International Conference on Recent Innovations in Science, Technology, Management and Environment
Indian Federation of United Nations Associations, New Delhi (India) (ICRISTME-16)
19th June 2016, www.conferenceworld.in ISBN: 978-81-931039-1-3

CASE STUDY ON - DATA MINING SECURITY ISSUES AND REMEDIES IN PRIVACY PRESERVATION

Shivali Yadav[1], Dr. K P Yadav[2]

In partial Fulfillment of the Degree of PHD (Computer Science and Application)

Department of Computer Science n Application

ABSTRACT

In recent years, data mining towards privacy-preserving has been deliberate widely; it is because of the wide explosion of susceptible in sequence on the internet. Numeral algorithmic methods have been intended for privacy-preserving data mining. In this paper, we discuss and examine methods for privacy.

Keywords: Data mining, Privacy-preserving, randomization

I. INTRODUCTION

Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Knowledge discovery is needed to make sense and use of data. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. [1,2,3]

Usually, data mining e.g. data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information from many different dimensions or angles, categorize it, and summarize the relationships identified [6]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.[4] Although data mining is a comparatively new term but the technology is not. Companies have used powerful computers to filter through volumes of superstore scanner data and analyze market research reports for many years [6]. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.[5] Data mining, the discovery of new and interesting patterns in large datasets, is an exploding field. One aspect is the use of data mining to improve security, e.g., for intrusion detection. A second aspect is the potential security hazards posed when an adversary has data mining capabilities. Privacy issues have attracted the attention of the media, politicians, government agencies, businesses, and privacy advocates [6].

II. REVIEW OF LITERATURE

There has been much interest recently on using data mining for counter-terrorism applications. For example, data mining can be used to detect unusual patterns, terrorist activities and fraudulent behavior. While all of these applications of data mining can benefit humans and save lives, there is also a negative side to this technology, since it could be a threat to the privacy of individuals. This is because data mining tools are available on the web

or otherwise and even naïve users can apply these tools to extract information from the data stored in various databases and files and consequently violate the privacy of the individuals. Recently we have heard a lot about national security vs. privacy in newspapers, magazines and television talk shows. This is mainly due to the fact that people are now realizing that to handle terrorism; the government may need to collect information about individuals. This is causing a major concern with various civil liberties unions.

We are beginning to realize that many of the techniques that were developed for the past two decades or soon the inference problem can now be used to handle privacy. One of the challenges to securing databases is the inference problem. Inference is the process of users posing queries and deducing unauthorized information from the legitimate responses that they receive. This problem has been discussed quite a lot over the past two decades. However, data mining makes this problem worse. Users now have sophisticated tools that they can use to get data and deduce patterns that could be sensitive. Without these data mining tools, users would have to be fairly sophisticated in their reasoning to be able to deduce information from posing queries to the databases. That is, data mining tools make the inference problem quite dangerous. While the inference problem mainly deals with secrecy and confidentiality we are beginning to see many parallels between the inference problem and what we now call the privacy problem.

2.1 Security Concern In Data Mining

Databases are imperative and indispensable mechanism of dissimilar government and private association. To defend the data of the databases worn in information warehouse and then data mining is innermost theme of security structure. The necessities of data mining sanctuary alarmed with the following character.

- Access Control
- Logical Database Integrity
- Element Integrity
- User Authentication
- Physical Database Integrity
- Auditability

III. TAXONOMY OF PRIVACY PRESERVING TECHNIQUES [12]

There are many methodologies which have been accepted for privacy preserving data mining. We can categorize them based on the following measurements:

- Data Distribution
- Data Modification
- Data Mining Algorithm
- Data or Rule hiding
- Privacy Preservation

The first dimension discusses to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to these

cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places.

The second dimension discusses to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection [7, 8]. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include:

- Perturbation, which is accomplished by the alteration of an attribute value by a new value, blocking, which is the replacement of an existing attribute value with a "?",
- Aggregation or merging which is the combination of several values into a coarser category,
- Swapping that refers to interchanging values of individual records, and sampling, which refers to releasing data for only a sample of a population.

IV. PRIVACY PRESERVING ALGORITHMS

4.1 Heuristic-Based Techniques

A number of techniques have been developed for a number of data mining techniques like classification, association rule discovery and clustering, based on the premise that selective data modification or sanitization is an NP-Hard problem, and for this reason, heuristics can be used to address the complexity issues.

