Secure Computation in Privacy Preserving Data Mining

Mr. Rajdeep Brar
M. Tech. *(Student)
Department of Computer Science & Engineering
Marudhar Engineering College, Bikaner (Raj.), India.
Email: rajdeep33@gmail.com

Mrs. Monika Soni
Asst. Prof. Department of Computer Science & Engineering,
Marudhar Engineering College, Bikaner (Raj.)
Email: 12.monika@gmail.com

Abstract— Data mining is the process of analyzing data from different perspectives and summarizing it into useful information used to feedback, increase revenue, cuts costs, or all. A number of freeware and shareware data mining software resources are available for analyzing data. Data Mining allows users or organizations to analyze the extracted data from many different dimensions or angles, systematically categorize it, and summarize the relationships identified. Privacy preserving data mining means the "mining" of knowledge from distributed data without disrupt the privacy of the parties involved in contributing the data. Data mining causes the social and ethical problem by acknowledge the data requiring privacy. Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community and for the government statistical agencies. Hence, the security and privacy are the issues which become much more important area of research in data mining recently.

Keywords- *Data Mining, Privacy, Security, data.*

*****

I. INTRODUCTION

Here the term data mining implies mining the data. Mining means to extract. The verb usually refers to mining operations that extract from the Earth her hidden, precious resources. The combination of this word with data suggests an in-depth search to find multiple information which previously not noticed in the huge available data. From the viewpoint of scientific research, data mining is a relatively new discipline that has developed mainly from studies carried out in other disciplines such as computing, marketing, statistics, medical, business etc [1].

Data mining is a combination of Multiple Disciplines. Figure 1 below show the different disciplines that take part in data mining.

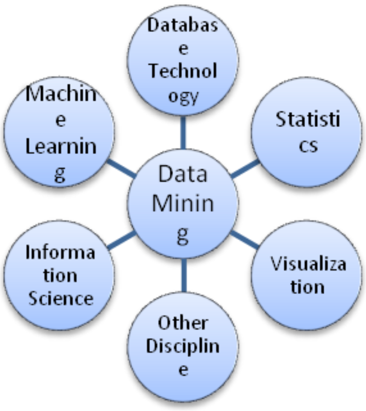

Figure 1 –Data Mining as Combination of Multiple Disciplines for data mining.

Data mining is one of the software tool from which the business analysts finding patterns and relationships in the data — it does not tell the value of the patterns to the organization. Furthermore, the patterns uncovered by data mining must be established in the real world [9].

II. KNOWLEDGE DISCOVERY IN DATABASES (KDD)

We can say that knowledge discovery is the process which identifying interesting new patterns in data. These patterns can be, e.g., relations, events or trends, and they can admit both precision and exceptions.

Due to the Internet the online data is growing very fast and the overall use of databases have created an endless need for KDD methodologies. KDD has emerge from interaction and cooperation among such different fields as machine learning, pattern recognition, database, statistics, artificial Intelligence, knowledge representation, and knowledge acquisition for intelligent systems. The main concept of knowledge discovery process is to discover a high level knowledge from lower levels of relatively raw data, or to discover a higher level of interpretation and abstraction than those previously known [2].

Knowledge Discovery and Data Mining both motivates the latest research—in statistics, databases, machine learning, and artificial intelligence —that are part of the exciting and fast growing field of Knowledge Discovery and Data Mining. KDD has evolved from interaction and cooperation among such different fields as machine learning, pattern recognition, database, statistics, artificial Intelligence, knowledge representation, and knowledge acquisition for intelligent systems [3].

III. Applications of Data Mining

1. Financial Data Analysis

The financial data in banking and financial industry is generally safe and of high quality which promote the systematic data analysis and data mining. Here are the few typical cases:

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

2. Retail Industry

Data Mining has much application in Retail Industry because it gather huge amount of data from, customer purchasing history, sales, consumption and goods transportation. It is natural that the quantity of data collected will continue to enlarge fast because of increasing ease, availability and popularity of web.

The Data Mining in Retail Industry helps in identifying customer buying patterns and trends. That starts to improved quality of customer service and good customer confinement and satisfaction. Here is some examples of data mining in retail industry:

- Design and Construction of data warehouses based on benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

3. Telecommunication Industry

Now in these days the Telecommunication industry is one of the most rising industries providing various services such as fax, pager, cellular phone, Internet messenger, images, e-mail, web data transmission etc. Due to the development of new computer and communication technologies, the telecommunication industry is speedily expanding.

