ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com

International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

Privacy Preserving Data Mining Techniques – A General Survey

P.SARASWATHI[1]
Research Scholar (Part-Time),
Bharathiyar University,
Coimbatore,
Tamil Nadu, India.
saisaraswathi.research@gmail.com

Dr.Mrs. N. NAGADEEPA[2]
Principal
Karur Velalar College of Arts and Science for Women
Kuppam (po), Karur 639 111, India.
nagadeepa1012@gmail.com

Abstract:

Data mining is the method of exploring and analyzing large quantities of data by automatic or semiautomatic means in order to discover meaningful patterns and rules. At present information like PAN NUMBER data collection is omnipresent, and every operation are recorded some place. Mounting data collection along with the advent of scrutiny tools capable of managing enormous volumes of information, has led to privacy concerns. Defending private data is an important concern for society. Several laws now require explicit consent prior to analysis of an individual's data, but its value is not limited to individuals: corporations' needs to protect their information's privacy, even though sharing it for analysis could benefit the company. Clearly, the trade-off between distribution of information for analysis and keeping it secret to preserve corporate trade secrets and customer privacy is a growing challenge. This paper provides a basic survey of different privacy preserving data mining methods and points out their Pros and Cons of techniques.

Keywords: - Data Mining, Information, PPDM, Data Transformation, Association Rule

1. INTRODUCTION

The span of information technologies and the internet in the past decades has brought prosperity of entity information into the hands of marketable companies and government agencies. Data owners persistently seek out to make enhanced use of the information they own, and exploit data mining tools to mine useful knowledge and patterns from the data. Privacy Preserving Data Mining (PPDM) is a research area concerned with the privacy driven from individually identifiable information when measured for data mining. Therefore, PPDM has become an increasingly vital field of research. PPDM is a novel study direction in data mining. A number of methods and techniques have been developed for privacy preserving data mining. The set of criteria has been recognized based on which a PPDM algorithm can be evaluated.

- ✓ Privacy level
- ✓ Hiding failure
- ✓ Data quality
- ✓ Complexity

The major challenges of PPDM method for association rule hiding are high information loss, expensive, difficult to recover original data after hiding and should be efficient enough for very large datasets. PPDM is a research area concerned with the

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

privacy driven from personally identifiable information when considered for data mining. This work addresses the privacy problem by considering the privacy and algorithmic requirements simultaneously.

The objective of this work is to implement a distortion algorithm using association rule hiding for privacy preserving data mining which would be efficient in providing confidentiality and improve the performance. The debate on PPDM has received special attention as data mining has been widely adopted by public and private organizations. The excessive number of techniques is leading to confusion among developers, practitioners, and others interested in this technology.

2. CHALLENGES AND GOALS IN PPDM

One of the most important challenges in PPDM now is to establish the groundwork for further research and development in this area. A Privacy Violation in Data Mining Understanding privacy in data mining requires understanding how privacy can be violated and the possible means for preventing privacy violation. In general, one major factor contributes to privacy violation in data mining: the misuse of data. Users' privacy can be violated in different ways and with different intentions. It can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected. One of the sources of privacy violation is called data magnets.

Data magnets are techniques and tools used to collect personal data. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected. In particular, collected personal data can be used for secondary usage largely beyond the users' control and privacy laws. This scenario has led to an uncontrollable privacy violation not because of data mining itself, but fundamentally because of the misuse of data.

In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. It refer to the former as individual privacy preservation and the latter as collective privacy preservation, which is related to corporate privacy. The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual. Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation.

The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to protect sensitive knowledge that can provide competitive advantage in the business world. In the case of collective privacy preservation, organizations have to cope with some interesting conflicts. For instance, when personal

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

information undergoes analysis processes that produce new facts about users' shopping patterns, hobbies, or preferences, these facts could be used in recommender systems to predict or affect their future shopping patterns.

