Privacy Preservation in Data Mining with Cyber Security

Jisha M M
St. Joseph's College,
Irinjalakuda,

***Abstract* – Correlation of the Privacy Preserving Data Mining techniques and the algorithms used in cyber security provides high level protection. Minnesota Intrusion Detection System (MINDS) is introduced in this paper. Data mining can play an increasingly important role in ensuring cyber security, as latest facilities are building into the present data mining techniques to provide solutions such as intrusion detection and auditing. This paper presents the idea of applying data mining techniques to intrusion detection systems to maximize the effectiveness in recognizing attacks, thereby helping the users to build more protected information systems.**

Keywords: Network intrusion detection; anomaly detection; summarization;, profiling; scan detection

I. INTRODUCTION

The main terrorist threats posed to our nation today are cyber attacks. Malicious mobile code circulation that can injure or leak sensitive files or other data and intrusion upon computer networks are the main divisions. Data mining is practiced on problems in intrusion detection and auditing. It is used to find unusual patterns and behaviors. Categorization may be used to group various cyber attacks and then use the profiles to detect an attack when it occurs. Estimations may be used to resolve potential future attacks depending in a way on information learnt about terrorists through email and phone conversations. A deep knowledge about data mining and applications of it, the techniques used in cyber security and the existing devices used in this field gives us an idea to connect these two. MINDS - The Minnesota Intrusion Detection System, which uses data mining in cyber security is a model for it. A deep study about MINDS helps to find out the fields where updation can be performed for a proposed system to convert network traffic data into useful features, for building real-time intrusion detection system, modification of standard data mining algorithms for low frequency of computer attacks etc.

II. LITERATURE SUVERY

a) The paper entitled "What is Data Mining, and How is it Useful for Power Plant Optimization? (and How is it Different from DOE, CFD, Statistical Modeling)"[1] describes data mining methods that are broadly accepted in a variety of business areas. This paper will present an opening to data mining, and in particular contrast the methods used in data mining with usual optimization techniques.

b) The paper entitled "Cyber Security: Threats, Reasons, Challenges, Methodologies and State of the Art Solutions for Industrial Applications"[5] highlights the general cyber threats and thorough analysis of open system and style used for its industrial resolutions. Several significant industrial functions is also examined in this paper.

c) The paper entitled "Data Mining and Cyber Security"[6] find out that the web applications were getting reputation for data storing and data sharing. Greater part of web applications has some type of design or progress fault which can be easily broken by the cyber criminals. At the moment the public networks, mobiles with internet connectivity, personal privacy, and the linked configuration of entities such as banks are the most enticing targets for cyber criminals. This paper highlights the common cyber threats and detailed analysis of existing system and methodology used for its industrial solutions. Some important industrial application is also analyzed in this paper.

d) The paper entitled "MINDS - Minnesota Intrusion Detection System"[3] is an outline of the Minnesota Intrusion Detection System (MINDS), which apply a set of data mining based algorithms to deal with diverse features of cyber security. The different mechanisms of MINDS like the anomaly detector, scan detector and the profiling module recognize dissimilar types of attacks and intrusions on a computer network. The intrusions noticed by MINDS are opposite to those of usual signature based systems, such as SNORT, which suggests that they both can be jointed to enlarge total attack coverage. MINDS has revealed enormous equipped success in identifying network intrusions in two live organizations the Interrogator architecture at the US Army Research Labs and the University of Minnesota.

e) The paper entitled "Minds: Architecture & Design"[2] introduces the Minnesota Intrusion Detection System (MINDS), which utilizes a collection of data mining techniques to mechanically detect attacks in opposition to computer networks and systems. Although the continuing aim of MINDS is to deal with all aspects of intrusion detection, this paper points on two definite contributions: (i) an unverified anomaly revealing technique (ii) an involvement pattern study. In addition

a lot of advanced characteristics when match up to with SNORT. Besides, given the very great volume of links experimented per unit time, association pattern support summarization of original attacks is rather useful in permitting a security analyst to recognize and characterize rising threats.

