Privacy Preservation in Data Mining: A Stat-of-Art

Neha Patel
Research Scholar, Department of Computer Science & Engineering, RKDF IST, RGPV University, Bhopal, INDIA.

Prof. Shrikant Lade
HOD, Department of Computer Science & Engineering, RKDF IST, RGPV University, Bhopal, INDIA.

Abstract – There is a rapid enhancement in the development of data mining. Data mining refer as a technique for extraction of data from huge database. It becomes a crucial topic in research community is Privacy preserving data mining (PPDM). It is necessary to keep the ratio in between privacy protection and knowledge discovery. The motive is to conceal sensitive item sets in order to prevent them from any intentional changes made by hackers in the database. For solving such issues some of algorithms shown through various authors worldwide. The motive of this survey paper is to become familiar with the existing privacy preserving data mining techniques and to get the efficiency.

Index Terms – Privacy, Data Mining, Clustering.

1. INTRODUCTION

Privacy preserving data mining (PPDM) associates with the field of data mining which is looking for safeguard for crucial information from the unsanctioned or unsolicited disclosure. Most conventional data mining method observe and model their dataset statistically, in a combine summation, On the other hand privacy preservation having the motive for securing against the disclosure of separate data records. Such domain separation attribute refer as the technical feasibility of PPDM.

In History the issues associates to PPDM were like initially studied through the national statistical agencies curious in aggregating private social and economic data, like as census and tax records, and available for observe through public servants, companies, and researchers. Building accurate socio-economical models is important for business strategy and public policy. Still, there is not any solution for knowing what models might be required &, nor is it suitable for the statistical agency for performing every data processing for all, performing the role for a "trusted third party." On the place of, the agency offers the data in that sanitized form that permits an statistical processing and secured the secrecy of separate records, resoling a problem refer as privacy preserving data publishing. For a concert on statistical databases watch Adam & Wortmann and Willenborg & de Waal.

The word "privacy preserving data mining" was proposed in papers (Agrawal & Srikant, 2000) & (Lindell & Pinkas, 2000). These papers possess two fundamental faults of PPDM, privacy preserving data aggregation and mining a dataset separation across various private enterprises. Agrawal and Srikant (2000) proposed a randomization algorithm that permits a huge number of users contributing their private records for good centralized data mining at the time of limiting their disclosure of its values; Lindell and Pinkas (2000) discover a cryptographic protocol for the decision tree building over a dataset horizontally separated in between two parties. These techniques were subsequently refined and forwarded through various researchers worldwide.

2. DATA MINING

Data mining is the exploration and analysis of large data sets, in order to discover meaningful pattern and rules. The key idea is to find effective way to combine the computer's power to process the data with the human eye's ability to detect patterns. The objective of data mining is to design and work efficiently with large data sets. Data mining is the component of wider process called knowledge discovery from database. [4-1]. Data Mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It has been defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" The definition of data mining is closely related to another commonly used term knowledge discovery [2-2]. Data mining is an interdisciplinary, integrated database, artificial intelligence, machine learning, statistics, etc. Many areas of theory and technology in current era are databases, artificial intelligence, data mining and statistics is a study of three strong large technology pillars. Data mining is a multi-step process, requires accessing and preparing data for a mining the data, data mining algorithm, analyzing results and taking appropriate action. The data, which is accessed, can be stored in one or more operational databases.

David Hand and Heikki Mannila categorize data mining into five tasks:

2.1 EXPLORATORY DATA ANALYSIS (EDA)

Typically interactive and visual, EDA techniques simply explore the data without any preconceived idea of what to look for.

2.2 DESCRIPTIVE MODELING:

A descriptive model should completely describe the data (or the process generating it); examples include models for the data's overall probability distribution (density estimation), partitions

of the dimensional space into groups (cluster analysis and segmentation), and descriptions of the relationship between variables (dependency modeling).

2.3 PREDICTIVE MODELING: CLASSIFICATION AND REGRESSION:

The goal here is to build a model that can predict the value of a single variable based on the values of the other variables. In classification, the variable being predicted is categorical, where as in regression, it‘s quantitative.

2.4 DISCOVERING PATTERNS AND RULES:

Instead of building models, we can also look for patterns or rules. Association rules aim to find frequent associations among items or features, whereas outlier analysis or detection focuses on finding outlying records that differ significantly from the majority.

2.5 RETRIEVAL BY CONTENT:

Given a pattern, we try to find similar patterns from the data set.

