A Review of Privacy Preserving Data Mining Techniques

D Kumar[1*], Ankita[2]

[1]Amity Institute of Information Technology, Amity University, India

[2]Mewar University, Rajasthan, India.

ABSTRACT

Nowadays, privacy-preserving data mining (PPDM) is being studied comprehensively, because of the wide-ranging availability of crucial data available on the internet. There exists a variety of algorithmic techniques for privacy-preserving data mining. The main focus of these algorithms is the mining of required knowledge from large ocean of dataset, at the same time protecting the sensitive information. Privacy Preserving is vital feature of data mining and therefore study of accomplishing some data mining goals without compromising the privacy of the persons is not only challenging but also an assignment of realistic significance. This paper reviews different methods for privacy preserving data mining such as Randomization, K-anonymization, Association rules, Cryptographic technique which are required for maintaining Information sharing and privacy. Studies show that tradeoff between privacy and information loss creates a bottleneck while developing generic solutions. In this paper we present a review of the existing well-organized methodologies in the framework of privacy preservation in data mining.

KEYWORDS

Data mining, PPDM, privacy, condensation, perturbation, cryptography

INTRODUCTION

DATA MINING

Data Mining is the method for understanding enormous informational indexes so as to discover designs that can disengage key factors to make prescient models which will help in taking decisions by the management [1]

One of the most basic and most referred definitions of the data mining process, which focuses on its distinguishing characteristics, is given by Fayyad, Piatetsky-Shapiro and Smyth (1996), who define it as "the nontrivial development in order to find valid, novel, potentially useful, and eventually clear patterns in statistics."

PROBLEM STATEMENT

The main inspiration for privacy-preserving data mining, starts from the necessity to dig out new and helpful knowledge from huge repositories of data, and has turn out to be an efficient logical and decision-making means in corporations. The distribution of data for data mining can fetch a bunch of reward for study and industry collaboration. The mishandling of these methods may lead to the exposé of crucial information. However, large repositories of data contain private data that must be protected, before being published.

HOW DATA MINING IS CARRIED OUT IN RETAIL SECTOR

These days plenitude of information is accessible; be it disconnected or on the web. Every single part utilizes the information for reasons unknown or the other. Like retail segment utilizes the client's information to comprehend their decision inclinations, their shopping propensities, recurrence of purchasing and so on. This thusly causes the organization to settle on their vital choices up to the imprint so as to develop the organization right way.

So as to gather the client's information, an organization may pursue any of the strategies, like at the time of checkout or through direct conversation while shopping.

DOES PRIVACY OF CUSTOMER AT STAKE?

The scope of privacy can be viewed from 4 categories:

1. Information: which deals with the management of accumulation of individual information
2. Bodily: which identifies with physical damages from intrusive techniques
3. Interactions: which deals with any form of interactions
4. Territory limits: which identifies with the interference of physical restrictions

This paper will concentrate on the information classification, which covers the frameworks that gather, examine and distribute data.

After collecting the entire customer’s data, one might think that whether the data stored in the database is safe in terms of privacy or not.

Here comes the mainly significant concern, not only of the customer but of company as well. Keeping the private information of a customer safe is the foremost responsibility of any organizations failing which may lead them to trouble.

WHY DATA MINING IS REQUIRED IN RETAIL SECTOR

1. PROCURING AND ENGAGING CUSTOMER

It is harder to get novel clients than to hold current one [2]. After knowing, current customers purchasing habits, one can predict their respective interest and requirements for buying a specific product.

This sort of action encourages the retailer to hold existing clients by offering different plans [3].

2. MARKET BASKET ANALYSIS

Market basket analysis is a method in understanding what things are in high likelihood to be purchased together as indicated by association rule [4]. It gives slight idea about client's buying behavior by showcasing relations between varieties of purchased products.

Such sort of relation analysis helps in deciding the display of items, and promoting the combination of items, so that customers can find each item of their interest easily and this helps the organization in selling (a different product or service) to an existing customer..

