Privacy Preservation Techniques in Data Mining: A Review

Priyanka Nalawade [1], Dhananjay Shivaji Gawali[2]
[1] Dr. D Y Patil School of Engineering & Technology, Pune
[2] New Art Commerce & Science College, Shevgaon,India

Abstract- *Data mining tools aims to find useful patterns from large amount of data. These patterns represent information and are conveyed in decision trees, clusters or association rules. The knowledge discovered by various data mining techniques may contain private information about people or business. Preservation of privacy is a significant aspect of data mining and thus study of achieving some data mining goals without losing the privacy of the individuals' .The analysis of privacy preserving data mining (PPDM) algorithms should consider the effects of these algorithms in mining the results as well as in preserving privacy. Within the constraints of privacy, several methods have been proposed but still this branch of research is in its formative years .The success of privacy preserving data mining algorithms is measured in terms of its performance, data utility, level of uncertainty or resistance to data mining algorithms etc. However no privacy preserving algorithm exists that outperforms all others on all possible criteria. Rather, an algorithm may perform better than another on one specific criterion. So, the aim of this paper is to present current scenario of privacy preserving data mining framework and techniques.*

Keywords- Anonymization, Condensation, Cryptography, Distributed Data Mining, Perturbation, Privacy Preserving Data Mining (PPDM), Randomized Response.

I. INTRODUCTION

Data mining is one of the core processes in knowledge discovery of databases [1]. Data mining research deals with the extraction of potentially useful information from large collections of data with a variety of application areas such as customer relationship management, market basket analysis. The mined information can be a patterns, rules, clusters or classification models. During the whole process of data mining (from gathering of data to discovery of knowledge) these data, which typically contain sensitive individual information such as medical and Financial information, often get exposed to several parties including collectors, owners, users and miners. The huge amount of data available means that it is possible to learn a lot of information about individuals from public data. Privacy preserving has originated as an important concern with reference to the success of the data mining. Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. People have become well aware of the privacy intrusions on their personal data and are very unwilling to share their sensitive information. In current years, the area of privacy has realized fast advances because of the increases in the ability to store data. In particular, recent advances in the data mining field have lead about privacy [2].The aim of privacy preserving data mining(PPDM) algorithms is to mined appropriate information from huge amounts of data while protecting at the same time thoughtful information.

The main goals a PPDM algorithm is:

1. A PPDM algorithm should have to thwart the discovery of sensible information.
2. It should be resistant to the various data mining techniques.
3. It should not compromise the access and the use of non-sensitive data.
4. It should not have an exponential computational complexity.

Many secure protocols have been proposed so far for data mining and machine learning techniques for decision tree classification, clustering, association rule mining, Neural Networks, Bayesian Networks. The main concern of these algorithms is to preserve the privacy of parties' sensitive data, while they gain useful knowledge from the whole dataset. One of the most studied problems in data mining is the process of discovering frequent item sets and, consequently association rules. Association rule mining are usually used in various area.

Most of the privacy-preserving data mining techniques apply a transformation which reduces the usefulness of the underlying data when it is applied to data mining techniques or algorithms. Privacy concerns can avoid building of centralized warehouse – in scattered among several places, no one are allowed to transfer their data to other place. In preserving privacy of data, the problem is how securely results are gained but not with data mining result but. As a simple example, suppose some hospitals want to get useful aggregated knowledge about a specific diagnosis from their patients' records while each hospital is not allowed, due to the privacy acts, to disclose individuals' private data. Therefore, they need

to run a joint and secure protocol on their distributed database to reach to the desired information.

In many cases data is distributed, and fetching the data collected in one place for analysis is not possible due these privacy acts or rules. Mining association rules requires iterative scanning of database, which is quite costly in processing. These techniques can be demonstrated in centralize as well as distributed environment [3, 4] where data can be distributed among the different sites. Distributed database scenario can be classified in horizontally partitioned data and vertically partitioned data.