The first three tasks outputs that essentially summarize the data in various ways; the last two find specific patterns, but they‘re often generalized and don‘t reflect particular data items. Because these models generally don‘t contain individual data values, they don‘t present an immediate threat to privacy. However, there‘s still the issue of inference. A perfect classifier, for example, would enable discovery of the target class, even if the individuals’ target classes weren‘t directly disclosed. Practically probabilistic inferences are more likely, giving the model‘s possessor aprobabilistic estimate of private values. The privacy problem with data mining does not depend on its results and the methods used to get those results.

3. PRIVACY VIOLATION IN DATA MINING

Getting familiar with the privacy in data mining needs understanding like how privacy can be discussed and the exact definition for preventing privacy violation. Basically, one major factor which contributes in privacy violation in the data mining: data misuse.

Users' privacy could be destructing various ways and with distinct intentions. Moreover data mining would be extremely important in many applications (e.g., business, medical analysis, etc.), it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be destructed when personal data are implemented for some other purposes as subsequent to that actual transaction in between an individual and the organization the time information was taken. One of their sources of the privacy violation refer as data magnets [17-3]. Data magnets are methods and instruments used for collecting personal data. These data magnets include large collecting information by on-line registration, identifying users by IP addresses; software downloads that needs registration, and indirectly taking information for secondary used. In various cases, users might be not aware that their information is taken or not aware that how those information is achieved [7-4, 13--5].Worse part in privacy invasion occurred through secondary used of data at the time individuals are not familiar of “behind the scenes” implementation of data mining methods [11-6]. Moreover, private data can be implemented for secondary utilization out of the users' control and privacy laws. This trend has become a reasoned for uncontrollable privacy violation not only because of data mining, but also fundamentally due to the misuse of data.

4. DEFINING PRIVACY PRESERVATION IN DATA MINING

In basic, privacy preservation takes place in two main dimensions: users' personal information and information anxious about their collective activity. We have taken as individual privacy preservation and afterwards as collective privacy preservation, which is associate with corporate privacy in [3-7].

4.1 Individual Privacy Preservation:

The ultimate goal of the data privacy is the securing of private identifiable information. Generally, information is taken as private identifiable when it is being attached, directly or indirectly, to an individual person. Hence, when personal data being subjected towards mining, the quality values relates with individuals are private and have to be secured from the disclosure. Miners become capable of learning from global models instead of the characteristics of a specific individual.

4.2 Collective privacy preservation:

Securing the private data might be not enough. Therefore Some times, we required to secure against learning sensitive knowledge showing few activities of some group. We discuss about protection of sensitive information as collective privacy preservation. The motive here is pretty similar as for statistical databases, where security control mechanisms offers aggregate information related to groups (population) and, together, should stop disclosure of sensitive information about the individuals. Somehow, not like as in case for statistical databases, distinct objective of the collective privacy preservation for conserving the strategic patterns which are paramount for planning decisions, instead of reducing the distortion of every statistics (e.g., bias and precision). We can also say that the motive here is not to secure private identifiable information but to some patterns also and patters must not be discovered.

Key Points:

- The geometric data transformation methods (GDTMs) that distort confidential numerical attributes for copping up with privacy protection in clustering analysis.

- End users are able to use their own tools so that the constraint for privacy has to be applied before the mining process on the data by data transformation.
- Data owners should not only cop up with privacy needs but also certain about valid clustering results.
- One major disadvantage is that the privacy preservation of individuals when data is shared for analysis is very complex.

5. VARIOUS ASPECTS OF PRIVACY PRESERVING

Data mining techniques are applicable for transformation, which decrease the effects of the underlying data while applying towards data mining techniques or algorithms. Moreover there is a natural tradeoff in between privacy and accuracy; by this tradeoff is influenced by that particular algorithm which implemented for privacy preservation. A main problem is to maintain highest utility of that data in absence of compromising the underlying privacy constraints. A wide overview of distinct utility depends on techniques for privacy-preserving data mining is shown. The issue of generates utility depends on algorithms to perform effectively with specific kinds of data mining issues is addressed.

