Exploiting the Data Mining Methodology for Cyber Security

Hisham S. Katoua
Management Information Systems Dept.
Faculty of Economics & Administration
King Abdulaziz University ,Jeddah, Kingdom of Saudi Arabia
dr_hisham_sa@hotmail.com

Abstract

Data mining methodology (DMM) aims to extract useful information and discover some hidden patterns from huge amount of databases, which statistical approaches cannot discover. It is a multidisciplinary field of research includes: machine learning, databases, statistics, expert systems, visualization, high performance computing, rough sets, neural networks, and knowledge representation, etc. Data mining is supported by a host that captures the character of data in several different ways (e.g. clustering, classification, link analysis, summarization, regression models, and sequence analysis).

Researchers are investigating the use of DMM in both national security (e.g. detecting the bad guys) and in cyber security (e.g. intrusion detection and auditing). This paper presents the application of data mining tasks and techniques in cyber security problems. The paper discusses the following topics; (a) profiling networks traffic using clustering,(b) tracing viruses to the perpetrators using link analysis ,(c) detecting unusual behaviors and patterns using anomaly detection techniques,(d) grouping various cyber attacks and then use the profiles to detect an attack when it occurs using classification, and(e) determining future attacks using prediction models.

Keywords: *Data mining, Cyber security, National Security, Machine learning*

1. Introduction

Data mining (DM) deals with the discovery of *hidden knowledge, unexpected patterns* and new rules from *large databases*. It is currently regarded as the key element of a much more elaborate process called *knowledge discovery in databases* (KDD), which is closely linked to another important development - *data warehouse*. A data warehouse is a central store of data that has been extracted from operational data. The information in a data warehouse is *subject-oriented, non-volatile*, and of an *historic nature*, so data warehouses tend to contain extremely large data sets. The combination of *data warehousing*, *decision support* and *data mining* indicates an innovative and totally new approach to information management. Until now, information systems have been built and operated mainly to support the operational processes of an organization. KDD and data warehousing view the information in an organization in an entirely new way- as a strategic source of opportunity.

There is confusion about the exact meaning of the terms 'data mining and "KDD" with many authors regarding them as synonymous. At the first international "KDD" conference in Montreal in 1995, it was proposed that the term "KDD" be employed to describe the whole process of extraction of knowledge from data. Knowledge means relationships and patterns between data elements. It was further proposed that the term "data mining should be used exclusively for the discovery stage of the KDD process. A more or less official definition of KDD is: 'the non-trivial extraction of implicit, previously unknown and potentially useful Knowledge from data'. So the knowledge must be new, not obvious, and one must be able to use it. KDD is not a new technique but rather a multi-disciplinary field of research including *machine learning, statistics, database technology, expert systems* and *data visualization*.

2. Knowledge Discovery Processes

Figure 1 shows the main functional phases of the knowledge discovery process. This is arranged into a stream of steps:

- understanding the domain in which the discovery will be carried out.
- forming the data set, its cleaning, and warehousing.
- extracting patterns, this is essence of DM.
- post-processing of the discovery knowledge
- putting the results of knowledge discovery into use.

Fundamental issues in knowledge discovery arise from the very nature of databases and the objects (data) they deal with. They are characterized as follows: (a) huge amounts of data, (b) dynamic nature of data, (c) incomplete or imprecise data, (d) noisy data,(e) missing attribute values, and (f) redundant or insignificant data.

The knowledge discovery process is dynamic, highly interactive, iterative, and fully visualizable. Its main goals are to:

- extract useful reports
- spot interesting events and trends
- support decision-making processes
- exploit the data to achieve scientific, business, or operational goals.

In spite of the diversity of the application areas, there are several common characteristic features:

- availability of massive sets of data
- high underutilization of data.
- access domain experts fully familiar with the area becomes crucial during the development of the knowledge discovery system.
- lack of expertise of the end users (thus knowledge discovery is a greatly welcomed activity).

DM exhibits a plethora of algorithmic tools such as statistics, regression models, neural networks, fuzzy sets and evolutionary model. DM is supported by a host that captures the character of data in several different ways (see table 1)

