International Journal of Information Technology and Management

Vol. IX, Issue No. XIV, November-2015, ISSN 2249-4510

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

International Journal of Information Technology and Management
Vol. IX, Issue No. XIV, November-2015, ISSN 2249-4510

Case Study on Data Mining Security Issues and Remedies in Privacy Preservation

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

[1]Research Scholar, Jodhpur National University, Rajasthan

[2]Director of KCC Institute of Technology and Management

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

- - - - - - - - - - - - - - X - - - - - - - - - - - - - -

1. 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]

2. 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.

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

3. 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.

4. PRIVACY PRESERVING ALGORITHMS

➢ 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.

➢ 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].

➢ 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].

5. 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.

6. DISTRIBUTED PRIVACY-PRESERVING DATA MINING TECHNIQUE:

The key goal in most “distributed methods for privacy-preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within the different participants. Thus, the participants may wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. For this purpose, the data sets may either be *horizontally partitioned* or be *vertically partitioned*. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of which has the same set of attributes. In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records, both kinds of partitioning pose different challenges to the problem of distributed privacy-preserving data mining” [24].

“The problem of distributed privacy-preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations.” “A broad overview of the intersection between the fields of cryptography and privacy-preserving data mining may be found in [14]. The broad approach to cryptographic methods tends to compute functions over inputs provided by multiple recipients without actually sharing the inputs with one another. For example, in a 2-party setting, Alice and Bob may have two inputs x and y respectively, and may wish to both compute the function $f\ (x, y)$ without revealing x or y to each other. This problem can also be generalized across k parties by designing the k argument function $h(x_1\ .\ .\ .\ x_k\ )$. Many data mining algorithms may be viewed in the context of repetitive computations of many such primitive functions such as the scalar dot product, secure sum etc. In order to compute the function $f\ (x,\ y)$ or $h(x_1\ .\ .\ .\ ,\ x_k\ )$, a protocol de-signed for exchanging information in such a way that the function is computed without compromising privacy.” We note that the robustness of the protocol de-pends upon the level of trust one is willing to place on the two participants Alice and Bob. This is because the protocol may be subjected to various kinds of adversarial behavior:

- Semi-honest Adversaries: In this case, the participants Alice and Bob are curious and attempt to learn from the information received by them during the protocol, but do not deviate from the protocol themselves. In many situations, this may be considered a realistic model of adversarial behavior.

- Malicious Adversaries: In this case, Alice and Bob may vary from the protocol, and may send sop Histicated inputs to one another to learn from the information received from each other.

A key building-block for many kinds of secure function evaluations is the 1 out of 2 oblivious-transfer protocol. This protocol was proposed in [15-16] and involves two parties: a *sender*, and a *receiver*. The sender’s input is a pair (x_0, x_1), and the receiver’s input is a bit value $\sigma \in \{0, 1\}$. At the end of the process, the receiver learns x_σ only, and the sender learns nothing. A number of simple solutions can be designed for this task. In one solution [16-17], the receiver generates two random public keys, K_0 and K_1, but the receiver knows only the decryption key for K_σ . The receiver sends these keys to the sender, who encrypts x_0 with K_0, x_1 with K_1, and sends the encrypted data back to the receiver. At this point, the receiver can only decrypt x_σ , since this is the only input for which they have the decryption key. We note that this is a semi-honest solution, since the intermediate steps require an assumption of trust. For example, it is assumed that when the receiver sends two keys to the sender, they indeed know the decryption key to only one of them. In order to deal with the case of malicious adversaries, one must ensure that the sender chooses the public keys according to the protocol. An efficient method for doing so is described in [18]. In [18], generalizations of the 1 out of 2 oblivious transfer protocols to the 1 out *N* case and *k* out of *N* case are described.

Since the oblivious "transfer protocol is used as a building block for secure multi-party computation, it may be repeated many times over a given function evaluation. Therefore, the computational effectiveness of the approach is important. Efficient methods for both semi-honest and malicious adversaries are discussed in [18]. More complex problems in this domain include the computation of probabilistic functions over a number of multi-party inputs [19]. Such powerful techniques can be used in order to abstract out the primitives from a number of computationally intensive data mining problems. Many of the above techniques have been described for the 2-party case, though generic solutions also exist for the multiparty case. Some important solutions for the multiparty case may be found in [20]."

The oblivious "transfer protocol can be used in order to compute several data mining primitives related to vector distances in multi-dimensional space. A classic problem which is often used as a primitive for many other problems is that of computing the scalar dot-product in a distributed environment [21]. A fairly general set of methods in this direction are described in [22]. Many of these techniques work by sending changed or encrypted versions of the inputs to one another in order to compute the function with the different alternative versions followed by an oblivious transfer protocol to retrieve the correct value of the final output. A systematic framework is described in [22] to transform normal data mining problems to secure multi-party computation problems. The problems discussed in [22] include those of clustering, classification, associ-ation rule mining, data summarization, and generalization. A second set of methods for distributed privacy-preserving data mining is discussed in [23] in which the secure multi-party computation of a number of important data min-ing primitives is discussed. These methods include the secure sum, the secure set union, the secure size of set intersection and the scalar product. These techniques can be used as data mining primitives for secure multi-party computation over a variety of horizontally and vertically partitioned data sets."

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: Research Issues & Preparation, International Journal of Computer Trends and Technology- volume4Issue2- 2013

[7] 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.

[8] 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.

[9] L. Sweeney, (2002)."k-anonymity: a model for protecting privacy ", International Journal

on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570.

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

[11] 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.

[12] 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.

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

[14] Pinkas B.: Cryptographic Techniques for Privacy-Preserving Data Min-ing. ACM SIGKDD Explorations, 4(2), 2002.

[15] Rabin M. O.: How to exchange secrets by oblivious transfer, Technical Report TR-81, Aiken Corporation Laboratory, 1981.

[16] Even S., Goldreich O., Lempel A.: A Randomized Protocol for Signing Contracts. Communications of the ACM, vol 28, 1985.

[17] Goldreich O.: Secure Multi-Party Computation, Unpublished Manu-script, 2002.

[18] Bayardo R. J., Agrawal R.: Data Privacy through Optimal k-Anonymization. Proceedings of the ICDE Conference, pp. 217–228, 2005.

[19] Yao A. C.: How to Generate and Exchange Secrets. FOCS Conferemce, 1986.

[20] Chaum D., Crepeau C., Damgard I.: Multiparty unconditionally secure protocols. ACM STOC Conference, 1988.

[21] Ioannidis I., Grama A., Atallah M.: A secure protocol for computing dot products in clustered and distributed environments, International Con-ference on Parallel Processing, 2002.

[22] Du W., Atallah M.: Secure Multi-party Computation: A Review and Open Problems.CERIAS Tech. Report 2001-51, Purdue University, 2001.

[23] Clifton C., Kantarcioglou M., Lin X., Zhu M.: Tools for privacy-preserving distributed data mining. ACM SIGKDD Explorations, 4(2), 2002.

[24] Charu C. Aggarwal, A General Survey of Privacy-Preserving Data Mining Models and Algorithms, IBM T. J. Watson Research Center