5.1 MINING ASSOCIATION RULES UNDER PRIVACY CONSTRAINTS

As association rule mining refers as important issue in data mining, we have committed many chapters to such problems. We have two aspects to this privacy preserving association rule mining issues. When the input gets perturbed, it becomes a big problem to properly find out the association rules on that perturbed data. A differ issue is of output association rule privacy. Here in such case, we tried to ensure that none of the association rules in that output result is disclosure sensitive data. This issue is taken as association rule hiding [5.1-8] through database community, and that contingency table privacy-preservation through the statistical community. The issue of output association rule privacy is discussed. A brief survey of association rule hiding from that Perspective of that database community is being studied.

5.2 CRYPTOGRAPHIC METHODS

In various cases, multiple parties might desired to share overall private data, in absence of leaking any sensitive information from it [6.1-9]. For example, distinct superstores having sensitive sales data desired to coordinate among themselves in familiar aggregate trends in absence of leaking the trends of its individual stores. This needs protective and cryptographic protocols for displaying the information across the distinct parties. The data might be spread in two ways in between two sites: In the field of privacy preserving data mining refer as those data streams, where data grows very fast at an unlimited flow rate. In these cases, the error of privacy-preservation is pretty challenging as the data is rapidly flowing. Additionally, the rapid nature of the data streams ignores the possibility of implementing the last history of its data.

6. RELATED WORK

Data mining in a fair information practices perspective was initially discussed in [15-10]. O'Leary studied the field of the OECD guidelines in the knowledge discovery. The crucial search of this study was like the OCDE guidelines could not address in various important issues based on knowledge discovery, and hence various principles are much general or unenforceable. Here our work is orthogonal [15-10]. We analyze the influence of the OECD principles in that context of PPDM categorizing them in distinct groups of relevance. Here we show that that OECD guidelines are taken world-wide and hence they shown the primary elements for PPDM standardization. We talk about how the community in the PPDM can draw some principles and policies from those OECD guidelines.

Recently, Clifton et al. talked about the meaning of the PPDM like a foundation for upcoming research in this stream [3-11]. That work proposed some definitions for the PPDM and talked few metrics for information leaked in data mining. The main motive of our work is to offer standardization issues in the PPDM. Our effort generates design of privacy principles and policies, and needs for developing technical solutions for the PPDM.

7. CONCLUSION

The above discussion about the privacy preserving data mining techniques is much satisfactory, but there is always a crazy about more modifications in it. Here in this paper PPDM can be important tool for finding the loopholes and errors of existing data mining techniques. This survey is certain about the efficient privacy preserving of data. The implementation of existing algorithms works to decrease the impact of PPDM on the source database directly. A comparative study assists in generating a latest system which combines every advantage and eradicates the drawbacks of such systems.

REFERENCES

[1] Han, J., Kamber, M., Data Mining Concepts and Techniques, Morgan Kaufmann Publisher, 2001.
[2] PavelBerkhin, A Survey of Clustering Data Mining Techniques, pp.25-71, 2002.
[3] A. Rezgur, A. Bouguettaya, and M. Y. Eltoweissy. Privacy on the Web: Facts, Challenges, and Solutions. IEEE Security & Privacy, 1(6):40-49, Nov-Dec 2003..
[4] M. J. Culnan. How Did They Get My Name?: An Exploratory Investigation of Consumer Attitudes Toward Secondary Information. MIS Quartely, 17(3):341-363, September 1993.
[5] K. C. Laudon. Markets and Privacy. Communication of the ACM, 39(9):92-104, September 1996.
[6] G. H. John. Behind-the-Scenes Data Mining.Newletter of ACM SIG on KDDM, 1(1):9-11, June 1999.

[7] C. Clifton, M. Kantarcioglu, and J. Vaidya. Defining Privacy For Data Mining. In Proc. of the National Science Foundation Workshop on Next Generation Data Mining, pages 126-133, Baltimore, MD, USA, November 2002.
[8] Stanley R. M. Oliveira, and Osmar R. Zaiane, "Revisiting Privacy Preserving Clustering by Data Transformation," Journal of Information and Data Management, vol. 1, no. 1, 2010.
[9] Aggarwal C., Pei J., Zhang B. A Framework for Privacy Preservation against Adversarial Data Mining.ACM KDD Conference, 2006.
[10] D. E. O'Leary. Some Privacy Issues in Knowledge Discovery: The OECD Personal Privacy Guidelines. IEEE EXPERT, 10(2):48-52, April 1995.
[11] Wang Qiang, Megalooikonomou, Vasileios, A dimensionality reduction technique for efficient time series similarity analysis, Inf. Syst. 33, 1 (Mar.2008), 115- 132.