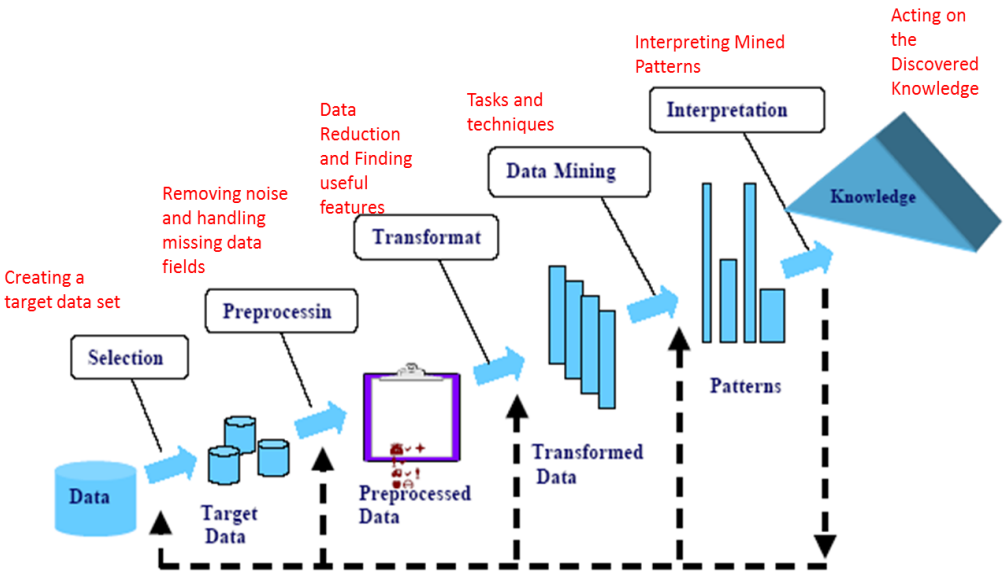

Fig.1: Phases of knowledge discovery process.

Table 1: Data Mining Tasks and Techniques

Data Mining Task	The Appropriate Data Mining Technique
Classification	Neural Networks Support Vector Machine Decision Trees Genetic Algorithms Rule induction
Clustering	K-Means
Regression and prediction	Support Vector Machine Decision Trees Rule induction, NN
Association and Link Analysis (finding correlation between items in a dataset)	Association Rule Mining
Summarization	Multivariate Visualization

4. Data Mining Uses and Applications

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. This section presents a brief account of the well known DM uses and applications (see table 2).

Table 2: Data Mining Applications in Private Sector

Application domain	Task
Insurance and banking industries	detect fraud and assist in risk assessment (e.g., credit scoring)
Medical community	predict the effectiveness of a procedure or medicine
Pharmaceutical firms	use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases.
Retailers	assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together.
Telephone service providers	to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor

DM approaches have grown also to be used for purposes such as measuring and improving program performance. It has been reported that data mining has been able to assess crime patterns and adjust resource allotments. In addition, data mining can be used to predict demographic changes in the constituency it serves so that it can better estimate its budgetary needs. Data mining can be used also to review plane crash data to recognize common defects and recommend precautionary measures.

Recently, data mining has been increasingly cited as an important tool for homeland security efforts. Some observers suggest that data mining should be used as a means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records.

Based on the analysis of published papers during the five years, figure 2 shows the emerging data mining research areas

• **Graph mining** • **Data mining in bioinformatics** • **Privacy-aware data mining** • **Large scale data mining** • **Temporal pattern mining** • **Stream data mining** **Mining moving object data, RFID data, and data from sensor networks** • **Ubiquitous knowledge discovery**	• **Mining multi-agent data** • **Mining and link analysis in networked settings: web, social and computer networks, and online communities** • **Mining the semantic web** • **Data mining in electronic commerce** • **Data mining in e-Learning** •**Web search, advertising, and marketing task**

Fig. 2 Emerging applications of data mining techniques

5. Cyber Security, Cyber Warfare and Digital Forensic

DM is mainly applied to cyber security also known as information security problems. This includes problems such as Intrusion detection and auditing

- Anomaly detection techniques to detect unusual pattern and behavior
- Link Analysis maybe used to trace the viruses to the perpetrators.
- Classification is used to group various cyber attacks and then use the profiles to detect an attack when it occurs
- Prediction maybe used to determine potential future attacks in a way on information learned about terrorists though email and phone conversations
- Analyzing the audit data, one could build a repository or data warehouse containing the audit data and the conduct an analysis using various data mining tools to see if there are potential anomalies
- Insider threat analysis is also a problem both for national security as well from cyber security perspective. That is those working in a corporation who are considered to be trusted could commit espionage. Similarly those with proper access to the computer system could plant Trojan horses and viruses, catching such terrorists is far more difficult than catching terrorists outside of an organization. One may need to monitor the access patterns of all the individuals of a corporation even if they are system administrators to see whether they are carrying out cyber terrorism activities.

