# Role and Importance of Association Mining For Preserving Data

Mr. **Mayank Chavda**
Department of Computer Science
*Assistant Professor*
*Sarvoday College of Mgt. and Technology, Limbdi. Gujarat*

**Dr. MijalkumarMistry**
Department of Computer Science
*Assistant Professor*
*ISTAR, VallabhVidhyanagar, Gujarat*

## Abstract

In Universe data is in form of basic raw material, which is used to develop any system. One has to gather and process the data for getting the information. In today's scenario one of the most existing problems in Data Mining is the process of discovering frequent data and consequently Association Rules. Data can be in form of text, audio, image files, and video. Discovering hidden dataitems(patterns) from complex and big data plays an important role in Marketing,Business,Medical Analysis, and other domain areas where these patterns are useful for strategic decision making. The concept of privacy preserving data mining has recently been proposed in response to the concerns of preserving privacy information from data mining algorithms. Proposed review paper described two types of method in which the first type of privacy, called output privacy, inwhich data is altered so that the mining result will preserve certain privacy. The second type of privacy called input privacy, inwhich the data is manipulated.

**Keywords: Data Mining, Security, Algorithms, Privacy Preservation, Association Mining.**

## I. INTRODUCTION

Association rule mining scrutinized valuable associations and established a correlation relationship between large set of data items[1]. Association rules describe attribute value conditions that occur frequently together in a given data sheet. A typical and widely used example of association rule mining is Market Basket Analysis[2][3][4]. For example, data are collected using bar-code scanners in supermarkets. Such market basket databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Manager would be interested to know if certain groups of items are consistently purchased together. They could use this data for adjusting store layouts, for cross-selling, for promotions, for catalog design and to identify customer segments based on buying patterns. People are scared to provide personal information while using websites as they feel the organization would misuse the information. To increase the confidence of the user, a system called Mining Associations with Secrecy Konstraints(MASK) was developed, where the information can be distorted at the user end using a simple probabilistic distribution instead of any third party or the organization doing the same. [5] They show that the efficiency of the Privacy Preserving Data Mining can be well with an order or magnitude with respect to data mining by maintaining a satisfactory level of privacy and accuracy.

## II. LITERATURE REVIEW

The sensitive association rule hiding problem is very common in a collaborative association rule mining project, in which one company may decide to disclose only part of knowledge contained in its data and hide strategic knowledge represented by sensitive rules. These sensitive rules must be protected before its data shared. Besides, by hiding some association rules, data owners can prevent the rule-based vicious inferences used for unwarrantable purposes,e.g. uncovering private data, as discussed in [6].

Verykios et al.[7] provide a survey of the existing privacy preserving data mining techniques. They classify the techniques based on the following dimensions: data distribution, data modification, data mining algorithm, data or rule hiding and privacy preservation. They define a parameter "transversal endurance" which is used to evaluate the sanitization algorithms designed for various privacy preserving techniques in different databases.

Verykios et al.[8] provide two approaches:

(1) Hiding the frequent sets to prevent the rules from being generated and
(2) Reducing the importance of the rules by keeping the confidence below a threshold value.

They provide five algorithms that are built on these two approaches. These strategies or algorithms perform minimal perturbation on data values in the data set.

## III. ANALYSIS OF EXISTING PROBLEM

On the basis of published paper described that from Non-sensitive information or unclassified data, one is able to infer sensitive information, including personal information, facts, or even patterns that are not suppose to be disclosed.

## IV. PROBLEM STATEMENT

Formally, the problem has inputs a database D and a privacy confidence threshold. Let R(D,s) be the set of rules with confidence s in D. It is given B $\subseteq$ R as the set of sensitive rules that must be hidden based on some privacy policy. The task is to lower the confidence of the rules in B below s and keep the impact on the non-sensitive rules A=R(D,s)\B at a minimum. The goal is to transform D into a database D' so that the most association rules in R can still be mined from D' while others, representing sensitive rules, will be hidden. In this case, D' becomes the released database.

In a previous classification of PPDM techniques,Oliveria et al.[9] classified the existing sanitizing algorithms into two major classes

(1) Data sharing techniques
(2) Pattern sharing techniques

In the previous classification, PPDM problems are classified based on the techniques used to protect sensitive data. When the classification is based on a privacy-preserving technique, this is called as "classification by what". Classification by what can be divided into two distinct categories. First, hiding data or showing it exactly. Solutions that fall into this category are: limiting access, augment the data, swapping, and auditing. Usually, the approaches under this category have less privacy but better accuracy in terms of results. Second, perturbing the data which means changing attributes values with new values. This can be accomplished by adding noise, or replacing selected values with a question mark. Approaches under this category have greater privacy but less accuracy in terms of results[1][3].