III. EXISTING SYSTEM

A. Data mining

What is Data Mining- Data mining discover correlations or patterns and trends that go beyond simple analysis by searching among dozens of fields in large comparative databases. Algorithms in Mathematics are used for this to segment the data and evaluate the probability of future events. It is also known as Knowledge Discovery in Data (KDD). For getting useful information, analyzing data from different perspectives and summarizing. From many different dimensions or angles, it allows users to analyze data and group it, and review the relationships identified.

Classes, Clusters, Associations and Sequential patterns are the four types Data Mining relationships. Presentation the data in a helpful format, such as a table or graph, study the data using application software, Providing data access to business analysts and information technology professionals, Mine and control the data in a complicated database system, mine, convert, and load transaction data onto the data warehouse system are the five major elements of Data mining.

Applications of Data Mining - Job understanding-Determining the job, Data understanding-collect and verify the quality of data, Data preparation-data selection, consolidation and formating, Process Modeling-generate test designs, model assessments, Process Evaluation-evaluates results and approve the model and Deployment-produce final reports, present documentation are the Data mining processes.

Data mining is widely used in the areas of Cyber Security, Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Other Scientific Applications, Data warehouses, Intrusion Detection, Sales/Marketing, Banking, Health Care and Insurance.

B. Cyber security

Fundamentals of Cyber Security- 1. Confidentiality: Controlling who gets to read data; 2.Integrity: encouraging that information and programmes are changed only in a specified and authorized manner; and 3.Availability: assuring that authorized users have continued access to information and resources. Common Reasons of Cyber Attacks are simple to access, ability to accumulate data in moderately small space, Complexity of code, Negligence, Loss of evidence are the common reasons of cyber attacks.

Overview Of Cyber Security Solutions- System And Methodologies - For cyber security, a variety of states of art mechanics exist in the form of Scanners, Intrusion Prevention System, Intrusion Detection System, Network and Application Firewall. These solutions are *Vulnerability Scanners, Intrusion Prevention System, Intrusion Detection System.*

An intrusion is defined as any set of actions that attempt to compromise the openness, privacy or ease of use of a resource. Two classifications for Intrusion detection. First one Misuse detection and the second one Anomaly detection. Ontology based IDS solutions are used in information security.

C. MINDS

The MINDS Project-A data mining based system for detecting network intrusions is The Minnesota Intrusion Detection System (MINDS).

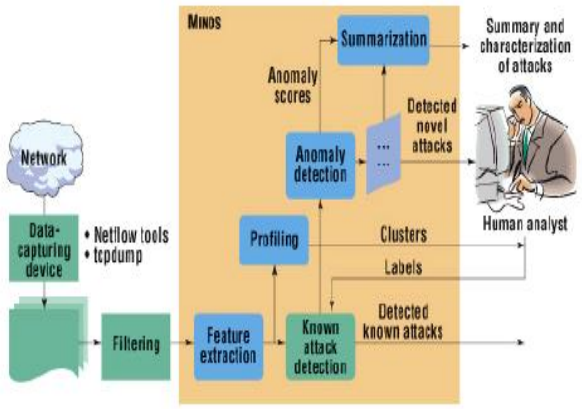

Figure 1. The Minnesota Intrusion Detection System (MINDS) [15]

In order to detect the successful data sources and to collect the full details of the data, The time-windows based features, extracted connection-window based features and Anomaly detection features can be used.

MINDS Anomaly Detection Module - The density based outlier detection scheme used in anomaly detection module is described in this section. The local outlier factor (LOF) is an outlier to each data point in MINDS anomaly detection module. Manipulating pair wise distances among all data points, is an $O(n^2)$ process. It makes calculations infeasible for millions of data points. Test a training set from the data and evaluate all data points to this small set, which reduces the complexity to $O(n*m)$. Where m is the size of the sample and n is the size of the data.

MINDS Anomaly Detection Results on Real Network Data

Figure 2 demonstrates a classic MINDS output after anomaly detection and summarization. The system sorts the links according to the keep count that the anomaly detection algorithm accredits them. MINDS examine anomalous links with the most scores, by the patterns that the link analysis module generates. Every line contains the middling anomaly count, the number of links indicated by the line, essential eight link faces, and the relative input of each basic and resulting anomaly detection feature. In figure 2 the second line represents 138 anomalous links. From this review, analysts can simply deduce that this is a

backscatter from a denial-of-service assault on a computer that is outside the network being analyzed. Figure 2 indicates the analysts' measurements of several other summaries the system found.

