International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 10, October-2015

# Techniques for Privacy Preservation in Data Mining

Mrs. Suchitra Shelke
ME Computer
PHCET, Rasayani
Dist Raigad, India

Prof. Babita Bhagat
Asst. Professor, Computer Department
PHCET, Rasayani
Dist Raigad, India

***Abstract*— Data Mining deals with automatic extraction of previously unknown patterns from large amounts of data sets. These data sets typically contain sensitive individual information or critical business information, which consequently get exposed to the other parties during Data Mining activities. This creates barrier in Data Mining process. Solution to this problem is provided by Privacy preserving in data mining (PPDM). PPDM is a specialized set of Data Mining activities where techniques are evolved to protect privacy of the data, so that the knowledge discovery process can be carried out without barrier. The objective of PPDM is to protect sensitive information from leaking in the mining process along with accurate Data Mining results. The goal of this paper is to present the review on different privacy preserving techniques which are helpful in mining large amount of data with reasonable efficiency and security.**

***Keywords*— *Data mining, Privacy Preservation, PPDM.***

## I. INTRODUCTION

In modern days organizations are extremely dependent on Data Mining results to provide better service, achieving greater profit, and better decision-making. For these purposes organizations collect huge amount of data. This data includes sensitive data about Individuals or organizations. While running Data Mining algorithm against such data, the algorithm not only extracts the knowledge but it also reveals the information which is considered to be private. The real threat is that once information gets exposed to unauthorized party, it will be impractical to stop misuse. Privacy can for instance be threatened when Data Mining techniques uses the identifiers which themselves are not very sensitive, but are used to connect personal identifiers such as addresses, names etc., with other more sensitive personal information. Privacy is very important for trusted collaboration and interactions. Because of these privacy and data security concerns in data mining, the data owner hesitates while sharing data for data mining activities. And this creates barrier in data mining task. Privacy preserving data mining technique gives new direction to solve this problem.

Privacy preserving in data mining (PPDM) is a new area of research in Data Mining process. Its ultimate goal is to allow one to extract relevant knowledge from large amount of data and provide accurate data mining result, while prevent sensitive information from disclosure or inference. In PPDM, new techniques are invented to provide privacy for the knowledge discovered in Data Mining. It also takes care that knowledge discovery process should not be banned because of privacy reason. Figure I shows the framework for PPDM process.

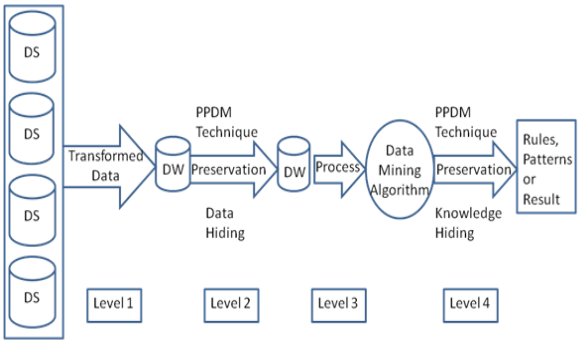


Figure I: Framework of PPDM

In level 1, Data from different sources is grouped together and preprocessed. This preprocessed data is stored in to the data warehouse. The same data which is stored in the Data warehouse is used for Data Mining. In level 2, data hiding techniques are applied to provide privacy for the sensitive data. Different data hiding techniques are applied in order for the users not to compromise with privacy of the other user's data. In level 3, Data Mining algorithms are used to find patterns and discover knowledge from the historical data. After mining, in level 4 privacy preservation techniques are applied on data mining results to protect it from unauthorized access.

## II. CLASSIFICATION SCHEME OF PPDM TECHNIQUES

The PPDM techniques can be classified based on following 3 characteristics: data distribution, purposes of hiding, Data Mining algorithms [1].