4.2 Centralized Data Perturbation-Based Association Rule Confusion

A formal proof that the optimal sanitization is an NP Hard problem for the hiding of sensitive large item sets in the context of association rules discovery, have been given in [9].

4.3 Centralized Data Blocking-Based Association Rule Confusion

One of the data modification approaches which have been used for association rule confusion is data blocking [10]. The approach of blocking is implemented by replacing certain attributes of some data items with a question mark. It is sometimes more desirable for specific applications (i.e., medical applications) to replace a real value by an unknown value instead of placing a false value. An approach which applies blocking to the association rule confusion has been presented in [11]. The introduction of this new special value in the dataset imposes some changes on the definition of the support and confidence of an association rule. In this regard, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly. As long as the support and/or the confidence of a sensitive rule lies below the middle in these two ranges of values, then we expect that the confidentiality of data is not violated. Notice that for an algorithm used for rule confusion in such a case, both 1-values and 0-values should be mapped to question marks in an interleaved fashion; otherwise, the origin of the question marks will be obvious. An extension of this work with a detailed discussion on how effective is this approach on reconstructing the confused rules, can be found in [11].

V. EVALUATION OF PRIVACY PRESERVING ALGORITHMS

An important aspect in the development and assessment of algorithms and tools, for privacy preserving data mining is the identification of suitable evaluation criteria and the development of related benchmarks. It is often

the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. Rather, an algorithm may perform better that another one on specific criteria, such as performance and/or data utility. It is thus important to provide users with a set of metrics which will enable them to select the most appropriate privacy preserving technique for the data at hand; with respect to some specific parameters they are interested in optimizing. A preliminary list of evaluation parameters to be used for assessing the quality of privacy preserving data mining algorithms is given below:

- the *performance* of the proposed algorithms in terms of time requirements, that is the time needed by each algorithm to hide a specified set of sensitive information;
- the *data utility* after the application of the privacy preserving technique, which is equivalent with the minimization of the information loss or else the loss in the functionality of the data:
- the *level of uncertainty* with which the sensitive information that have been hidden can still be predicted;
- the *resistance* accomplished by the privacy algorithms to different data mining techniques.

VI. CONCLUSION

This paper discussed taxonomy of privacy preserving data mining approaches. Along with the expansion of data psychiatry and dispensation technique, the solitude revelation trouble about personage or company is unavoidably uncovered when releasing or allocation data to colliery useful conclusion information and acquaintance and then provide a confinement to the research scenery on privacy preserving data mining.

REFERENCE

[1]. 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.

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

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

[4]. Larose, D. T., “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[5]. L. Getoor, C. P. Diehl. “Link mining: a survey”, ACM SIGKDD Explorations, vol. 7, pp. 3-12, 2005.

[6]. Dileep Kumar Singh, Vishnu Swaroop, Data Security and Privacy in Data Mining:

[7]. Research Issues & Preparation, International Journal of Computer Trends and Technology- volume4Issue2- 2013

[8]. Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim, and Vassilios S. Verykios, Disclosure Limitation of Sensitive Rules, In Proceedings of the IEEE Knolwedge and Data Engineering Workshop (1999),45–52.

[9]. Chris Clifton and Donald Marks, Security and privacy implications of data mining, In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.

[10]. L. Sweeney, (2002)."k-anonymity: a model for protecting privacy ", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570.

[11]. Evfimievski, A.Srikant, R.Agrawal, and GehrkeJ(2002),"Privacy preserving mining of association rules". In Proc.KDD02, pp. 217-228.

[12]. DakshiAgrawal and Charu C. Aggarwal, On the design and quantification of privacy preservingdata mining algorithms, In Proceedings of the 20th ACMSymposium on Principles of Database Systems (2001), 247–255.

[13]. Md. Riyazuddin, .Dr.V.V.S.S.S.Balaram, An Empirical Study on Privacy Preserving Data Mining, International Journal of Engineering Trends and Technology- Volume3Issue6- 2012

[14]. Dileep Kumar Singh, Vishnu Swaroop, Data Security and Privacy in Data Mining: Research Issues & Preparation, International Journal of Computer Trends and Technology- volume4Issue2- 2013