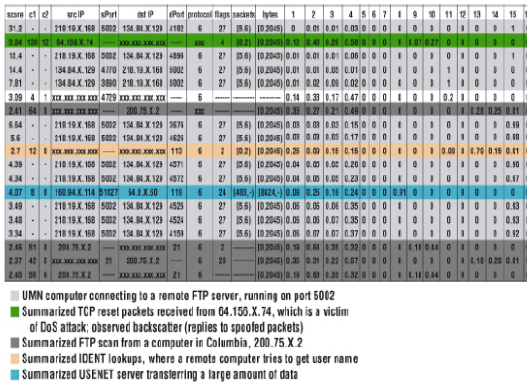

Figure 2. MINDS output of the summarization module

Summarizing Anomalous Connections in MINDS Module, Association Rules

In common, an association rule is an inference expression of the form X ⇒ Y, where X and Y are sets of binary features. To guess the happening of certain features in a record given the incidence of other features an association rule can be used. For instance, the rule {Bread, Butter}⇒{Milk} point outs that the majority of the transactions that enclose bread and butter also entail the purchase of milk. The sets of objects or binary features are famous as item sets in association rule terminology. Support calculates the small part of transactions that follow the rule while confidence is an estimate of the conditional probability P(Y|X). Sets or association rules are used for analyzing network traffic data association patterns.

Mining related patterns in network traffic data

It is good task due to the following reasons:

- *Unwarranted class distribution.*
- *Binarization and assembling of attribute values.*
- *Pruning the unneeded patterns.*
- *Finding discriminating patterns.*
- *Grouping the discovered patterns.*

Figure3: shows the general architecture of association analysis module.

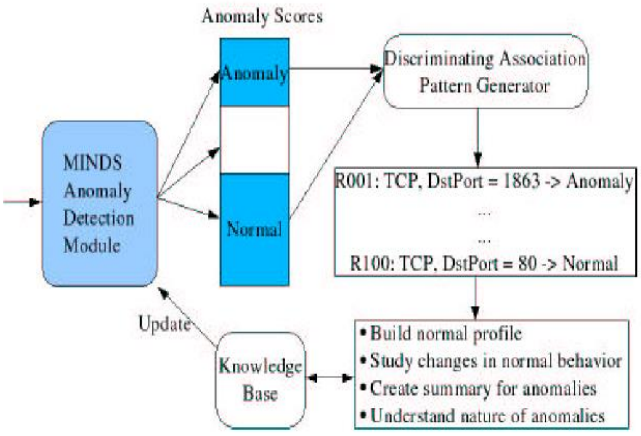

Figure 3. Overall architecture [3]

Assessment of attack summaries on real network data

In this part, we report some of the uppermost ranked (mainly discriminative) patterns produced by the our association pattern analysis module. These patterns placed for an outline of the majority regularly happening and discriminating anomalous traffic standard by MINDS anomaly detection module.

Profiling

We can apply clustering, a data mining technique for assembling related items, to discover related network connections and thus determine leading modes of activities. When data is high-dimensional and noisy (for example, network data), MINDS uses the Shared Nearest Neighbor clustering algorithm. SNN is extremely computationally demanding of the order $O(n2)$, where n is the number of network connections. Thus, we require to use similar computing to scale this algorithm to huge data sets. Our group has refined a parallel establishment of the SNN clustering algorithm for performance modeling, creating it possible to analyze very big amounts of network data.

The majority large clusters be similar to normal behavior modes, such as implicit confidential network traffic. Though, quite a lot of smaller clusters be identical to minor deviant behavior modes connecting to not design computers, insider attack, and strategy disobediences untraceable by other methods. Such clusters offer analysts knowledge they can act on instantaneously and can aid them find out their network traffic behavior.

Encountering distributed attacks

An intrusion detection system (IDS) which is running at one site do not have adequate data by itself to detect the attack. Immediately detecting such distributed cyber attacks depend upon an interconnected system of IDSs that can absorb network traffic data in adjacent real-time, detect anomalous connections, interact their conclusion to other IDSs, and assimilate the information from other systems to enlarge the anomaly scores of such threats. Such a system expressed of several autonomous IDSs that share their knowledge bases with each other to quickly detect wicked, large-scale cyber attacks.

