Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-6, 2016
ISSN: 2454-1362, http://www.onlinejournal.in

A Review on Security Measures in Data Mining

Soumen Bhowmik[1], Rabisankar Chattopadhyay[2] &
Uddalok Chatterjee[3]

[1,3]Assistant Professor, Department of Computer Science & Engineering, Bengal Institute of Technology & Management, Santiniktan

[2]M. Tech Student, Department of Computer Science & Engineering, Bengal Institute of Technology & Management, Santiniktan.

Abstract*: In our computer science world the biggest challenge is to secure our information i.e. how to protect our information from outer threats. Information is the meaningful combination of data. To store the data we need a huge database or a data mine. Data mining is a process to search the data from a huge data base for different work. As data are available in the different formats so not only to analyze these data but also take a good decision and maintain the data. In this paper we have discuss on various technique, approaches and security measures for the data mining.*

Index Terms: *Data exploration, Knowledge Discovery in Databases (KDD), Rule Indication, k-nearest neighbor*

1. Introduction

Now a days the computer system store huge amounts of data. Data mining is special technical term for extraction of hidden predictive information of new and interesting pattern of data from large data sets. In International Federation for Information Processing (IFIP) conference motive was that how to secure the data mining process and how to protect the future databases and legality of those databases. So, that large organization can securely extract their secret information from their large data warehouse and that help them to focus in achieving their goal [1].

The KDD Process:

Data mining, known as Knowledge Discovery in Databases (KDD) [2-4], it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. To obtain useful knowledge from data, the following steps are performed in an iterative way-

Problem Definition:
A data-mining project starts with the understanding of the business problem then translated into a data mining problem definition.

Data Exploration:
Domain experts understand the meaning of the metadata, describe and explore the data and identify quality problems of the data.

Data Preparation:
Domain experts collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. Here meaning of the data is unchanged.

Modeling:
Data mining experts select and apply various intelligent methods to extract data patterns.

Evaluation:
Data mining experts evaluate the model. If the model does not satisfy their required, they go back to the modeling phase and rebuild the model by changing its parameters until ultimate values are gain.

Deployment:
Data mining experts use the mining results by exporting the results into database tables or into other applications i.e. known as knowledge.

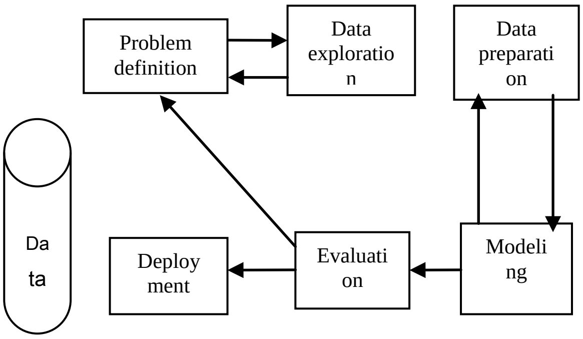

Figure 1: An overview of KDD process

2. Data Mining Architecture

Data mining is a technique to dig the data (Interesting knowledge) from the large databases. This knowledge useful for business strategies, medical research etc. The architecture contains modules for efficient data analysis for generating global mining model [5].