Data Mining in Telecommunication industry helps in establish the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is some examples for which data mining improve telecommunication services:

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

4. Biological Data Analysis

We can see that there is huge growth in field of biology. For example genomics, proteomics, functional Genomics and biomedical research. Biological data mining is very important part of Bioinformatics. Following are some examples in which Data mining takes important part for biological data analysis:

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.

5. Other Scientific Applications

Large amount of data have been possessed from scientific domains such as geosciences, astronomy etc. There is huge amount of data sets being accomplished because of the fast numerical simulations in various fields such as climate, and ecosystem modeling, chemical engineering, fluid dynamics etc. Following are some applications of data mining in field of Scientific Applications:

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

6. Intrusion Detection

Intrusion means any kind of action that threatens integrity, confidentiality, or availability of network resources. In this world of connectivity security has become the major issue. With increased usage of internet and availability of tools and tricks for intruding and attacking network prompted intrusion detection to become a critical part of network administration. Here is some examples list in which data mining technology may be applied for intrusion detection:

- Development of data mining algorithm for intrusion detection.

- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools [10].

IV. Privacy Issues in Data Mining

Data mining is a powerful data analysis tool permissive discovery of useful patterns in several applications. The availability of large data warehouses and associated technologies acknowledge that the usefulness of data feel necessity for protection of every key attributes such as patient's customer preferences, condition information, personal background information etc. Therefore, how to do provide security without losing the privacy is the main issue in data mining. Privacy preserving data mining (PPDM) conception with this issue. Currently, there are many approaches of privacy preserving data mining to transform the original data. The methods of privacy preserving data mining are estimate based on applicability, privacy protection metric, the accuracy, computation, etc [4].

V. Privacy Preserving Data Mining

The problem of privacy preserving data mining has become more significant in present years because of the increasing ability to store personal data about users and the increasing composure of data mining algorithm to leverage this information. A number of techniques such as, k-anonymity, association, classification, clustering, rule mining have been suggested in recent years in order to work with privacy preserving data mining. Furthermore, the problem has been emulated in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. Data mining techniques have been refined successfully to extracts knowledge in order to support a variety of domains retailing, weather forecasting, medical diagnosis, and national security. But it is still a important task to mine certain kinds of data without resist the data owner's privacy .For example, how to mine patient's private data is an ongoing problem in health care applications. As data mining become more pervasive, privacy concerns are increasing [5].

A. Framework for PPDM

In data mining or knowledge discovery from databases (KDD) process the data is collected by single/various enterprises and stored at various databases. Then, it is transfer to a format suitable for analytical purposes, stored in large data warehouses and then data mining algorithms are applied on it for the generation of knowledge or information. With the riveted of protecting privacy the model has to be derive. Privacy constraints cannot be applied at one step; it needs to be kept in mind along with the data mining process all the way from data collection to the generation of knowledge or information. There suggests three levels where privacy concerns are taken care of. At level 1, the raw data collected from a single or multiple databases or even data marts is transformed into a format that is well suited for analytical purposes. Even at this stage, privacy concerns are needed to be taken care of. Researchers have applied various techniques at this stage but most of them deal with making the raw data suitable for analysis [6].

B. Classification of PPDM

According to [7] work done in PPDM can be classified according to various categories. These are Data Distribution-the PPDM algorithms can be first divided into two major categories, centralized and distributed data, based on the distribution of data. In a centralized database environment, data are all stored in a single database; while, in a distributed database environment, data are stored in various databases. Distributed data scenarios can be further classified into horizontal and vertical data distributions.

VI. Confidentiality Issues in Data Mining.

One of the problems that arise in any large collection of data is that of confidentiality. The need for privacy is sometimes due to law (e.g., for medical databases) or can be inspire by business interests. However, there are situations where the sharing of data can lead to mutual gain. A key utility of large databases today is research, whether it is scientific or economic and market oriented. For example, the medical field has much to gain by merge data for research; as can even competing businesses with mutual interests. Despite the potential gain, this is often not possible due to the confidentiality issues which arise [11].

