A SURVEY PAPER OF DIFFERENT TECHNIQUES FOR PRIVACY PRESERVING DATA MINING

Nidhi Joshi [1], Shakti V. Patel [2]

[1] *Computer Engineering, Computer, SPCE, Gujarat, India*

[2] *Computer Engineering, Computer, SPCE, Gujarat, India*

Abstract

Nowadays Data Mining has many privacy challenges when transforming data from database or data warehouse to the users. Because of publishing large amount of data everyday there is high risk of data loss and this losses of data sometimes create high risk for users for their sensitive data. Data mining comes with lots of techniques which are very necessary for privacy preserving. For preserve the privacy in data mining efficiency, Time, Cost, Accuracy are very necessary parameters. For obtain high privacy users have to compromise accuracy, time and cost. In this paper we studied many techniques in the direction of privacy preserving in data mining(PPDM) and after that describe the disadvantages of different techniques in privacy preserving in data mining.

Key Words :- *Data Mining, Privacy, PPDM Techniques*

1. Introduction

Data mining is a process of knowledge extraction from large data sets [1]. Data is passed through many phases during the life cycle of data management. There should be privacy is very necessary in each stage of life cycle. Data contains lots of sensitive information which are very necessary for users so there is privacy required.

In recent years privacy, security and data integrity are considered as challenging problem in data mining. Data mining is extensively used for knowledge discovery from large datasets. There are numbers of techniques and algorithms are available for this purpose. Privacy preserving is very necessary in secure multi party computation. Despite its benefit in a wide range of applications, data mining techniques also have raised a number of ethical issues. Some such issues include those of privacy, data security, and many others. Data mining incorporate privacy as a functional component for gain information and knowledge.

Clustering is widely used data mining techniques such as customer behavior analysis,targeted marketing and many others.Achiving privacy preservation when sharing data for clustering challenging problem.To address this problem, the system must not only meet privacy requirement of data owners but also gurantee valid clustering results [3].

1.2 Introduction about privacy:-

Privacy is an important concern while disclosing various categories of electronic data including business data and medical data for data mining. Especially for doing medical data mining the original data should be available for making accurate predictions otherwise lead to impractical solutions. Any kind of disclosure related to the person-specific information leads to many problems including ethical issues. Therefore extra care should be taken to protect privacy of individuals before publishing such data[3].

The privacy can be interpreted as preventing unwanted disclosure of information while performing data mining on aggregate results. Thus, privacy can be addressed at various levels in the process of data mining. For entire database security both privacy and security are required.

1.3 Objective Of Privacy:

The objective of privacy preserving data mining is to build algorithms for transforming and hide the original information in some way, so that the private data and private knowledge remain confidential even after the mining process. Privacy in data mining is very necessary in clustering process.

2. Related Work:-

There are various techniques and algorithms available in data mining for preserving the privacy. The hybrid approach includes combine two or more sanitization techniques and also various algorithm for PPDM. Privacy is an important issue when one wants to make use of data that involve individual sensitive information. As the increasing use of data mining, large volumes of personal data are regularly collected and analyzed so for that various techniques are available. Some of the technique may reduce the granularity and effectiveness of clustering Problem. Some PPDM techniques are given below[3]:-

K- Anonymity:-

When releasing micro data for research purposes, one needs to limit disclosure risks to an acceptable level while maximizing data utility. To limit disclosure risk, Sweeney introduced the *k*-anonymity privacy requirement, which requires each record in an anonymzed table to be indistinguishable with at least *k*-1 other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the *k*-anonymity requirement, they used both generalization and suppression for data anonymization.

Random Perturbation:-

It can deal with character type, Boolean type, classification type and number types of discrete data, and to facilitate conversion of data sets, it is necessary to preprocess the original data set. The data preprocessing is divided into discrete data, attribute coding, data sets coded data set.