Figure 4 emphasizes the distributed aspect of this problem. It displays the two-dimensional global Internet Protocol space such that every IP address designated in the world is expressed in some block. The black region shows unallocated IP space.

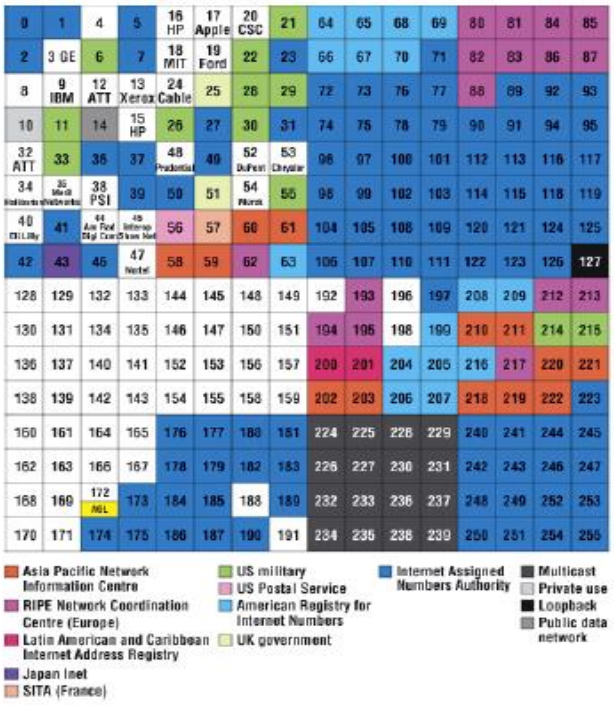

Figure 4. Map of the global IP Space

Figure 5 illustrates a graphical demonstration of suspicious connections activated from the outside (box on the right)to machines inside the University of Minnesota's IP space (box on the left) in a typical time window of 10 minutes. Each red dot in the right-hand box shows a doubtful connection made by a machine to an internal machine on port 80. The box which is on right hand pointed out that most of these potential attackers are clustered in specific Internet address blocks. A close checkup shows that most of the dense areas exist to the network blocks of cable and AOL users based in the US or to blocks allocated to Asia and Latin America. It's hard to tag a source as malicious on the basis of just one link. If multiple sites working the equivalent analysis beyond the IP space report the equivalent external source as suspicious, it would assemble the classification much more definite.

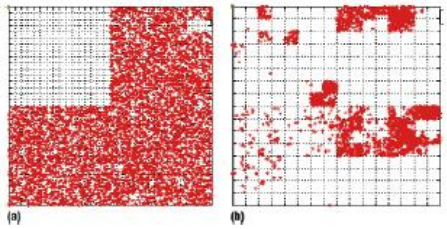

Figure 5. Doubtful traffic on port 80. (a) Destination IP addresses of doubtful connections within the University of Minnesota. (b) Source IPs of doubtful connections in the global IP space.

IV. PROPOSED SYSTEM

The excellent scheme for the future would be that we bring the data collected at these various sites to one place and then analyze it. But this isn't appropriate because the data is commonly distributed and more appropriate for distributed analysis, the amount of combining huge amounts of data and executing analysis at one site is very large and confidentiality, safeness, and certainty issues derive in distributing network data in the midst of different organizations. The necessity for a circulated framework is that in which these various sites can separately analyze their data and then share remarkable patterns and conclusions while appreciating the individual sites' data privacy. Enabling such a system would require directing distributed data, forwarding privacy issues, and using data mining tools, and would be much smooth if a middleware contributed these actions.