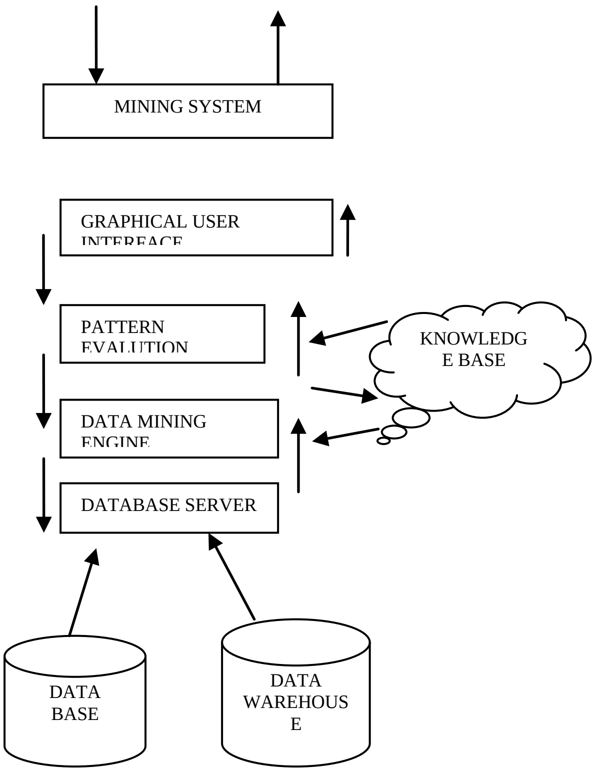

Figure 2: Block Diagram of DMA

3. Security Concern in Data Mining

For every government and private organization database are very sensitive and important component. Data Mining is associated with data-ware house and database. As Knowledge extract from the Data Warehouse is very confidential and monumental for the organization so it should be protected from the outside threats .The necessity of data mining security concern with the following characteristics-

Physical Database Integrity:

Due to some problems occur like system crash, power failure data may be lost .We cannot retrieve the information. This physical database integrity concerned with the data is read from and write to the disk. For this reason Data mining becomes unable to predict pattern by given applications.

Logical Database Integrity:

When there is a need of updation on database this type of integrity indicates that for any modification of data no other data will affect. If it occurs, no data-mining algorithm cannot be able to predict correct information.

Auditability:

The modification of records and fields of the database are taken the system responds immediately to user requests (i.e. OLTP).This ensures that modification implemented under the data warehouse are error free.

User Authentication:

Authorization primarily includes two processes:
1) Permitting only certain users to access, process, or alter data.
2) Applying varying limitations on users' access or actions. The limitations placed on (or removed from) users can apply to objects, such as schemas, tables, or rows; or to resources, such as time (CPU, connect, or idle times)

Human related error:

Database Management system requires expert engineer to overcome the human related error and mishandling over the data.

Password Lifetime and Expiration:

The database administrator can specify for a valid user a lifetime for passwords, after which they expire and must be changed before account login is again permitted. It is very powerful scheme for security of data mining.

4. Security Measure and Performance

Data mining is one of the most popular combinations of many tools for data abstraction and getting meaningful items. To keep extract information confidential and secure from outside threats we use security measures. These type of security measures are based on the characteristic of data Mining [6]

Privacy:

Every organization should maintain the privacy of the data so that no one can gain benefits. To maintain this organization should have to train their employees time to time to aware about the privacy of data.

Sensitivity:

The data warehouse keep the whole information about the organization, in all those data their exist some access control for sensitive and general information .Not every user can access the whole database.

Consistent and error free Data:

It must ensure that data entered into the database is correct and consistent otherwise data mining tools will produce incorrect data which vital issue for any organization.

Domain Integrity:

If data numeric field is in mode of character then it produces the incorrect result of mathematical operations during data mining. Once a integrity constraint is enforced on data items then user should not have to right about removal of that integrity constraint.

Data Integrity:

To remove redundant data integrity of data is important. Integrity is applied by applying various constraints on the database which helps to link the tables of databases and also to improve security in the database. Once a integrity constraint is enforced on data items then user should not have to right about removal of that integrity constraint.

Elimination of duplicate data:

There should be a mechanism that finds the duplicate and incorrect data to be corrected before the storing into the large databases. The correction should be automated not manual.

Elimination of False Matches:

In the process of data mining the extraction of information from databases may produce wrong matching output. This false information matching is eliminated by automated filtering so that accidentally any confidential is not leaked. If manual system is applied then proper security aspects of leakage of information should be defined on behalf of the company policies.

5. Popular Techniques of data Mining

Artificial Neural Networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure. [19]

Decision Trees: Tree-shaped structures that represented set of decisions. These decisions generate rules for the classification of a dataset under the large databases. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution. [18]

k-nearest Neighbor Technique: This method classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). [20]

Rule Induction: The extraction of useful if-then rules from data based on statistical significance between different records of database.