Blocking Based Method:-

Blocking technique applies to applications where we can store unknown values for some attributes, when actual values are not available or confidential [1] .This method replaces the 1's or 0's by unknowns ("?") in selected transactions. So, that rule will not be generated from the dataset. The goal of the algorithm presented here is to obscure a given set of sensitive rule by replacing known values with unknown values. For each sensitive rule, it scans the original database and find outs the transactions supporting sensitive rules.

Cryptographic Technique:-

Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. Recent work has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation.This approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records.

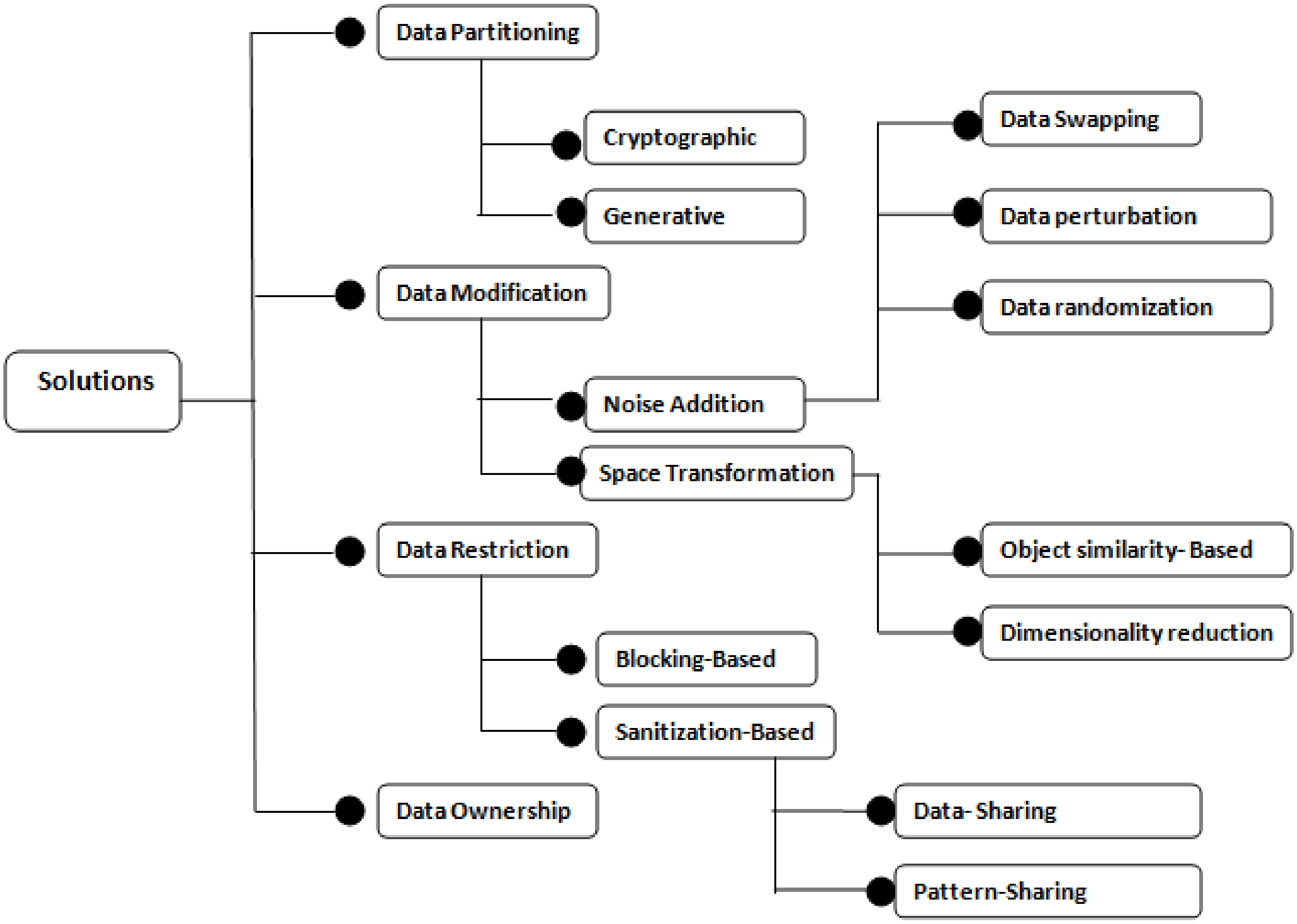

Fig-1 Taxonomy Of PPDM Techniques[4]

PPDM algorithms can further be divided according to privacy preservation techniques used.
Four techniques – sanitation, blocking, distort, and generalization -- have been used to hide data items for a centralized data distribution.