3. CLIENTS SEGMENTATION AND TARGET ADVERTISING

Segmentation refers to partitioning the marketplace into various partitions on the basis of some characteristics. In order to form group or clusters on the basis of behavior, data mining can be used [5]. With the help of these clusters, customers with similar interests can be identified and simultaneously we can find customers for target marketing.

PRIVACY AND PPDM

PRIVACY DEFINED

Data protection alludes to the desire of people to be in charge of or have some control over information regarding them. Advancement in IT has hiked uncertainties about data safety and its consequences, and has encouraged Information Systems specialists to look into data safety issues, including specific replies for resolving various issues. [6]

Ways to maintain privacy of customer's data in retail

It is usually not possible that you want to protect customer data and also use it at the same time.

1. *Start a dedicated data safekeeping role within your organization* – this person entire movement ought to revolve in the region of information safety and ensuring protection of client data. They ought to be conversant in the fundamentals of information security and must be efficient on the majority of latest advancements.
2. *Make use of an intermediary service to provide external consulting and assistance* – Information defense organizations and external advisors can provide vital advice to allow you to review and address security issues that exists currently and that might come into sight shortly. They can likewise allow you to maintain a data safety plan with succeeding risk assessment. They will stay aimed in their evaluation of your safety conventions as they're not a part of your organization's way of life or law issues.
3. *Put into practice privacy preserving data mining techniques* – These security measures will let you to keep sensitive data safe at the same time as maintaining usability. In return, your data will be safe even as you analyze it to give you a tactical benefit in the market.

4. *Create a culture that highly prioritizes cyber security* – make employees and staff at all levels conscious that data protection is each and every person's responsibility, and that even one slight breach may lead to serious penalty for everyone within an organization.

The insights you put on by accumulating and analyzing customer data can give you added benefit in the retail market, but still you need to look after that data as well [7]

PRIVACY PRESERVING DATA MINING

PRIVACY PRESERVING DATA MINING TECHNIQUES

There is an immense growth in the investigation of data mining. Data mining is the strategy of extraction of information from gigantic warehouses. The hugest degree in research system is Privacy preserving data mining (PPDM). It is particularly essential to keep up an extent between maintaining privacy and information disclosure. The goal is to shroud personal data with the objective that the outsider can't extricate the real data from the database. To deal with such issues there are various algorithms established by various researchers across the globe.All together, those algorithms are termed as Privacy preserving data mining (PPDM) techniques.

Agencies need to alter values of sensitive data to maintain confidentiality and build trust.

More the data is altered, lower is the risk of disclosure.

DATA HIDING TECHNIQUES

1. Data perturbation

Strategies that try to achieve masking of individual private information while keeping up basic total connections of the records are referred as data perturbation techniques. These techniques amend genuine data figures to 'hide' exact secret entity record information. [8]

The main objective of data perturbation technique is to keep the customer's personal data safe like his/her buying habits, time of visits etc.

These techniques work by adopting either of the following methods

A. Noise inclusion

Noise inclusion methods modify secret attributes by including noise in order to achieve confidentiality. In this technique, a hypothetical or randomized number is added or multiplied to secret computable attributes. The hypothetical value is taken from a normal distribution having mean value as zero and a very negligible standard deviation. [9] [10]

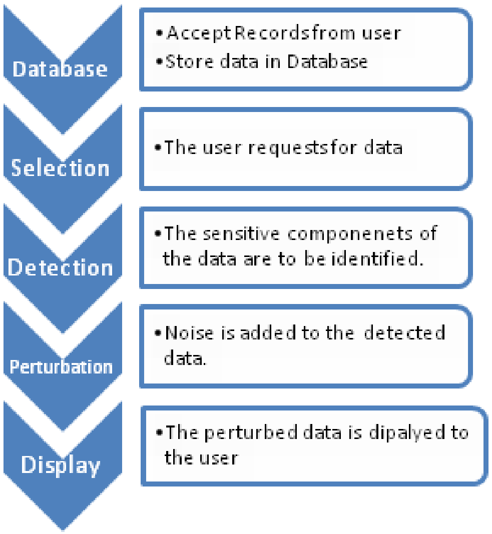
Database
•Accept Records from user
•Store data in Database
Selection
•The user requests for data
Detection
•The sensitive componenets of the data are to be identified.
Perturbation
•Noise is added to the detected data.
Display
•The perturbed data is dipalyed to the user