Challenges:

- Threats imposed by data mining techniques to privacy/security and possible remedies.
- Statistical approaches to ensure privacy in data mining
- New methodologies for privacy preserving data mining
- Data quality, privacy, and security measures

6. Conclusion

How to protect sensitive information from the security threats brought by data mining has become a hot topic in recent years. In this paper we review the privacy issues related to data mining. To achieve the privacy-preserving goals of different research are required. We hope that the review presented in this paper can offer researchers different discernment into the issue of privacy-preserving data mining and help for new solutions to the security of sensitive information.

7. References

[1] Abdelsalam, H. et al., 2001. Drishti: An Integrated Navigation System for Visually Impaired and Disabled. IEEE Fifth International Symposium on Wearable Computers. P 149.

[2] *Introduction to Data Mining and Knowledge Discovery*, Third Edition ISBN: 1-892095-02- 5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[3] Dunham, M. H., Sridhar S., "*Data Mining: Introductory and Advanced Topics*", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st
Edition, 2006

[4] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "*From Data Mining to Knowledge Discovery in Databases*," AI Magazine, American Association for Artificial Intelligence, 1996.

[5] Mafruz Zaman Ashrafi, David Taniar, Kate A. Smith, ''*Data Mining Architecture for Clustered Environments*'' , *Proceeding PARA '02 Proceedings of the 6th International Conference on Applied Parallel Computing Advanced Scientific Computing*, Pages 89-98, Springer- Verlag London, UK ©2002.

[6]Chris Clifton. Using sample size to limit exposure to data mining. Journal of Computer Security, 8(4):281{307, November 2000. URL http://iospress.metapress.com/openurl.asp?genre=artcle&issn=0926227X&volume=8&issue=4&spage=281.

[7]Data Mining Techniques: http://www.ibm.com/developerworks/library/ba-data-mining-techniques.

[8]Wikipedia "http://en.wikipedia.org/wiki/Wearable_technology"

[9] IBM (1995) Data Mining - An IBM Overview, IBM AlmadenResearchCentre. URL:http://www.almaden.ibm.com/stss/papers/overview.html

[10] Morgenstern, M., "Security and Inference in Multilevel Database and Knowledge Base Systems," Proceedings of the ACM SIGMOD Conference, San Francisco, CA, June 1987.

[11] S. A. Demurjian and J. E. Dobson, "Database Security IX Status and Prospects Edited by D. L. Spooner ISBN 0 412 72920 2, 1996, pp. 391-399.

[12] Lin, T. Y., "Anamoly Detection -- A Soft Computing Approach", Proceedings in the ACM SIGSAC New Security Paradigm Workshop,Aug 3-5, 1994,44-53.,1994

[13] Scott W. Ambler, "Challenges with legacy data: Knowing your data enemy is the first step in overcoming it", Practice Leader, Agile Development, Rational Methods Group, IBM, 01 Jul 2001.

[14] Agrawal, R, and R. Srikant, "Privacy-preserving Data Mining," Proceedings of the ACM SIGMOD Conference, Dallas, TX, May 2000.

[15] Clifton, C., M. Kantarcioglu and J. Vaidya, "Defining Privacy for Data Mining," Purdue University, 2002 (see also Next Generation Data Mining Workshop, Baltimore, MD, November 2002.

[16] Evfimievski, A., R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, July 2002.

[17] Fung B., Wang K., Yu P. "Top-Down Specialization for Information and Privacy Preservation. ICDE Conference, 2005. [22] Wang K., Yu P., Chakraborty S., " Bottom-Up Generalization: A Data Mining Solution to Privacy Protection.", ICDM Conference, 2004

[18] Tan Jun-shan1, He Wei1, Qing Yan2 "Application of Genetic Algorithm in Data Mining". 2009 First International Workshop on Education Technology and Computer Science

[19] Ms. Aruna J. Chamatkar, Dr. P.K. Butey "Implementation of Different Data mining Algorithms with Neural Network" 2015 International Conference on Computing Communication Control and Automation.

[20] Wang Tongwen; Guan Lin. "A data mining technique based on pattern discovery and k-nearest neighbor classifier for transient stability assessment" 2007 International Power Engineering Conference (IPEC 2007)