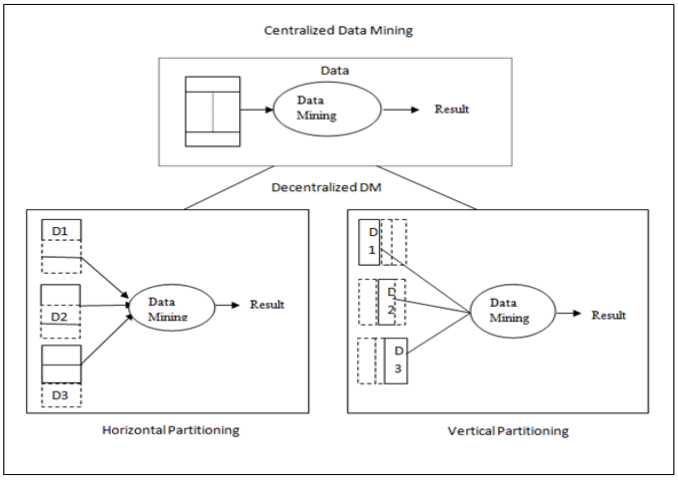

Fig.1 Distributed Database Scenario

1. Horizontally partitioned data: It divides database into a number of non-overlapping horizontal partitions. In this scenario different places have different record about same entities or people which are used for mining purposes. Many of these methods use specialized versions of the general approaches discussed for various problems.

2. Vertically partitioned data: In Vertically partitioned data sets; each site has different number of attributes with same number of transaction. The approach of vertically partitioned mining has been extended to a variation of data mining applications such as decision trees, SVM Classification, Naïve Bayes Classifier, and k-means clustering [5].

II. PPDM FRAMEWORK

The framework for PPDM is shown in fig.2. In data mining or knowledge discovery from databases (KDD) process the data (mostly transactional) is collected by single/various organization/s and stored at respective databases. Then, it is transformed to a format suitable for analytical purposes, stored in large data warehouse/s and then data mining algorithms are applied on it for the generation of information/knowledge [6]. The first level is raw data or databases where transactions exist in. The second level is data mining algorithms and techniques that ensure privacy. The third level is the output of different data mining algorithms and methods [2].

At level 1, the raw data collected from a single or multiple databases or even data marts is transformed into a format that is well suited for analytical purposes. Even at this stage, privacy concerns are needed to be taken care of. Researchers have applied different techniques at this stage but most of them deal with making the raw data suitable for analysis [6].

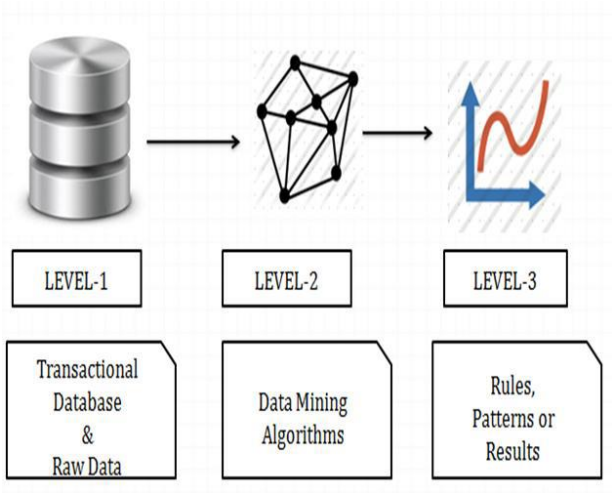

Fig.2 PPDM Framework

At level 2, the data from data warehouses is subjected to various processes that make the data sanitized so that it can be revealed even to untrustworthy data miners. The processes applied at this stage are blocking, suppression, perturbation, modification, generalization, sampling etc. Then, the data mining algorithms are applied to the processed data for knowledge/information discovery. Even the data mining algorithms are modified for the purpose of protecting privacy without sacrificing the goals of data mining [6].

At level 3, the information/knowledge so revealed by the data mining algorithms is checked for its sensitiveness towards disclosure risks. We have described the embedding of privacy concerns at three levels, but any combination of these may be used [6].