Figure 1: Noise addition procedure

B. Data swapping

Data swapping is a renowned and famous data perturbation technique. Data swapping can be defined as, the process of swapping of sensitive information among two persons by maintaining the sensitive information about the individuals [11]. In this method, actual individual records are changed with new values so that original dataset is entirely replaced so that the confidential attributes in a dataset are preserved. Through this method, data mining process achieved much accuracy when compared with existing noise addition methods with no breach in the privacy of the individuals.

The main reason for using data swapping technique is that it can be functional all along with additional privacy preserving data mining techniques , for example k-anonymity and randomization [12].

2. Cryptography

Cryptography is an extensively used method that is used for encrypting a plain text to get cipher (encrypted) text. Clear text or plaintext is defined as the data that is written by the user and can be easily examined and understood with no algorithm. The method of covering normal in order to mask its actual meaning is known as encryption. After applying encryption on a plain text, the random data which is generated is known as cipher text. Cryptography simply muddles data in order to achieve secrecy and/or accuracy of information and facilitates transmission of data among unsure networks so that it cannot be read by any mediator apart from the legal receiver. [13, 14].

3. Anonymization technique: Masking of personal identifiers

A. Suppression: In this strategy, a few estimations of the properties are supplanted by an asterisk '*'. Its may be possible, that we need to replace the entire data of the column. In the table below, all the data in the columns 'Name' and 'Religion' has been replaced by an asterisk.

B. Generalization: In this technique, specific values of attributes are changed with a more general type.

For example, the value '22' of the attribute 'Age' may be changed by ' ≤ 25', the value '33' by '30 < Age ≤ 40' , etc.

Following table depicts the anonymized database.

Name	Age	Gender	State	Religion	Disease
*	30 < Age ≤ 40	Female	Hyderabad	*	Cancer
*	30 < Age ≤ 40	Female	Bengaluru	*	Viral infection
*	30 < Age ≤ 40	Female	Hyderabad	*	TB
*	30 < Age ≤ 40	Male	Pune	*	No illness
*	30 < Age ≤ 40	Female	Bengaluru	*	Heart-related
*	30 < Age ≤ 40	Male	Pune	*	TB
*	Age ≤ 25	Male	Bengaluru	*	Cancer
*	30 < Age ≤ 40	Male	Pune	*	Heart-related
*	Age ≤ 25	Male	Bengaluru	*	Heart-related
*	Age ≤ 25	Male	Bengaluru	*	Viral infection

Table 1: Anonymized database

Data given in the above table has 2-anonymity with the attributes 'Age', 'Gender' and 'State' as for any grouping of these attributes establish in any line of the table there are for all time minimum 2 rows with those precise attributes. The attributes existing to an adversary are known as quasi-identifiers. Every quasi-identifier record exists in minimum k records for a dataset with k-anonymity. [15]

4. Condensation approach

Condensation approach makes use of a method which encapsulates the data into n-number of gatherings of pre-characterized size. For each group, a particular level of factual data about distinct records is retained. This statistical information is enough to protect statistical information about the mean and correlations among the distinct dimensions. Within a group, all similar records are maintained, which in turn, makes difficult to differentiate records from each other. Each group keeps up a base size s, which is known as the indistinguishability level. Indistinguishability level is directly proportional to the privacy level. Higher the Indistinguishability level, higher will be the privacy. Simultaneously, because of condensation n-number of records are grouped into a single statistical group entity and hence large amount of information is lost. [16]

CONCLUSION

Organizations and establishments continually gather information to offer or improve their current administrations. A large number of these administrations require the gathering, investigation and once in a while distributing/sharing of private delicate information. Data security is especially basic with pervasive data frameworks equipped for get-together information from a few sources, along these lines raising protection worries regarding the disclosure of information.