The idea behind data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined.
The blocking approach replaces certain attributes of the data with a question mark. Generalization transforms and replaces each record value with a corresponding generalized value.

3. Techniques comparison:-

Sr. No.	Techniques	Parameter used	Advantages	Disadvantages
1	K-Anonymity	Quasi-identifier attribute	It limits disclosure Risk	Homogenity and Background Attack
2	Random Perturbation	Discrete formula A(max)-A(min)/n A-continuous attribute n-no. of discrete length	Reconstruct the original distribution	Does not reconstruct original data values.
3	Blocking based method	1's or 0's	Easy to implement	It gives information loss
4	Cryptographic Techniques	Different Keys	It offers well defined models for privacy and gives many tools for that	It is difficult to scale when two or more parties are involved
5	K-means Clustering		It limiting communication cost	Very slow Less effective
6	Generative models	"mean model"	It gives loss communication cost.	This approach achieves high quality distributed clustering with acceptable privacy loss
7	Data Swapping	Decision Tree	Swapping is performed over confidential attributes only, where confidential attribute is a class label.	This approach deal with trade off: statistical precision against security level

4. Research Gap:-

Nowadays K-Means clustering algorithm is popularly used for clustering process in PPDM.K-Means algorithm is centroid based algorithm which creates problem for clustering process,it also sometime reduce cluster efficiency and also take more time for clustering process. By observing all above technique there may be information loss, homogeneity attack and also creates linkage attack. So need to develop better clustering process for preserving the privacy.

5. Conclusion:-

Privacy preserving data mining has the potential to increase the reach and benefits of data mining technology. However, we must be able to justify that privacy is preserved. For this, we need to be able to communicate what we mean by “privacy preserving”. From the above techniques it is clear that present technologies have lots of advantages as well as disadvantages which creates high risk for PPDM. While creating cluster for attributes there is more accuracy is neededao better clustering algorithm is required.

6. References:-

[1]. Natarajan, R., et al. "A survey on Privacy Preserving Data Mining."International Journal of Advanced Research in Computer and Communication Engineering 1.1 (2012).

[2]. Lohiya,Savita, and Lata Ragha. "Privacy Preserving in Data Mining Using Hybrid Approach." Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on. IEEE, 2012.

[3]. Kalita, M., D. K. Bhattacharyya, and M. Dutta. "Privacy Preserving Clustering-A Hybrid Approach." Advanced Computing and Communications, 2008. ADCOM 2008. 16th International Conference on. IEEE, 2008.

[4]. Chidambaram, S., and K. G. Srinivasagan. "A combined random noise perturbation approach for multi level privacy preservation in data mining."Recent Trends in Information Technology (ICRTIT), 2014 International Conference on. IEEE, 2014.

[5]. R.Hemalatha, M.Elamparithi,”Privacy Preserving Data Mining Using Sanitizing Algorithm”, (IJCSIT) International Journal of Computer Science and Information Technologies, 2015.

[6]. Wang, Xiao-Feng, and De-Shuang Huang. "A novel density-based clustering framework by using level set method." Knowledge and Data Engineering, IEEE Transactions on 21.11 (2009): 1515-1531.)

[7]. Abeysekara, Ruvan Kumara, and Weishi Zhang. "Hybrid framework for privacy preserving data sharing." Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on. IEEE, 2013.

8) Malik, Mohammad Bilal, M. Asger Ghazi, and Raian Ali. "Privacy preserving data mining techniques: current scenario and future prospects." Computer and Communication Technology (ICCCT), 2012 Third International Conference on. IEEE, 2012.