III. PPDM TECHNIQUES

For Recent years have witnessed extensive research in the field of PPDM. As a research direction in data mining and statistical databases, privacy preserving data mining received substantial attention and many researchers performed a good number of studies in the area. Since its inception in 2000 with the pioneering work of Agrawal & Srikant [7] and Lindell &

Pinkas [8], privacy preserving data mining has gained increasing popularity in data mining research community. PPDM has become an important issue in data mining research [10-11]. As a outcome, a whole new set of approaches were presented to allow mining of data, while at the same time leaving out the releasing any secretive and sensitive information. The majority of the existing approaches can be classified into two broad categories [9]:

(i) Methodologies that protect the sensitive data itself in the mining process, and
(ii) Methodologies that protect the sensitive data mining results (i.e. extracted knowledge) that were produced by the application of the data mining.

The first category refers to the methodologies that apply perturbation, sampling, generalization or suppression, transformation, etc. techniques to the original datasets in order to generate their sanitized counterparts that can be safely disclosed to untrustworthy parties. The goal of this category of approaches is to enable the data miner to get accurate data mining results when it is not provided with the real data. Secure Multiparty Computation methodologies that have been proposed to enable a number of data holders to collectively mine their data without having to reveal their datasets to each other.

The second category deals with techniques that prohibits the disclosure sensitive knowledge patterns derived through the application of data mining algorithms as well as techniques for downgrading the effectiveness of classifiers in classification tasks, such that they do not reveal sensitive information. In difference to the centralized model, the Distributed Data Mining (DDM) model accepts that the individual's information is distributed across multiple places. Algorithms are developed within this area for the problem of efficiently receiving the mining results from all the data through these distributed sources. A simple method to data mining over multiple sources that will not share data is to run existing data mining tools at each place independently and combine the results [12].

However, this will often fail to give globally valid output. Issues that cause a difference between local and global results include:

(i) Values for a single entity may be divided across sources. Data mining at individual sites will be unable to detect cross-site correlations.
(ii) The same item may be duplicated at different sites, and will be over-biased in the results.
(iii)At a single site, it is likely to be from a similar population.
PPDM tends to transform the original data so that the result of data mining task should not defy privacy constraints. Following is the list of five dimensions on the basis of which different PPDM Techniques can be classified [13]:

i. Data distribution
ii. Data modification
iii. Data mining algorithms
iv. Data or rule hiding
v. Privacy preservation

Data or Rule Hiding: This dimension refers to whether raw data or grouped data should be hidden. Data hiding means protecting sensitive data values, e.g. names, social security numbers etc. of some people. And Rule hiding means Protecting Confidential Knowledge in data, e.g. association rule. The difficulty for hiding aggregated data in the form of rules is very difficult, and for this purpose, typically heuristics have been developed.

Data Distribution: This dimension refers to the distribution of data. There are some of the approaches are developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can be divided as horizontal data partition and vertical data partition. Horizontal distribution refers to these cases where different sets of records exist in different places, while vertical data distribution refers where all the values for different attributes reside in different places.

Data Modification: Data modification is used with the aim of change the unique values of a database that wants to be allowed to the public and in this way to guarantee high privacy protection. Methods of data modification include:
i. **Perturbation:** which is able to replacing attribute value by a new value (changing a 1-value to a 0-value, or adding noise)
ii. **Blocking:** which is the replacement of an existing attribute value with a "?"
iii. **Swapping:** This refers to interchanging values of individual record.
iv. **Sampling:** This refers to losing data for only sample of a population.
v. **Encryption:** many Cryptographic techniques are used for encryption.

Data Mining Algorithm: The data mining algorithm for which the privacy preservation technique is designed:
1. Classification data mining algorithm
2. Association Rule mining algorithms
3. Clustering algorithm.

Privacy Preserving Techniques:

1. Heuristic-based techniques: It is an adaptive modification that modifies only selected values that minimize the effectiveness loss rather than all available values.

2. Cryptography-based techniques: This technique includes secure multiparty computation where a computation is secure if at the completion of the computation, no one can know anything except its own input and the results. Cryptography-based algorithms are considered for protective privacy in a distributed situation by using encryption techniques.