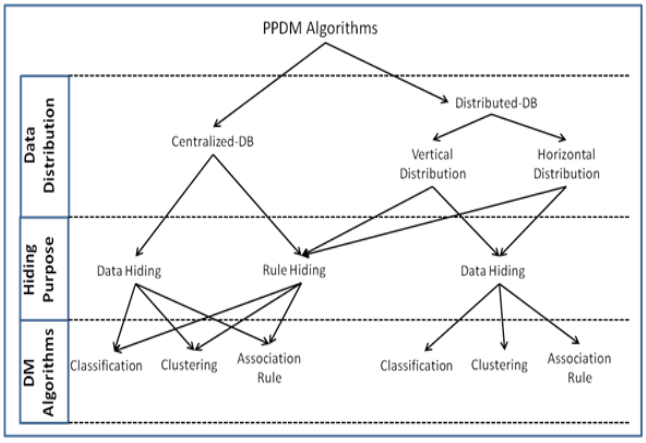


Figure II: Classification schema of PPDM techniques

Under Data distribution characteristic PPDM algorithms are broadly classified for Central Database and Distributed Database. Centralized Database was owned by the private party where as Distributed Database was owned by more than one parties which are interested to perform Data Mining on joint Data. In case of Distributed Database, the distribution may be vertical or horizontal. According to the Hiding purpose, PPDM algorithms are further classified in to Data hiding and Rule Hiding. Data hiding refers to hiding sensitive data from the individuals, where as rule hiding refers to hiding knowledge derived after applying Data Mining algorithms. Currently majority of the PPDM algorithms uses association rule method for mining data, followed by classification, and then clustering.

## III. TECHNIQUES OF PPDM

Techniques of PPDM can divided into following categories:

### A. *Data Hiding Techniques*

In this technique input data provided for data mining task is altered, trimmed, or blocked in such a way that sensitive information present in that will not be exposed to other parties. There are different ways of implementing these techniques which are explained in detail in section VI.

### B. *Knowledge Hiding Techniques*

In this technique sensitive knowledge extracted by data mining algorithm is excluded from use. These techniques are very important because the sensitive knowledge extracted by data mining process can be used to derive confidential information. There are different ways of implementing these techniques which are explained in detail in section V.

### C. *Hybrid Technique*

Each techniques mentioned above have some advantages and disadvantages. None of the technique is perfect. In hybrid technique, an effort has been made to combine any 2 techniques mentioned above in order to get one prefect technique. Some of the hybrid techniques are mentioned in the section VI.

## IV. DATA HIDING TECHNIQUES

In these techniques sensitive raw data is altered, blocked, or trimmed out from the original database, so that the users of the data will not be able to compromise another person's privacy. Following diagram shows the different approaches used for hiding the data.

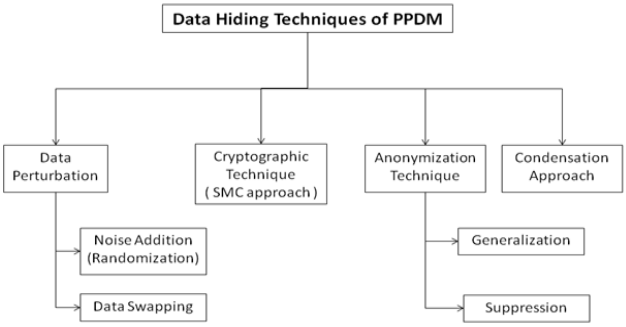


Figure III: Data Hiding Techniques of PPDM

### A. *Data Perturbation*

In this technique available data is modified before it is passed to Data Mining. There are number of ways to modify the data like swapping, adding noise etc. but after modification quality of the released data is maintained.

In case of noise addition technique, Data owner add some random number (noise) to input data. This random number is generally drawn from a normal distribution with zero mean and a small standard deviation, which preserves statistics of original data. Then data owner share this noisy data for Data Mining task. Data owner also shares distribution of the noise added to the original data. By using this distribution and noisy data, data miner reconstruct original data set's distribution. But data miner cannot retrieve actual data values. This enabled a Data Mining algorithm to construct a much more accurate result without revealing actual data [2].