5.1. Cyber Security

Cyber security research areas may be related to any security field such as:

- Privacy issues
- Formal Methods Application in Security
- Incident Handling and Penetration Testing
- Operating Systems and Database Security
- Security in Cloud Computing
- Security in Social Networks
- Multimedia and Document Security
- Hardware-Based security
- VOIP, Wireless and Telecommunications Network Security
- Security of Web-based Applications and Services
- Enterprise Systems Security
- SCADA and Embedded systems security
- Distributed and Pervasive Systems Security
- Secure Software Development, Architecture and Outsourcing
- Security for Future Networks
- Security protocols
- Legal Issues

5.2. Digital Forensic

Digital forensic application areas may be related to the following fields:

- Data leakage, Data protection and Database forensics
- Forensics of Virtual and Cloud Environments
- Network Forensics and Traffic Analysis Hardware Vulnerabilities and
- Device Forensics
- Information Hiding
- File System and Memory Analysis Multimedia Forensic
- Executable Content and Content Filtering
- Anti-Forensics and Anti-Anti-Forensics Techniques
- Malware forensics and Anti-Malware techniques
- Evidentiary Aspects of Digital Forensics
- Investigation of Insider Attacks
- Cyber-Crimes
- Large-Scale Investigations
- New threats and Non-Traditional approaches

5.3. Information Assurance and Security Management

Information assurance and security management application areas include the following topics:

- Corporate Governance
- Laws and Regulations
- Threats, Vulnerabilities, and Risk Management
- Business Continuity & Disaster Recovery Planning
- Critical Infrastructure Protection
- Digital Rights Management and Intellectual Property Protection
- Security Policies and Trust Management
- Identity Management
- Decidability and Complexity
- Economics of Security
- Fraud Management

5.4. Cyber warfare and Physical Security

Cyber warfare and physical security application areas related to the following fields:

- Surveillance Systems
- Cyber Warfare Trends and Approaches
- Social engineering
- Authentication and Access Control Systems
- Biometrics Applications
- Electronic Passports, National ID and Smart Card Security
- Template Protection and Liveliness detection
- Biometrics standards and standardization
- New theories and algorithms in biometrics

6. Conclusions

Data mining approach is a very promising methodology towards both national and cyber securities. Data mining techniques and algorithms can be used as a robust intelligent tool by network analysts to defend the network against attacks and emerging cyber threats. The data mining based intelligent systems can detect different types of attacks and intrusions on a computer network. The architecture of such systems and its algorithms are very effective in performing the following tasks; (a) detecting scans which are the precursors to any network attack, (b) detecting behavioral anomalies in the network traffic which typically translate to malicious activities (such as dos traffic, worms, policy violations and inside abuse), and (c) to understand the characteristics of the network traffic and detect any deviations from the normal profile. While data mining can be used to detect and prevent cyber attacks, data mining can cause some security problems such as the inference and privacy problems. With data mining techniques one could infer sensitive associations from the legitimate responses.

References

[1] Abdel-Badeeh M.Salem, and safia A. Mahmoud., "Mining patient Data Based on Rough Set Theory to Determine Thrombosis Disease", Proceedings of First Intelligence conference on Intelligent Computing and Information Systems, pp 291-296. ICICIS 2002, Cairo, Egypt, June 24-26,2002.

[2] Abdel-Badeeh M.Salem and Abeer M.Mahmoud, "A Hybrid Genetic Algorithm-Decision Tree Classifier", Proceedings of the 3rd International Conference on New Trends in Intelligent Information Processing and Web Mining, Zakopane, Poland, pp. 221-232, June 2-5, 2003.

[3] Agarwal, R., & Srikant, R. Mining sequential patterns. In Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan (pp. 3–14), 2005.

[4] Aleksandar Lazarevic, Levent ErtÄoz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network in-trusion detection. In *SIAM Conference on Data Mining (SDM)*, 2003.

[5] Anoop Singhal and Sushil Jajodia. Data mining for intrusion detection. In *Data Mining and Knowledge Discovery Handbook*, pages 1225{1237. Springer, 2005.

[6] Ansari, S., Kohavi, R., Mason, L., and Zheng, Z., "Integrating E-Commerce and Data Mining: Architecture and Challenges" .Proceedings of IEEE International Conference on Data Mining, 2001.

[7] B. Thuraisingham. Data mining, national security, privacy and civil liberties. SIGKDD Explorations, January 2003.

[8] B. Thuraisingham. Managing threats to web databases and cyber systems: Issues, solutions and challenges. In V. Kumar et al, editor, Cyber Security: Threats and Countermeasures. Kluwer.

[9] Cai, W. and Li L., "Anomaly Detection using TCP Header Information, STAT753 Class Project Paper, May 2004.".
Web Site:http://www.scs.gmu.edu/~wcai/stat753/stat753report.pdf.

[10] Cadez, D. Heckerman, and C. Meek. Visualization of navigation patterns on web site using model based clustering. In ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD'00), PP 280–284, Boston, USA, August 2000.