3. Reconstruction-based techniques: where the original distribution of the data is reassembled from the randomized data.
Based on these dimensions, different PPDM techniques may be classified into following five categories [13-15, 21, 22].

1. Anonymization based PPDM
2. Perturbation based PPDM
3. Randomized Response based PPDM
4. Condensation approach based PPDM
5. Cryptography based PPDM

We discuss these in detail in the following subsections.

3.1 Anonymization based PPDM

The basic form of the data in a table consists of following four types of attributes:
(i) Explicit Identifiers is a set of attributes containing information that identifies a record owner explicitly such as name, SS number etc.
(ii) Quasi Identifiers is a set of attributes that could potentially identify a record owner when combined with publicly available data.
(iii) Sensitive Attributes is a set of attributes that contains sensitive person specific information such as disease, salary etc.
(iv)Non-Sensitive Attributes is a set of attributes that creates no problem if revealed even to untrustworthy parties.

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden [23, 24]. It even assumes that sensitive data should be retained for analysis. It's obvious that explicit identifiers should be removed but still there is a danger of privacy intrusion when quasi identifiers are linked to publicly available data. Such attacks are called as linking attacks. For example attributes such as DOB, Sex, Race, and Zip are available in public records such as voter list[23,24].

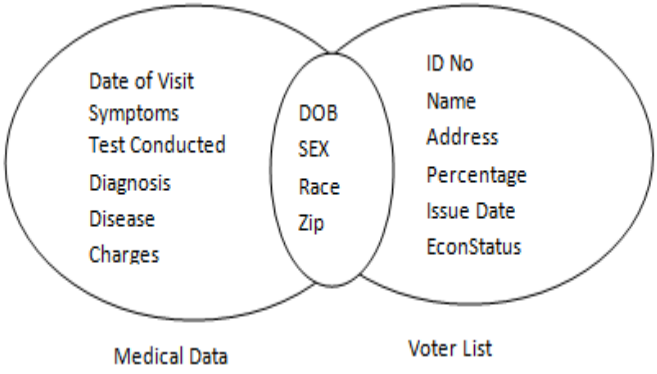

Fig.3 Linking Attack

Such records are available in medical records also, when linked, can be used to infer the identity of the corresponding individual with high probability as shown in fig.3.

Sensitive data in medical record is disease or even medication prescribed. The quasi-identifiers like DOB, Sex, Race, Zip etc. are available in medical records and also in voter list that is publicly available. The explicit identifiers like Name, SS number etc. have been removed from the medical records. Still, identity of individual can be predicted with higher probability. Sweeney [16] proposed k-anonymity model using generalization and suppression to achieve k-anonymity i.e. any individual is distinguishable from at least k-1 other ones with respect to quasi-identifier attribute in the anonymized dataset. In other words, we can outline a table as k-anonymous if the Q1 values of each raw are equivalent to those of at least k- 1 other rows. Replacing a value with less specific but semantically consistent value is called as generalization and suppression involves blocking the values. Releasing such data for mining reduces the risk of identification when combined with publically available data. But, at the same time, accuracy of the applications on the transformed data is reduced. A number of algorithms have been proposed to implement k-anonymity using generalization and suppression in recent years.

Although the anonymization method ensures that the transformed data is true but suffers heavy information loss. Moreover it is not immune to homogeneity attack and background knowledge attack practically [14]. Limitations of the k-anonymity model stem from the two conventions[23]. First, it may be very tough for the owner of a database to decide which of the attributes are available or which are not available in external tables. The second limitation is that the k-anonymity model adopts a certain method of attack, while in real situations; there is no reason why the attacker should not try other methods. However, as a research direction, k-anonymity in combination with other privacy preserving methods needs to

be investigated for detecting and even blocking k-anonymity violations.

3.2 Perturbation Based PPDM

Perturbation being used in statistical disclosure control as it has an intrinsic property of simplicity, efficiency and ability to reserve statistical information. In perturbation the original values are changed with some synthetic data values so that the statistical information computed from the perturbed data does not differ from the statistical information computed from the original data to a larger extent. The perturbed data records do not agree to real-world record holders, so the attacker cannot perform the thoughtful linkages or recover sensitive knowledge from the available data. Perturbation can be done by using additive noise or data swapping or synthetic data generation.

