ISSN:0975-9646

Survey on Privacy Preserving Data Mining

Krishna Pratap Rao[1] ,Adesh Chaudhary[2,] Prashant johri[3]

School of Computing Science and Engineering,
Galgotias University

Abstract: - Data mining is the process of extraction of data from large database. One of the most important topics in research community is Privacy preserving data mining (PPDM). Privacy preserving data mining has become increasingly popular because it allows sharing of privacy sensitive data for analysis purposes. It is essential to maintain a ratio between privacy protection and knowledge discovery. To solve such problems there are some algorithms presented by various authors worldwide. The primary goal of this survey paper is to understand the existing privacy preserving data mining techniques and to achieve efficiency.

Keywords: - Privacy Preserving, Utility Mining, Sanitization, Data Mining, randomization, distributed privacy preserving.

1. INTRODUCTION:-

There are two main approaches of previous work in privacy preserving data mining. Perturbing the data values for preservation of customer privacy is the first approach. Cryptographic tools to build data mining models are the other approach. The following section contains some of the recent researches in this field.

As the data mining deals with generation of association rules, the change in support and confidence of the association rule for hiding sensitive rules is done. A new concept named 'not altering the support' is proposed to hide an association rule. The support of sensitive item not being changed is the first characteristic of proposed algorithm. The position of the sensitive item is the only thing which changes. The efficiency of proposed algorithm is the second characteristic we analyzed. The reduction of the confidence of the sensitive rules without change in the support of the sensitive item is the approach for modifying the database transaction. This is in contrast to this existing algorithm, which either decreases or increases the support of the sensitive item to modify the database transactions.

One of the way of promotional business growth among the organization is information sharing. Intimidation of data sharing is majorly caused by recent trends in data mining. Balancing the privacy of the data as per the legitimate need of the user is the major problem. The original data is modified by the sanitization process to conceal sensitive knowledge before release so the problem can be addressed. Privacy preservation of sensitive knowledge is addressed by several researchers in the form of association rules by suppressing the frequent item sets.

2. 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.

A. Hiding Association Rules by Using Confidence and Support: -

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.

Advantages: -

- First advantage of proposed algorithm is that support for the sensitive item is unchanged. Instead, only the position of the sensitive item set is changed.
- It provide the use of a different technique for modifying the database transaction . to reduce the confidence of sensitive rules with out making any changement in the sensitive item.

Disadvantage: -

One of the main disadvantages of the existing approaches is that the approach in tries to hide every single rule from a given set of rules without checking if some of the rules could be pruned after modification of some transactions from the set of all transactions. This approach hides rules having sensitive items either in the right side or in the left side.

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. Stanley R. M. Oliveira, and Osmar R. Zaiane [1] 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.

Advantages: -

- The geometric data transformation methods (GDTMs) that distort confidential numerical attributes in order to meet 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.

Disadvantage: -

- The protection of the data values to clustering similarity between objects under analysis is hard to achieve
- One major disadvantage is that the individual's data is shared for clustering is very complex**.**

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 [2] 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.[3] We will discuss heuristic based algorithms. 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. [4] 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.

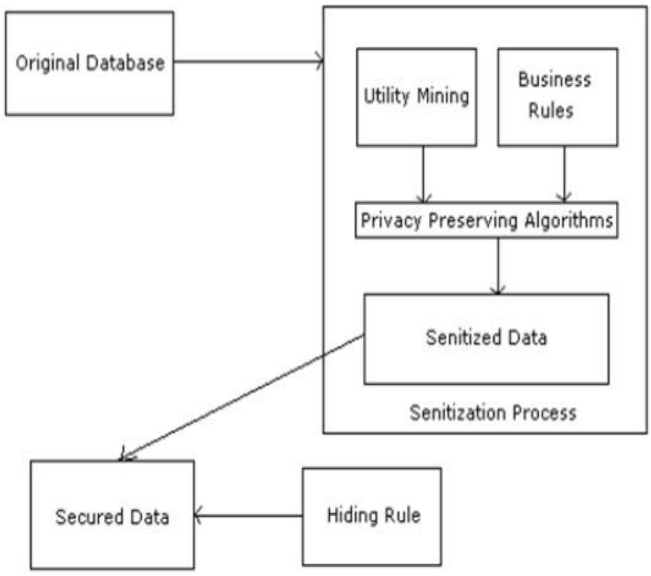

Fig. 1 **:** Block diagram of Privacy Preserving Data Mining Technique

3. CONCLUSIONS AND FUTURE WORK: -

In this paper, we carries out a wide survey of the different approaches for privacy preserving data mining, and analyses the major algorithms available for each method and points out the existing drawback. While all the purposed methods are only approximate to our goal of privacy preservation. To address this issue, we advise that 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. We are tried to develop more efficient algorithms and achieve a balance between disclosure cost, computation cost and communication cost.

REFERENCES: -

[1] 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.

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

[3] 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.

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

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