In general, this scenario is beneficial to both users and organizations. However, when organizations share data in a collaborative project, the goal is not only to protect personally identifiable information but also sensitive knowledge represented by some strategic patterns.

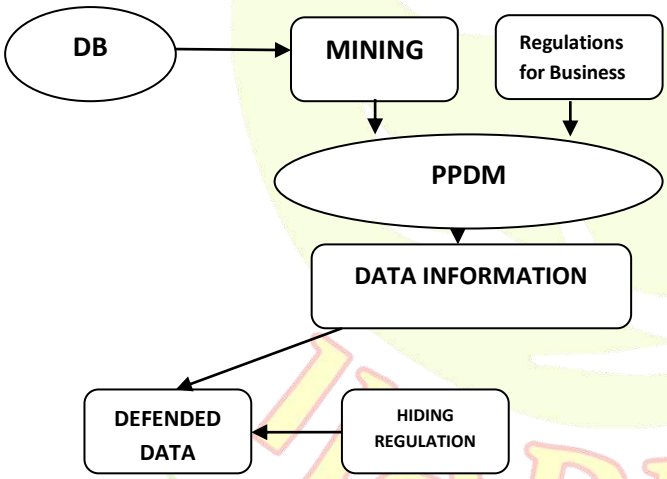

Fig 1: Block Diagram for PPDM

3. RELATED WORK:

Based on the concept of roles and permissions in the market, there are a number of existing systems on which we will take a brief look on. Previously two approaches of privacy preserving data mining are defined. In the first one, the aim is to preserve customer privacy by perturbing the data values and the other approach uses cryptographic tools to build various models for data mining process. In this section, some of the recent researches are described.

3.1 Methodology

A. Hiding Association Rules:

Author of the paper suggested some rules for hiding sensitivity by changing the support and the confidence of the association rule or frequent item set as data mining mainly deals with generation of association rules. In order to hide an association rule a new concept of 'not altering the support' of the sensitive item(s) has been proposed in this work. The proposed algorithm is that support for the secured data is unchanged. Instead, only the position of the secured data set can be changed by random method. It provide the use of a different technique for modifying the database transaction to reduce the confidence of secured rules. One of the main disadvantages of the existing approaches is the approach hides rules having sensitive items either in the right side or in the left side. In this proposed approach secured by random based method.

B. Privacy Preserving Clustering by Data Transformation:

Preserving the privacy of individuals when data are shared for clustering was a complex problem. The challenge was how to protect the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis.

In the ref [5] revisited a family of geometric data transformation methods (GDTMs) that distort numerical attributes by scaling, rotations, translations or by the combination of all above transformations. This method was designed to specify privacy-preserving clustering, in context where data owners must meet privacy requirements as well as guarantee valid clustering results. Authors also provided a particularized, broad and advanced picture of methods for privacy-preserving clustering by data transformation. End users are able to use their own tools so that the constraint for privacy has to be

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

applied before the mining process on the data by data transformation.

C. Cryptographic technique:

The ways of cryptography are used to data encryption. Many Cryptography-based approaches have been proposed in the context of privacy preserving data mining algorithms. Cryptography-based approaches like Secure Multi-party Computation (SMC) are secure at the end of the computations. No party knows anything except its own input and the results. SMC method is a typical technique. The ref [4] presents four secure multiparty computations based on the methods that can support privacy preserving data mining. SMC is mainly uses in distributed environment.

The purpose of SMC is that it is necessary to guarantee the correctness of the calculation, but also to protect their respective input and output data from leaking when two or more participants who are carrying out the cooperation calculation.

D. Privacy Preserving:

Privacy preserving data mining (PPDM) is a divide into varies categories. We will review the basic concepts of PPDM and different studies performed in the area of PPDM under various categories. We shall concentrate on metrics that are used to measure the side-effects resulted from privacy preserving process [5]. Although many different approaches are employed to protect important data in today's networked environment, these methods often fail. One way to make data less vulnerable is to deploy Intrusion Detection System (IDS) in critical computer systems. In case a computer system is compromised, an early detection is the key for recovering lost or damaged data without much complexity.