In the perturbation approach any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. Relevant information for data mining algorithms such as classification remains hidden in inter-attribute correlations. This is because the perturbation approach treats different attributes independently. Hence the distribution based data mining algorithms have an intrinsic disadvantage of loss of hidden information available in multidimensional records. Another branch of privacy preserving data mining that manages the disadvantages of perturbation approach is cryptographic techniques.

3.3 Randomized Response Based PPDM

Basically, randomized response is statistical technique introduced by Warner to solve a survey problem. In Randomized response, the data is twisted in such a way that the central place cannot say with chances better than a pre-defined threshold, whether the data from a customer contains correct information or incorrect information. The information received by each single user is twisted and if the number of users is large, the aggregate information of these users can be estimated with good quantity of accuracy. This is very valuable for decision-tree classification. It is based on combined values of a dataset, somewhat individual data items. The data collection process in randomization method is carried out using two steps [14]. During first step, the data providers randomize their data and transfer the randomized data to the data receiver. In second step, the data receiver rebuilds the original distribution of the data by using a distribution reconstruction algorithm. The randomization response model is shown in fig.4.

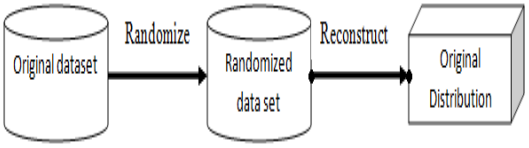

Fig.4 Randomization Response Model

Randomization method is relatively very simple and does not require knowledge of the distribution of other records in the data. Hence, the randomization method can be implemented at data collection time. It does not require a trusted server to contain the entire original records in order to perform the anonymization process [2]. The weakness of a randomization response based PPDM technique is that it treats all the records equal irrespective of their local density. These indicate to a problem where the outlier records become more subject to oppositional attacks as compared to records in more compressed regions in the data [8]. One key to this is to be uselessly adding noise to all the records in the data. But, it reduces the utility of the data for mining purposes as the reconstructed distribution may not yield results in conformity of the purpose of data mining.

3.4 Condensation approach based PPDM

Condensation approach constructs constrained clusters in dataset and then generates pseudo data from the statistics of these clusters [19]. It is called as condensation because of its approach of using condensed statistics of the clusters to generate pseudo data. It creates sets of dissimilar size from the data, such that it is sure that each record lies in a set whose size is at least alike to its anonymity level. Advanced, pseudo data are generated from each set so as to create a synthetic data set with the same aggregate distribution as the unique data. This approach can be effectively used for the classification problem. The use of pseudo-data provides an additional layer of protection, as it becomes difficult to perform adversarial attacks on synthetic data.

Moreover, the aggregate behavior of the data is preserved, making it useful for a variety of data mining problems [2]. This method helps in better privacy preservation as compared to other techniques as it uses pseudo data rather than modified data. Moreover, it works even without redesigning data mining algorithms since the pseudo data has the same format as that of the original data. It is very effective in case of data stream problems where the data is highly dynamic. At the same time, data mining results get affected as huge amount of information is released because of the

compression of a larger number of records into a single statistical group entity [14].

3.5 Cryptography Based PPDM

Consider a scenario where multiple medical institutions wish to conduct a joint research for some mutual benefits without revealing unnecessary information. In this scenario, research regarding symptoms, diagnosis and medication based on various parameters is to be conducted and at the same time privacy of the individuals is to be protected. Such scenarios are referred to as distributed computing scenarios [17].The parties involved in mining of such tasks can be mutual un-trusted parties, competitors; therefore protecting privacy becomes a major concern. Cryptographic techniques are ideally meant for such scenarios where multiple parties collaborate to compute results or share non sensitive mining results and thereby avoiding disclosure of sensitive information. Cryptographic techniques find its utility in such scenarios because of two reasons: First, it offers a well-defined model for privacy that includes methods for proving and quantifying it. Second a vast set of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms are available in this domain. The data may be distributed among different collaborators vertically or horizontally.