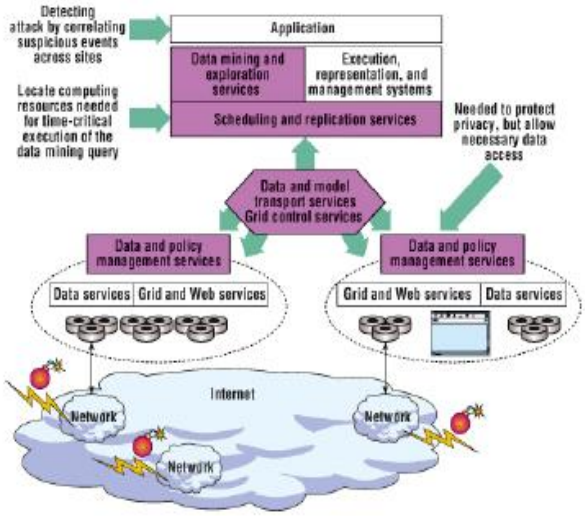

Figure 6. The distributed network intrusion detection system

V. CONCLUSION

The existing system of Minnesota Intrusion Detection System (MINDS), which apply a set of data mining based algorithms to deal with diverse features of cyber security is a very useful one. The *scan detector* intends at detecting scans which are the precursors to any network aggression. The algorithm of *anomaly detection* is very effective in detecting behavioral anomalies in the network traffic which typically translate to malicious activities such as denial-of-service (DoS) traffic, bounders and method violations. The module *profiling* helps a network analyst to understand the characteristics of the network traffic and detect any deviations from the normal profile. The intrusions noticed by MINDS are opposite to those of usual signature based systems, such as SNORT, which suggests that they both can be jointed to enlarge total attack coverage.

Due to the input data to the system, the privacy and security can be broken down. So we derived too many mechanisms to detect it. Since the hackers are finding out threats which can overcome the existing detecting mechanisms, we should develop a solution to check the input data which is entering into the system either by internal or external input mechanisms. A perfect mechanism is needed to detect the unusual input data and which prevents the suspected data to enter into the system. Data mining techniques can be used for deriving a solution to it.

VI. FUTURE WORKS

According to our first search, SNORT – a signature-based system is a complementary to the intrusions detected by MINDS. To increase overall attack coverage the two can be combined. MINDS is based on the analysis of unusual behavior, which is the key anomaly detection approach. This is suitable for detecting many types of threats. Figure shows three such types.

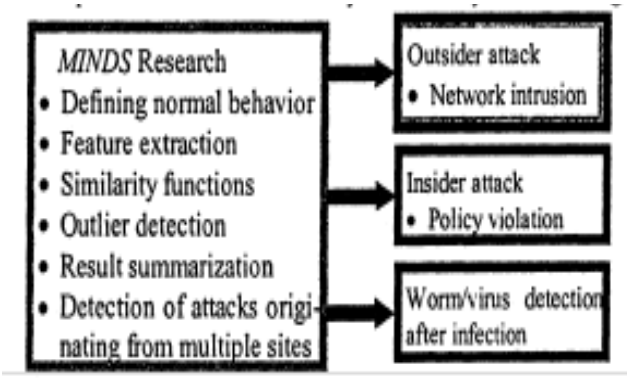

Figure 7. Three types of threats that can be detected by MINDS anomaly detection module [14]

On-line and scalable data mining algorithms and tools for detecting attacks and threats against computer systems are developed in MIND. But we need to be improved due to various challenges.

- For handling very large network traffic data sets there is a need for high performance data mining algorithms.
- Algorithms for mining data streams are necessary for building real-time intrusion detection system.
- Modification of standard data mining algorithms for low frequency of computer attacks is needed.
- In order to detect distributed attacks, analyze network data from several network locations

It is a complex task to convert network traffic data into useful features. We plan to use it in data mining algorithms. Security analysts can figure out the anomalous events and patterns extracted with the help of the graphical user interface tool of MINDS.

A number of applications early detection of unusual medical conditions - e.g. cardiac arrhythmia, detecting credit card and insurance blackmailers, new signs of hidden disasters in industrial process control, etc are outside of intrusion detection which have similar characteristics. We plan to use the data mining with cyber security techniques to such problems.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my guide Ms. Reesha P.U. and to Ms. Nisha Peter of Department of Computer Science, St. Joseph's College, Irinjalakuda.