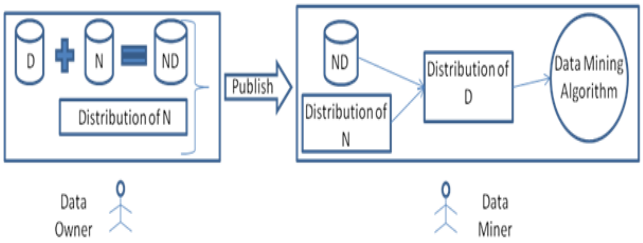


Figure IV: Noise Addition Technique

In case of data swapping technique, it keeps all original value in the data set, while at the same time makes the record re-identification very complex. Data swapping interchange the attribute values between different records. Similar attribute values are interchanged with higher probability. The unique feature of this approach is all original values are kept back within the data set and only the positions are swapped. Records are exchanged in such a way that low-order frequency counts are maintained [3].

### B. *Cryptographic Technique*

In distributed data mining, privacy can be achieved by using cryptographic and secure multiparty computation (SMC) techniques. The basic idea of SMC is that parties hold their own private data and they cooperate in computation to get the final result, but at the same time ensures that no more information is revealed to a participant in the computation other than participant's own input and final output. There are different SMC techniques for different type of computation. Secure Sum is one SMC technique which is used parties to find sum of their local values securely. Secure Set union is another example of SMC technique which is used by parties to securely compute union of all private items owned by parties, without revealing the owner of the item.

Secure set sum and secure set union methods can be used for securely mining association rule on horizontally distributed data. An association rule is an implication of the form, X =>Y. The rule X =>Y holds in the transaction set D with confidence c if c percent of transactions in D containing X also contain Y. The rule X => Y has support s in the transaction set D if s percent of transactions in D contain X < Y. In association rule mining, we need to find all association rules in D with support s > st and confidence c > ct, where st

and ct are user-defined thresholds for "interesting" rules. By combining Secure Set Union and Secure Sum method this association rule mining can be carried out securely on horizontally distributed data [4].

Each party Pi can find all possible association rules and local confidence, support for that rules using its local data. The rules having support and confidence greater than threshold (ct and st) will be added to the local set of association rule, say LRi (for ith party). Then all sites can give their local rule set as an input to secure set union algorithm and participate in the algorithms to find global set of association rule say GR. After getting GR, each party can calculate local confidence and local support for each rule in global set GR. Next step is to find global support and global confidence for each rule in GR. For this secure sum method can be used. One rule from GR can be selected at a time and for that rule each party can give their local support and confidence as input to secure sum method. The method returns global support and global confidence for selected rule [5]. Association rule having global confidence and support greater than threshold (ct and st) can be added in final output set.

### C. *Anonymization Technique*

Anonymization means removing identifying information from the original data to protect personal or private information. There are many ways for performing data anonymization basically this method uses k-anonymization approach. If each row in the table cannot be distinguished from at least other k-1 rows by only looking a set of attributes, then this table is K-anonymized on these attributes. Example: While trying to identify a person from a table the only input information given is person's birth date and gender, then there will be k people meeting the requirement.

There are two anonymization methods, suppression-based and generalization-based [6]. In suppression-based anonymization method, subsets of original data records are formed by masking the values of some well-chosen attributes. This masking is done by using some special values like *. And in generalization-based anonymization method, subsets of original data records are formed by replacing original values with more general ones in the database, like replacing value present in age column ( 33, 37 ) with range ( 30 – 40 ).

### D. *Condensation Technique*

Condensation approach compresses and packs the raw input data into multiple groups or clusters. Each group or cluster has constraint which is defined for it in terms of its size. This size is referred as the level of that privacy-preserving approach. Greater is the level, the greater will be the amount of privacy. This size is chosen in a way so as to preserve k-anonymity. After condensing data into clusters, statistics of data in each group is analyzed and maintained separately for each cluster. This statistics from each cluster is used further to generate pseudo data for corresponding clusters. In the process of data mining, data owner publish this pseudo data instead of original data. Various data mining tasks use this pseudo data as input. In this way actual data remains hidden from other parties [7].