All these methods are almost based on a special encryption protocol known as Secure Multiparty Computation (SMC) technology. SMC used in distributed privacy preserving data mining consists of a set of secure sub protocols that are used in horizontally and vertically partitioned data: secure sum, secure set union, secure size of intersection and scalar product. Although cryptographic techniques ensure that the transformed data is exact and secure but this approach fails to deliver when more than a few parties are involved. Moreover, the data mining results may breach the privacy of individual records. There exist a good number of solutions in case of semi-honest models but in case of malicious models very less studies have been made. Table 1. Contains advantages and limitations of PPDM techniques.

IV. EVALUATION CRITERIA OF PRIVACY PRESERVING ALGORITHM

Privacy preserving data mining an important characteristic in the development and evaluation of algorithms is the identification of suitable evaluation criteria and the development of related principles. In some case, there is no privacy preserving algorithm exists that beats the other entire algorithm on all possible measures. Relatively, an algorithm may perform better that another one on specific measures, like performance and/or data utility. It is vital to deliver users with a set of metrics which will allow them to select the best suitable privacy preserving technique for the data; with respect to some specific parameters [13].

An introductory list of evaluation parameters to be used for evaluating the quality of privacy preserving data mining algorithms is given below: [13]

(i) Performance: the performance of a mining algorithm is measured in terms of the time required to achieve the privacy criteria.

(ii) Data Utility: Data utility is basically a measure of information loss or loss in the functionality of data in providing the results, which could be generated in the absence of PPDM algorithms.

Table 1. Advantages and Limitations of PPDM Techniques

Technique	Advantages	Limitations
Anonymization based PPDM	Anonymization based PPDM	Anonymization based PPDM
Identity or sensitive data about record owners are to be hidden.	Identity or sensitive data about record owners are to be hidden.	Identity or sensitive data about record owners are to be hidden.
Randomized Response based PPDM	Randomized Response based PPDM	Randomized Response based PPDM
It is relatively simple useful for hiding information about individuals.	It is relatively simple useful for hiding information about individuals.	It is relatively simple useful for hiding information about individuals.
Better efficiency compare to cryptography based PPDM technique [20].	Better efficiency compare to cryptography based PPDM technique [20].	Better efficiency compare to cryptography based PPDM technique [20].

(iii) Uncertainty level: It is a measure of uncertainty with which the sensitive information that has been hidden can still be predicted.

(iv)Resistance: Resistance is a measure of tolerance shown by PPDM algorithm against various data mining algorithms and models.

As such, all the criteria that have been discussed above need to be quantified for better evaluation of privacy preserving

algorithms but, two very important criteria are quantification of privacy and information loss. Quantification of privacy or privacy metric is a measure that indicates how closely the original value of an attribute can be estimated. If it can be estimated with higher confidence, the privacy is low and vice versa. Lack of precision in estimating the original dataset is known as information loss which can lead to the failure of the purpose of data mining. So, a balance needs to be achieved between privacy and information loss. Dakshi Agrawal and Charu Agrawal in [18] have discussed quantification of both privacy and information loss in detail.

V. CONCLUSION

The main objective of privacy preserving data mining is developing algorithm to hide or provide privacy to certain sensitive information so that they cannot be disclosed to unauthorized parties or intruder. Although a Privacy and accuracy in case of data mining is a pair of ambiguity. Succeeding one can lead to adverse effect on other. In this, we made an effort to review a good number of existing PPDM techniques. Finally, we conclude there does not exists a single privacy preserving data mining algorithm that outperforms all other algorithms on all possible criteria like performance, utility, cost, complexity, tolerance against data mining algorithms etc. Different algorithm may perform better than another on one particular criterion.

REFERENCES

[1] J. Han, M. Kamber. “Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers.

[2] Charu C. Aggarwal, Philip S. Yu “Privacy-Preserving Data Mining Models and algorithm”advances in database systems 2008 Springer Science, Business Media, LLC.

[3] Vaidya, J. & Clifton, C. W, "Privacy preserving association rule mining in vertically partitioned data," In Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton, Canada 2002.