## V. PROPOSED WORK

In the new classification, PPDM problems are classified based on "classification by where". New classifications is general, comprehensive and gives better understanding to the field of PPDM in terms of placing each problem in the right category. The new classification is as follows: PPDM can be attempted at three levels as shown in Figure 1.1. The first level is raw data or databases where transactions reside. The second level is data mining algorithms and techniques that ensure privacy. The third level is the output of different data mining algorithms and techniques.

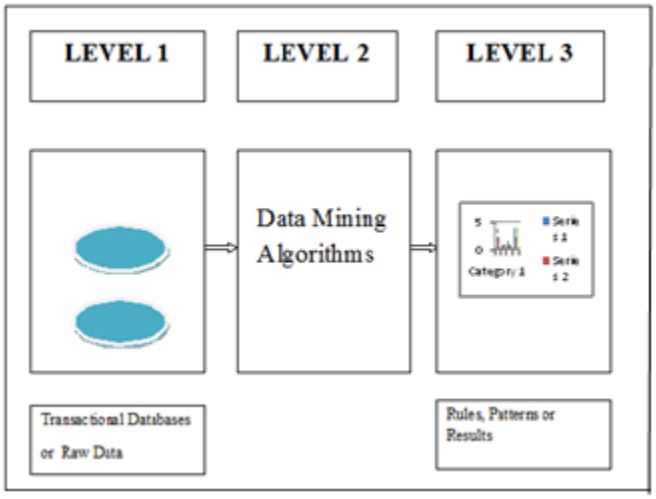


Fig. 1.1: Levels of Privacy Preserving

Finally, the review can be concluded that most of the research in privacy preserving association rule mining; the ideas of selecting the sensitive rules that need to be hidden are not given. The sensitive patterns (items) too identified are not given. In large amount of database manual selection of the sensitive items and sensitive rules is difficult. The proposed work will be done under level 1and the algorithm is introduced to identify the sensitive items which then give the sensitive rules for hiding.

## VI. EVALUATING PPDDM TECHNIQUES

At present, the privacy Preserving Distributed Data Mining study is in development stage. Then most current PPDDM techniques are on the theory level and are developed for specific application against some certain aspects. Therefore so far, therefore there is no technique to effectively achieve the PPDDM goals. So the evaluation framework recommended for assessing and evaluating PPDDM techniques, is in accordance with the following criteria:

- Efficiency:It is defined based on techniques running time (computational cost) and cost of information exchange between sites (communication cost).
- Privacy Level: PPDDM a computation is called secure if the information obtained by any party can be obtained through only its own input and output.
- Mining Accuracy:It is defined based on amount of data mining result accuracy that achieved in PPDDM techniques.
- Scalability: scalability of the technique refers to the ability to efficient handle many participant sites, when the number of participant site increases.
- Security model: is defined based on assumptions of site's behavior that is considered in techniques.
- Applicable areas: is defined based on appropriate distributed areas that these techniques are applicable.

## VII. CONCLUSION

In this paper a new taxonomy of PPDM techniques is described. PPDM problems in association rule mining is discussed. Classification of privacy preserving and the state of art in privacy is described. A technique is introduced to identify the sensitive rules and experiments show their effectiveness in identifying sensitive items.

## REFERENCES

[1] R.Agarwal, T. Imielinski, and A.swami. Mining association rules between sets of items in large databases. In proceedings of the ACM SIGMOD Conference on Management of Data, pages 207-216, New York, NY, USA, May 1993. ACM Press.
[2] M.Dunham. Data Mining: Introductory and Advanced Topics (book). Prentice Hall,1st edition,2003.
[3] A. K Pujari Data Mining Techniques (book) University Press (India) limited, 2001.
[4] R.Roiger and M.Geatz. Data Mining: A Tutorial Based Primer (book). Addidion-wesiey, 2003.
[5] S.Agrawal, V.Krishnan and J.R. Harista. On addressing Efficinecy Concerns in privacy-preserving Mining.In Proceedings of the 9th International Conference on Database Systems for Advanced Applications (DASFAA-2004).Jeju Island, Korea March-2004.
[6] S. Fienberg and A.slavkovic, preserving the confidentiality of categorical statistical data bases when releasing information for association rules. Data Mining and Knowledge Discovery, 11(2): 155-180, 2005.
[7] V.S.Verykios,E.Bertino,I.N.Fovino,L.p.provenza,Y.saygin,Y.Theodoridis. State-of-the-art in privacy preserving Data Mining. In SIGMOD Record, 33(1): 50-57 March 2004.
[8] V.S.Verykios, A.K Elmagarmid, E.Bertino, Y.Saygin, E.Dasseni, Association Rule Hiding. IEEE Transactions on Knowledge and Data Engineering volume: 16, Issue: 4. April 2004. pp.434-447.
[9] S.R.M. Oliveria, O.R Zaiane, and Y.Saygin. Secure association rule sharing. In Proceedings of the 8th PAKDD Conference, volume 3056, pages 74-85, Sydney,Australia,May 2004.