This technique is referred as condensation because of its approach of using condensed statistics of the clusters in order to generate pseudo-data. In this approach, a greater amount of information is lost because of the condensation of a larger number of records into a single statistical group entity. Following figure shows steps in followed in the condensation approach.

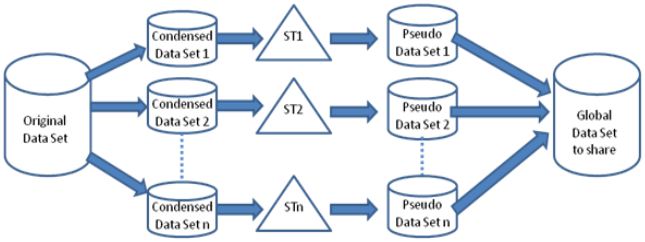


Figure V: Condensation Approach

## V. KNOWLEDGE HIDING TECHNIQUES

In this approach, sensitive knowledge extracted from the Data Mining process is excluded for use. These techniques are as important as data hiding techniques because knowledge extracted in data mining process can be used to derive confidential information. This problem is also commonly called the "database inference problem". Following are different ways used for hiding knowledge.

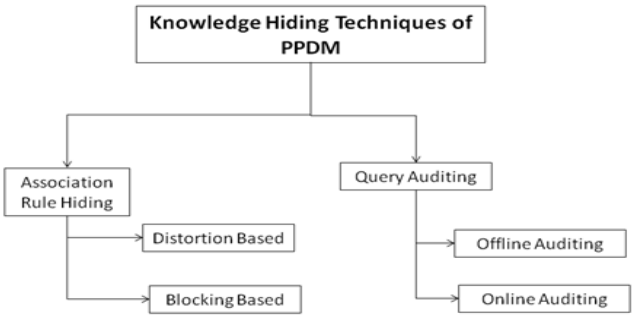


Figure VI: Different Knowledge Hiding Techniques

### A. *Association Rule Hiding*

Association rule hiding refers to the process of modifying the original database in such a way that certain sensitive association rules disappear without seriously affecting the data and the non-sensitive rules. The primary goal is to protect access to sensitive information that can be obtained through non-sensitive data and inference rules. Here it is considered that it is not the data but the sensitive rules that create a breach to privacy. Association rule mining algorithms scan the database of transactions and calculate the support and confidence of the rules and retrieve only those rules having support and confidence higher than the user specified minimum support and confidence threshold. So for hiding sensitive rules, the support or confidence of the sensitive association rules need to be decreased below the threshold value [8].

There are different approaches developed to decrease the support of association rule for hiding association rule. Distortion is one such method. Data distortion is done by the alteration of attribute value by a new value. To decrease support of an item, the system will modify one item at a time by changing from 1 to 0 or from 0 to 1 in a selected transaction. There are 2 algorithms namely ISL (Increase Support of Left hand side) and DSR (Decrease Support of

Right hand side) to hide useful association rule from transactions data using distortion scheme. In ISL method, confidence of a rule is decreased by increasing the support value of Left Hand Side (LHS) of the rule. For this purpose, only the items from LHS of a rule are chosen for modification. For example if rule is X=>Y, increases the support of X, the LHS of the rule. In DSR method, confidence of a rule is decreased by decreasing the support value of Right Hand Side (R.H.S.) of a rule. For this purpose, only the items from R.H.S. of a rule are chosen for modification. For example if rule is X=>Y, increases the support of Y, the RHS of the rule.

Another method for association rule hiding is blocking. This technique inserts unknown values in the data to fuzzify the rules. In some applications where publishing wrong data is not acceptable, then unknown values may be inserted to blur the rules. When unknown values are inserted, support and confidence values would fall into a range instead of a fixed value and this in hiding sensitive association rules from final result set.