VII. Secure Multiparty Computation

Distributed computing revolve the scenario where a number of distinct, yet connected, computing devices (or parties) wish to carry out a joint computation of some function. For example, these devices may be servers who hold a distributed database system, and the function to be computed may be a database update of some kind. The objective of secure multiparty computation is to prepare parties to carry out such distributed computing tasks in a secure manner. Whereas distributed computing classically deals with questions of computing under the threat of machine crashes and other unwitting faults, secure multiparty computation is concerned with the possibility of consciously malicious behavior by some adverse entity. That is, it is assumed that a protocol execution may come under attack by an external entity, or even by a subset of the participating parties. The motive of this attack may be to learn private information or cause the result of the computation to be incorrect. Thus, two important essentials on any secure computation protocol are correctness and privacy. The privacy needs states that nothing should be learned beyond what is absolutely necessary; more exactly, parties should learn about their output and nothing else. The correctness needs states that each party should receive its correct output. Therefore, the adversary must not be able to cause the result of the computation to vary from the function that the parties had set out to compute.

The setting of secure multiparty computation beset tasks as simple broadcast, and it is also as complex as, private information retrieval schemes, electronic voting, , anonymous transactions, contract signing, electronic cash schemes and electronic auctions. Take the example of voting and auctions. The privacy needs for an election protocol assure that no parties learn anything about the individual votes of other parties, and the correctness needs assure that no affiliation of parties can essence the outcome of the election beyond just voting for their preferred candidate. Likewise, in an auction protocol, the privacy requirement ensures that only the winning bid is revealed, and the correctness requirement ensures that the highest bidder is indeed the party to win. Due to its generality, the setting of secure multiparty computation can model almost every cryptographic problem [8].

VIII. Secure Computation and Privacy-Preserving Data Mining

Secure multiparty computation tells that for any functionality, it is possible to compute it without acknowledge anything beyond the output. However, it does not consider the question of how much information about the input is acknowledged by that output. We take the example of computing the “average". A secure protocol can evaluate the average of parties' salaries without acknowledge anything except output. However, if two parties run the protocol, then each party can evaluate the other party's salary exactly. Thus, even though the protocol acknowledges nothing, the output itself reveals everything. This shows that although secure computation is an extremely powerful tool and it is very helpful in the field of privacy-preserving data mining, it can only be applied once it has been decided that the function in question is “safe". This latter question of what functions can be safely computed is the focus of the field of “privacy". It emphasis that do not criticize the role of secure computation in privacy-preserving data mining in any way. Rather, it see the fields of privacy and secure computation as complementary: the first is needed for the safety purpose and the second is required to compute the function so that it remains safe (i.e., by using secure computation we are assured that only the output is acknowledge and so the resolve that the function is safe suffices for saying that it can be computed).

IX. Conclusion

The wide use of data mining tools in both the public and private sectors increasing concerns respecting the potentially sensitive nature of much of the data being mined. Privacy preserving data mining objective is to achieve the complicated property of enabling a data mining algorithm to use data without ever actually “seeing” it. Thus, the benefits of data mining can be gain, without compromising the privacy of concerned individuals.

our objective is the success of privacy preserving data mining may depend on the ability to find new areas that provide both the accurate security that are unavoidable and it should efficient in secure manner.

X. References:

[1] Giudici Paolo, “Applied Data-Mining: Statistical Methods for Business and Industry” 28 January 2003 page no. 1.

[2] http://www.usc.edu/dept/ancntr/Paris-in-LA/Analysis/discovery.html.

[3] Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy, “Advances in Knowledge Discovery and Data Mining” January 1996.

[4] G.Rama Krishna, G.V.Ajaresh, I.Jaya Kumar Naik, Parshu Ram Dhungyel, D.Karuna Prasad “A New Approach to Maintain Privacy And Accuracy In Classification Data Mining” IJCSET Volume 2, Issue 1, January 2012 Y.

[5] Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam, “A study of data mining tools in knowledge discovery process”, IJSCE, Volume-2, Issue-3, July 2012.

[6] Jasbir Malik, Rajkumar, “A Hybrid Approach Using C Mean and CART for Classification in Data Mining”, IJCSMS, Vol. 12, Issue 03, Sept 2012.

[7] Malik, M.B.,“Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”, ICCCT, Nov. 2012.

[8] Yehuda Lindell, Benny Pinkas, “Secure Multiparty Computation for Privacy-Preserving Data Mining”, JPC, Volume-1, Issue-1, 2009.

[9] Two Crows Corporation, “Introduction to Data Mining and Knowledge Discovery” Third Edition 2005 page no. 1.

[10] http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm.

[11] Brands, Estelle and Gerritsen, Rob. Assocation and Sequencing. “DBMS, Data Mining Solutions Supplement”, Miller Freeman, Inc. 1998 page no 54.