Privacy-Preserving Data Mining (PPDM) methods have been expected to allow the drawing out of knowledge from data at the same time as maintaining the privacy of persons. Nowadays, there is no only most constructive PPDM technique. The perfect decision is regularly only weighting the distinctive exchange offs among the perfect shelter level, the data misfortune, which is estimated by data efficacy measurements, the unpredictability & practical feasibility of the techniques.

The development of PPDM is aggravated by the security necessities of utilizations and fields/areas that manage data. Each application domain has its own assumptions, requirements and concerns with respect to privacy. All at once this heterogeneity opens a way for vast variety of algos. and methods, the fundamental approaches are generally intersecting.

REFERENCES

1. Prof. P. Balaram Babu, Prof. Prasanata Kumar Padhy, Dr.D.Rajeswara Rao, ―Data mining: A Source For Creative Decision Making‖, Asia Pacific Journal of Marketing & Management Review, Vol.1 No. 2, October 2012.
2. “The 6 biggest challenges retailer Face today", www.onstepretail .com, retrieved on June 2011
3. Jayanthi Ranjan, IMT Ghaziabad, “A Review of Data Mining Tools in Customer Relationship Management”, Journal of Knowledge Management Practice Vol 9, No 1, March 2008

4. Dhanabhakyam and Dr. M Punithavalli, “A Survey on Data mining Algorithm for Market Basket Analysis”: Global Journal of Computer Science and Technology” Vol-11 Issue –1 ver. 1.0 July 2011.ISSN: 0975-4350
5. Wedel, M. & Kamakura, W., 2000. Market segmentation Conceptual and Methodological Foundations 2nd ed., Kluwer Academic.
6. F Belanger, R E Crossler, Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems ,MIS Quarterly, volume 35, issue 4, p. 1017 – 1041
7. Travis Wright (2018, april 25). Top 4 ways retailers can protect customer data [Blog post]. Retrieved from https://www.protegrity.com/top-4-ways-retailers-can-protect-customer-data/
8. R. L. Wilson and P. A. Rosen. Protecting data through perturbation techniques: The impact on knowledge discovery in databases. Journal of Database Management (JDM), 14(2):14–26, 2003
9. Md Zahidul Islam, Privacy Preservation in Data Mining Through Noise Addition, phd Thesis, School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, New South Wales 2308, Australia, November 2007
10. Mohammad ali kadampur, somayajulu d.v.l.n., a noise addition scheme in decision tree for, privacy preserving data mining, journal of computing, volume 2, issue 1, January 2010, ISSN 2151-9617
11. Ienberg S., McIntyre J.: Data Swapping: Variations on a Theme by Dalenius and Reiss. Technical Report, National Institute of Statistical Sciences, 2003
12. Vaidya, Jaideep, basitshafiq, Wei Fan, Danish Mehmood, and David Lorenzi, "A random decision tree framework for privacy-preserving data mining", IEEE transactions on dependable and secure computing, no. 5, pp. 399-411, 2014.
13. Menezes, Alfred , Paul C van Oorschot ,Scott A. Vanstone, “ Handbook of Applied Cryptography. CRC Press”, October 1996 , ISBN 0-8493-8523-7.
14. William Stallings, “Cryptography and Network Security: Principles and practices”, Pearson education, Third Edition, ISBN 81-7808-902-5
15. *Narayanan, Arvind; Shmatikov, Vitaly.* "Robust De-anonymization of Large Sparse Datasets"
16. AGGARWAL, C. C. AND YU, P. S. 2004. A condensation approach to privacy preserving data mining. In Proceedings of the International Conference on Extending Database Technology (EDBT). 183– 199.