In recent years, researchers have proposed a variety of approaches for increasing the intrusion detection efficiency and accuracy [6]. But most of these efforts concentrated on detecting intrusions at the network or operating system level. They are not capable of detecting malicious data corruptions, i.e., what particular data in the database are manipulated by which specific malicious database transaction(s). Without this information, fast damage assessment and recovery cannot be achieved.

4. CONCLUSION:

This paper carries out a wide survey of the different approaches for privacy preserving data mining, and points out the existing drawback. While all the proposed methods are only approximate to the goal of privacy preservation. To address this issue, the following problems should be widely studied: (A) Privacy and accuracy is a pair of contradiction; improving one usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched. (B) In distributed privacy preserving data mining areas.

Reference

[1] Nivetha.P.R Nivetha.P.R et al, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 10, October- 2013, pg. 166-170.

[2] Bengbua China, International Journal of Digital Content Technology and its Applications Volume 4, Number 9, December 2010

[3] R. Agrawal and R. Srikant, ―Privacy-Preserving Data Mining,‖ Proc. ACM SIGMOD Conf. Management of Data, ACM Press, 2000, pp. 439–450; 7. A.C. Yao, ―How to Generate and Exchange Secrets,‖ Proc. 27th IEEE Symp. Foundations of Computer Science, IEEE CS Press, 1986, pp. 162–167.

[4] Pei, J., Han, J., Pinto, H., Chen, Q, Dayal, U., and Hsu, M-C. PrefixSpan: Mining Sequential Patterns Efficiently by PrefixProjected Pattern Growth. In Proceeding of 2001

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology Ecology Science and Technology (IJARBEST)
Vol. 2, Special Issue I, January 2016 in association with
KARUR VELALAR COLLEGE OF ARTS AND SCIENCE FOR WOMEN, KARUR
Conference on Emerging Trends and Functional Applications of Computer Technology - 12th January 2016

[5] 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.

[6] P.Samarati,(2001). Protecting respondent's privacy in micro data release. In IEEE Transaction on knowledge and Data Engineering, pp.010-027.

[7] O. Goldreich, S. Micali, and A. Wigderson, ―How to Play any Mental Game: A Completeness Theorem for Protocols with Honest Majority,‖ Proc. 19th ACM Symp. Theory of Computing, ACM Press, 1987, pp. 218–229;

[8] M. Franklin and M. Yung, ―Varieties of Secure Distributed Computing,‖ Proc. Sequences II, Methods in Communications, Security and Computer Science, Springer-Verlag, 1991, pp. 392– 417.

[9] M. Kantarcıog˘lu and C. Clifton, ―Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data,‖ IEEE Trans. Knowledge Data Eng., vol. 16, no. 9, 2004, pp. 1026– 1037;

[10] J. Vaidya and C. Clifton, ―Secure Set Intersection Cardinality with Application to Association Rule Mining,‖ to be published in J. Computer Security, 2005.

[11] Y. Lindell and B. Pinkas, ―Privacy Preserving Data Mining,‖ J. Cryptology, vol. 15, no. 3, 2002, pp. 177–206.

[12] C. Clifton et al., ―Tools for Privacy Preserving Distributed Data Mining,‖ SIGKDD Explorations, vol. 4, no. 2, 2003, pp. 28–34; www.acm.org/sigs/sigkdd/explorations/ issue4-2/contents.htm.

[13] N.R. Adam and J.C. Wortmann, ―Security-Control Methods for Statistical Databases: A Comparative Study,‖ ACM Computing Surveys, vol. 21, no. 4, 1989, pp. 515–556;

[14] T.H. Hinke, H.S. Delugach, and R.P. Wolf, ―Protecting Databases from Inference Attacks,‖ Computers and Security, vol. 16, no. 8, 1997, pp. 687–708.