[4] Murat Kantarcioglou and Chris Clifton, "Privacy-preserving distributed mi ni ng of association rules on horizontally partitioned data," In Proceedings of the ACM SI GMOD Workshop on Research Isuues in Data Mining and Knowledge Discovery, pp 24-31, 2002.

[5] Ahmed K. Elmagarmid, Amit P. Sheth “Privacy-Preserving Data Mining Models and algorithm” advances in database systems 2008.

[6] Ahmed HajYasien. Thesis on “PRESERVING PRIVACY IN ASSOCIATION RULE MINING” in the Faculty of Engineering and Information Technology Griffith University June 2007.

[7] R. Agrawal and R. Srikant. “Privacy Preserving Data Mining”,ACM SIGMOD Conference on Management of Data, pp: 439-450,2000.

[8] Y. Lindell and B. Pinkas, “Privacy Preserving Data Mining”, Journal of Cryptology, 15(3), pp.36-54, 2000.

[9] Aris Gkoulalas-Divanis and Vassilios S. Verikios, “An Overview of Privacy Preserving Data Mining”, Published by The ACM Student Magazine, 2010.

[10] Stanley, R. M. O. and R. Z Osmar, “Towards Standardization in Privacy Preserving Data Mining”, Published in Proceedings of 3rd Workshop on Data Mining Standards, WDMS’ 2004, USA, p.7-17.

[11] Elisa, B., N.F. Igor and P.P. Loredana. “A Framework for Evaluating Privacy Preserving Data Mining Algorithms”, Published by Data Mining Knowledge Discovery, 2005, pp.121-154.

[12] Andreas Prodromidis, Philip Chan, and Salvatore Stolfo, : “Metalearning in distributed data mining systems: Issues and approaches”. In “Advances in Distributed and Parallel Knowledge Discovery”, AAAI/MIT Press, September 2000.

[13] S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004, “State of the Art in Privacy Preserving Data Mining” Published in SIGMOD Record, 33, 2004, pp: 50-57.

[14] Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", ternational Journal of Engineering Science and Technology, Vol. 3 No. 3, 2127-2133, 2011.

[15] Wang P, "Survey on Privacy preserving data mining", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, 2010.

[16] Sweeney L, "Achieving k-Anonymity privacy protection uses generalization and suppression" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571-588, 2002.

[17] Benny Pinkas, "Cryptographic Techniques for Privacy preserving data mining", SIGKDD Explorations, Vol. 4, Issue 2, 12-19, 2002.

[18] D. Agrawal and C. Aggarwal, “On the Design and Quantification of Privacy Preserving Data Mining Algorithms”, PODS 2001. pp: 247-255.

[19] Aggarwal C, Philip S Yu, "A condensation approach to privacy preserving data mining", EDBT, 183-199, 2004.

[20] Helger Lipmaa,” Cryptographic Techniques in Privacy-Preserving Data Mining”, University College London, Estonian Tutorial 2007.

[21] Dharmendra Thakur and Prof. Hitesh Gupta,” An Exemplary Study of Privacy Preserving Association Rule

Mining Techniques", P.C.S.T., BHOPAL C.S Dept, P.C.S.T., BHOPAL India, International Journal of Advanced Research in Computer Science and Software Engineering ,vol.3 issue 11,2013.

[22] C.V.Nithya and A.Jeyasree,"Privacy Preserving Using Direct and Indirect Discrimination Rule Method", Vivekanandha College of Technology for WomenNamakkal India, International Journal of Advanced Research in Computer Science and Software Engineering ,vol.3 issue 12,2013.

[23] Priyanka Gawali and Dhananjay Gawali "Big data privacy preservation using K-Anonymization and l-Diversity" IJSART - Volume 2 Issue 11 –NOVEMBER 2016 ISSN [ONLINE]: 2395-1052

[24] Priyanka Shivaji Gawali and Prof. Arti Mohanpurkar "Scalable Privacy Preservation in Big Data-A Review" International Journal of Computer Applications (0975 – 8887) Volume No 120, June 2015