### *B. Query Auditing*

Auditing is the process of examining past actions to check whether they were in conformance with official policies. In the context of database systems with specific data disclosure policies, auditing is the process of examining queries that were answered in the past to determine whether answers to these queries could have been used by an individual to ascertain confidential information forbidden by the disclosure policies.

There are 2 variants of query auditing, offline and online auditing [9]. Offline auditors examine answers provided to all the queries in past to find if answers disclosed any sensitive information. This auditing can also be referred as retroactive auditing. On the other hand online auditing is used for detecting disclosures and could potentially also be used or extended to prevent disclosures. Given a sequence of queries, Q1, . . . , Qt-1 that have already been posed, corresponding answers A1, . . . , At-1 that have already been supplied, and a new query Qt, the task of an online auditor is to determine if the new query should be answered as such, or denied in order to prevent a privacy breach.

## VI. HYBRID TECHNIQUES

Privacy preservation is a very huge field. Many techniques have been proposed in this filed in order to secure the data. But there is no any single technique that is consistent in all domains. Each technique has some limitations and disadvantages. All methods perform in a different way depending on the type of data as well as the type of application or domain. Along with disadvantages, each technique has some advantages. In order to overcome limitations of the different PPDM techniques, two or more techniques can be merged together. This new approach is called as hybrid technique. Many algorithms have been proposed to combine 2 or more techniques.

The randomization and generalization techniques can be combined in hybrid technique. In this approach first randomization is applied on the raw data and then modified or randomized data is generalized. This technique protects private data with better accuracy; also it can reconstruct original data and provide data with no information loss. Murat Kantarcioglu and Chris Clifton proposed a hybrid technique to combine noise addition and SMC for securely mining of association rules over horizontally partitioned data [4]. In this technique while sharing encrypted rule set to other parties, a little noise is added in the rule in form of false rules. Also in [6], AES encryption technique is combined with Anonymization to provide higher level of security.

## VII. CONCLUSION

Data mining is very important tool used by organizations for providing better service, achieving greater profit, and better decision-making. But privacy and security concerns may create barrier in data mining task. These barriers can be removed by applying PPDM techniques and by ensuring security in data mining task. Many techniques have been proposed for PPDM, with each technique having some advantages over another in its own terms.

It is important to note that there are some key points which are not addressed in current research of PPDM. First is, there is no standard terminology used for PPDM. Second, most of algorithms are developed for centralized database. However, in today's global digital environment, data is often stored in different sites. Third, many algorithms concentrate on protecting privacy of individual information, but don't pay attention on security of sensitive information included in data mining result. There is no single method which can achieve data as well as rule hiding. Fourth, each algorithm concentrates only for on type data mining task. There is no single method present which can work for all type of data mining task. These all points can be used as a direction for future research in PPDM.

## REFERENCES

[1] K. Srinivasa Rao & B. Srinivasa Rao, 2013 ,"An Insight in to Privacy Preserving Data Mining Methods".

[2] Charu C. Aggarwal and Philip S. Yu, 2008, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms".

[3] Manish Sharma and Atul Chaudhary, 2013 ," A Review Study on the Privacy Preserving Data Mining Techniques and Approaches".

[4] The IEEE computer society, 2004, "Privacy-Preserving Data Mining: Why, How, and When".

[5] Murat Kantarcioglu and Chris Clifton, Senior Member, 2004 , "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data".

[6] Mahesh Dhande, N.A.Nemade and Yogesh Kolhe, 2013, "Privacy Preserving in K- Anonymization Databases Using AES Technique".

[7] Gayatri Nayak 2011, "A Survey on Privacy Preserving Data Mining: Approaches and Techniques".

[8] Vassilios S. Verykios and Aris Gkoulalas-Divanis, 2008 , "A Survey of Association Rule Hiding Methods for Privacy".

[9] Shubha U. Nabar, Krishnaram Kenthapadi and Nina Mishra, 2008, "A Survey of Query Auditing Techniques for Data Privacy".