, Application of Data Mining Techniques in Intrusion Detection, a Yang Institute of Technology.
28. Rothleder, Neal. “Data Mining for Intrusion Detection”. The Edge Newsletter
29. M. Siddiqui, M. C. Wang, and J. Lee, “Detecting Internet Worms Using Data Mining Techniques,” *Journal of Systemics, Cybernetics and Informatics*, vol. 6, no. 6, pp. 48–53, 2009.
30. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip*et al.*, “Top 10 Algorithms in Data Mining,” *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1– 37, 2008