[11] Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J. J., Wang, G., Zheng, R., Atabakhsh, H., Crime Data Mining: An Overview and Case Studies", A project under NSF Digital Government Programme, USA, "COPLINK Center: Information and Knowledge Management for Law Enforcement,", July 2000 -June 2003.

[12] Chen, H., Chung, W., Xu Jennifer, J., Wang, G., Qin, Y., Chau, M., "Crime Data Mining: A General Framework and Some Examples". Technical Report, Published by the IEEE Computer Society, 0018-9162/04, pp 50-56, April 2004.

[13] Cios K. J., Pedrycz, W. and Swiniarski, R. W. Data Mining Methods for Knowledge Discovery. Kluwer 1998.

[14] Cohen, J. J., Olivia, C., Rud, P., "Data Mining of Market Knowledge in The Pharmaceutical Industry". Proceeding of 13th Annual Conference of North-East SAS Users Group Inc., NESUG2000, Philadelphia Pennsylvania, September 24-26 2000.

[15] Daniel Barbara and Sushil Jajodia, editors. *Applications of Data Mining in Computer Security*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[16] Deng, B., Liu, X., "Data Mining in Quality Improvement". USA.ISBN1-59047-061-3.WebSite http://www2.sas.com/proceedings/sugi27/Proceed27.pdf.

[17] Duda, R. O., Hart, P. E., & Stork, D. G. Pattern classification. Wiley Interscience 2000.

[18] Elovici, Y., Kandel, A., Last, M., Shapira, B., Zaafrany, O., "Using Data Mining Techniques for Detecting Terror-Related Activities on the Web".WebSite:
www.ise.bgu.ac.il/faculty/mlast/papers/JIW_Paper.pdf
-12-

[19] Eric Eilertson, Levent ErtÄoz, Vipin Kumar, and Kerry Long. Minds { a new approach to the information security process. In 24th *Army Science Conference*. US Army, 2004.

[20] Feldman, R., & Sanger, J. The text mining handbook. Cambridge University Press 2006.

[21] Gyorgy Simon, Hui Xiong, Eric Eilertson, and Vipin Kumar. Scan detection: A data mining approach. Technical Report AHPCRC 038, University of Minnesota, Twin Cities, 2005.

[22] Gyorgy Simon, Hui Xiong, Eric Eilertson, and Vipin Kumar. Scan detection: A data mining approach. In *Proceedings of SIAM Conference on Data Mining (SDM)*, 2006.

[23] H. Chen et al. In Proceedings of the 1st Conference on Security Informatics, Tucson, AZ, June 2003.

[24] I.H. Witten and E. Frank, Data Mining – Practical Machine Learning Tools and Techniques. 2nd ed Elsevier, 2005.

[25] Jadhav, S. R., and Kumbargoudar, P., "Multimedia Data Mining in Digital Libraries: Standards and Features READIT, pp 54-59, 2007.

[26] Jain, A. K., Murty, M. N., and Flynn, P. J., Data clustering: A review. ACM Computing Surveys, 31(3), 264–323, 1999.

[27] Kirkos, E., Spathis, C., and Manolopoulos., Y., "Data Mining techniques for the detection of fraudulent financial statements." Expert Systems with Applications 32(4), 995-1003, 2007.

[28] Levent ErtÄoz, Eric Eilertson, Aleksander Lazarevic, Pang-Ning Tan, Vipin Ku-mar, Jaideep Srivastava, and Paul Dokas. MINDS - Minnesota Intrusion Detection System. In *Data Mining - Next Generation Challenges and Future Directions*. MIT Press, 2004.

[29] Romero, C., & Ventura, S. Data mining in e-learning. Southampton, UK: Wit Press 2006.

[30] Schultz, M. G., Eskin, Eleazar, Zadok, Erez, and Stolfo, Salvatore, J., "Data Mining Methods for Detection of New Malicious Executables". Proceedings of the 2001 IEEE Symposium on Security And Privacy, IEEE Computer Society Washington, DC, USA , ISSN:1081-6011, 2001.

[31] Smith, L., Lipscomb, B., and Simkins, A., "Data Mining in Sports: Predicting Cy Young Award Winners". Journal of Computer Science, Vol. 22, Page No. 115-121,April 2007.

[32] Spence, R. Information visualization. Addison-Wesley 2001.

-13-

[33] Vipin Kumar, Jaideep Srivastava, and Aleksander Lazarevic, editors. *Managing Cyber Threats{Issues, Approaches and Challenges*. Springer Verlag, May 2005.

[34] Varun Chandola, Eric Eilertson, Levent ErtÄoz, GyÄorgy Simon and Vipin Kumar, Data Mining for Cyber Security, Book chapter in *Data Warehousing and Data Mining Techniques for Computer Security*, Springer, 2006.