REFERENCES

[1] "What is Data Mining, and How is it Useful for Power Plant Optimization? (and How is it Different from DOE, CFD, Statistical Modeling)" by StatSoft White Paper, July 2007 available at http://www.statsoft.com /Portals/0 /Support/Download/White-Papers/ What Is Data Mining. pdf

[2] "Minds: Architecture & Design" by Varun Chandola, Eric Eilertson, Levent Ert • oz, Gy • orgy Simon and Vipin Kumaravailable at http://www-users.cs.umn.edu /~kumar/ papers/ minds-chapter.pdf

[3] "MINDS - Minnesota Intrusion Detection System" by by Levent Ert¨oz, Eric Eilertson, Aleksandar Lazarevic, Pang-Ning Tan, Vipin Kumar, Jaideep Srivastava, Paul available at www-users.cs.umn.edu/ ~kumar/MINDS/papers/minds_chapter.pdf

[4] "Data mining with big data" by Xindong Wu , Sch. of Comput. Sci. & Inf. Eng., Hefei Univ. of Technol., Hefei, China Xingquan Zhu , Dept. of Comput. & Electr. Eng. & Comput. Sci., Florida Atlantic Univ., Boca Raton, FL, USA, Gong-Qing Wu , Sch. of Comput. Sci. & Inf. Eng., Hefei Univ. of Technol., Hefei, China, Wei Ding , Comput. Sci. Dept., Univ. of Massachusetts Boston, Boston, MA, USA DOI Bookmark available at http://doi.ieeecomputersociety.org/10.1109/TKDE.2013.109

[5] "Cyber Security: Threats, Reasons, Challenges, Methodologies and State of the Art Solutions for Industrial Applications" by Abdul Razzaq, Ali Hur, H Farooq Ahmad, Muddassar Masood School of Electrical Engineering and Computer Science (SEECS) National University of Sciences and Technology, Islamabad, Pakistan available at http://www.computer.org/csdl/proceedings/isads/2013/5069/00/06513420.pdf

[6] "Data Mining and Cyber Security" By Nida Tahir available at http://nationalsecurityzone.org/site/data-mining-and-cyber-security/

[7] "Enabling Multilevel Trust in Privacy Preserving Data Mining" by Yaping Li , The Chinese University of Hong Kong, Hong Kong Minghua Chen , The Chinese University of Hong Kong, Hong Kong Qiwei Li , Rice University, Houston Wei Zhang , The Chinese University of Hong Kong, Hong Kong DOI Bookmark available at http:// doi.ieeecomputersociety.org/ 10.1109/TKDE.2011.124

[8] "Security Related Data Mining" by Mehrnoosh Monshizadeh, Zheng Yan, DOI Bookmark: available at http://doi.ieeecomputersociety.org/10.1109/CIT.2014.130

[9] "Data Mining for Security Applications" by Bhavani Thuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen DOI Bookmark: available at http://doi.ieeecomputersociety.org/10.1109/EUC.2008.62

[10] "Prevention in Data Mining" by Sara Hajian , Universitat Rovira i Virgili, Tarragona, Josep Domingo-Ferrer , Universitat Rovira i Virgili, Tarragona DOI Bookmark: available at http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.72

[11] "Compressive sensing based secure multiparty privacy preserving framework for collaborative data-mining and signal processing" by Jun Tian , Futurewei Technologies, Bridgewater, NJ 08807, USA, DOI Bookmark: available at http://doi.ieeecomputersociety.org/10.1109/ICME.2014.6890141

[12] "A Privacy Preserving Repository for Data Integration across Data Sharing Services", by Stephen S. Yau , Arizona State University, Tempe, Yin Yin , Arizona State University, Tempe, DOI Bookmark: available at http:// doi.ieeecomputersociety.org /10.1109 /TSC.2008.14

[13] "An Adaptive Privacy Preserving Data Mining Model under Distributed Environment" by Feng Li , Jin Ma , Jian-hua Li, DOI Bookmark: available at http:// doi.ieeecomputersociety.org/10.1109/SITIS.2007.139

[14] "Detection and Summarization of Novel Network attacks using Data Mining" by L. Ertoz, E. Elierston, A. Lazarevic, P. Ning-tan, P. Dokas, V. Kumar and J. Srivastava available at http://static.msi.umn.edu/rreports/2003/212.pdf

[15] "Data Mining for Cyber Security" by Varun Chandola, Eric Eilertson, Levent ErtÄoz, GyÄorgy Simon and Vipin Kumar available at http://minds.cs.umn.edu/ publications/ chapter.pdf