International Journal of Computer Applications Technology and Research
Volume 3-Issue7,403-408, July 2014

A Review on "Privacy Preservation Data Mining (PPDM)

Dwipen Laskar
Assistant Professor, Dept. of CSE
Girijananda Chowdhury Institute of Management & Technology,
Guwahati, Assam, India
laskardwipen@gmail.cm

Geetachri Lachit
Assistant Professor, Dept. of MCA
Girijananda Chowdhury Institute of Management & Technology,
Guwahati, Assam,India
lachit.geetashri@gmail.cm

Abstract: It is often highly valuable for organizations to have their data analyzed by external agents. Data mining is a technique to analyze and extract useful information from large data sets. In the era of information society, sharing and publishing data has been a common practice for their wealth of opportunities. However, the process of data collection and data distribution may lead to disclosure of their privacy. Privacy is necessary to conceal private information before it is shared, exchanged or published. The privacy-preserving data mining (PPDM) has thus has received a significant amount of attention in the research literature in the recent years. Various methods have been proposed to achieve the expected goal. In this paper we have given a brief discussion on different dimensions of classification of privacy preservation techniques. We have also discussed different privacy preservation techniques and their advantages and disadvantages. We also discuss some of the popular data mining algorithms like association rule mining, clustering, decision tree, Bayesian network etc. used to privacy preservation technique.. We also presented few related works in this field.

Keywords: perturbation data mining, bayesian network, privacy preservation, association rule mining, clustering

1. INTRODUCTION

data mining aims to extract useful information from multiple sources, whereas privacy preservation in data mining aims to preserve these data against disclosure or loss. Privacy preserving data mining (PPDM) [1,2] is a novel research direction in data mining and statistical databases [3], where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main consideration of the privacy preserving data mining is two-fold. First, sensitive raw data like identifiers, name, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy. The main objective of privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and the private knowledge remain private even after the mining process. In this paper, we provide a classification and description of the various techniques and methodologies that have been developed in the area of privacy preserving data mining. Agarwal and Srikant [3] and Lindell and Pinkas [4] introduced the first Privacy-preserving data mining algorithms which allow parties to collaborate in the extraction of knowledge, without any party having to reveal individual items or data. The goal of this paper is to give an review of the different dimension and classification of privacy preservation techniques used in privacy preserving data mining. Also aim is to give different data mining algorithms used in PPDM and related research in this field.

2. DIMENSIONS OF PRIVACY PRESERVATION DATA MINING

Different techniques are used in privacy preserving data mining. They can be classified based on the following six dimensions [5]: *Data Mining Scenario, Data Mining Tasks, Data Distribution, Data Types, Privacy Definition, Protection Method.*

The first dimension refers to the different data mining scenarios used in privacy preservation. They are basically of two major classes presently used. In the first one organization release their data sets for data mining and allowing unrestricted access to it. Data modification is used to achieve the privacy in this scenario. In the second one organization do not release their data sets but still allow data mining tasks. Cryptographic techniques are basically used for privacy preserving

The second dimension refers to the different data mining tasks due to the data set containing various patterns. Different types of data mining tasks used are like classification, association rule mining, outlier analysis, and clustering and evolution analysis [6]. The basic need of a privacy preservation technique is to maintain data quality to support all possible data mining tasks and statistical analysis.

The third dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to the cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places.

The fourth dimension refers to different types of data types which are basically of three classes: Numerical and Categorical and Boolean. Boolean data are the special case of categorical data which takes two possible values 0 and 1. Numerical data has a natural ordering inherent to them but which is lacking in Categorical data. This is the most basic difference between categorical and numerical values which forces the privacy preservation technique to take different approaches for them.

The fifth dimension refers to the different definitions of privacy in different context. The definition of privacy is context dependant. In some scenario individuals data values are private, whereas in other scenario certain group, association or classification rules are private. They are basically of two classes: *Individual privacy preservation* and *Collective privacy preservation* [7]. The primary goal of Individual privacy preservation is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. The goal of the Collective privacy preservation is to protect against learning sensitive knowledge representing the activities of a group. Depend on the privacy definition we work on different privacy preserving techniques.

The sixth dimension refers to different Protection Methods: Privacy in data mining is protected through different methods such as *data modification* and *secure multiparty computation* (SMC). In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include: *data perturbation, Data swapping, Aggregation, and Suppression.*

Data perturbation, which refers to a data transformation process typically performed by the data owners before publishing their data. The goal of performing such data transformation is two-fold. On one hand, the data owners want to change the data in a certain way in order to disguise the sensitive information contained in the published datasets, and on the other hand, the data owners want the transformation to best preserve those domain-specific data properties that are critical for building meaningful data mining models, thus maintaining mining task specific data utility of the published datasets. The major challenge of data perturbation is balancing privacy protection and data quality, which are normally considered as a pair of contradictive factors Two types of data Perturbation are available: *Additive Perturbation* and *matrix multiplicative Perturbation* [3][7].

In *Additive Perturbation is* a technique for in which noise is added to the data in order to mask the attribute values of records [4][7]. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions.

In the *matrix multiplicative perturbations* [7] can also be used to good effect for privacy-preserving data mining. The data owner replaces the original data X with $Y = MX$ where M is an $n' \times n$ matrix chosen to have certain useful properties. If M is orthogonal ($n' = n$n and $M^T M = I$), then the perturbation exactly preserves Euclidean distances, *i.e.*, for any columns x_1 ,x_2 in X ,their corresponding columns y_1 ,y_2 in Y satisfy $x_1 - x_2 = y_1 - y_2$. If each entry of M is generated independently from the same distribution with mean zero and variance σ 2 (n' not necessarily equal to n), then the perturbation approximately preserves Euclidean distances on expectation up to constant factor $2\sigma\ n'$. If M is the product of a discrete cosine transformation matrix and a truncated perturbation matrix, then the perturbation approximately preserves Euclidean distances.

In *data swapping* techniques, the values across different records are interchanged in order to perform privacy preserving in data mining. One advantage of this technique is that the lower order marginal totals of the data are completely preserved and are not perturbed at all. Therefore certain kinds of aggregate computation can be exactly performed without violating the privacy of the data [8].

In *Suppression* technique sensitive data value are removed or suppressed before published. Suppression is used to protect an individual privacy from intruders attempt to accurately predict a suppressed value. Information loss is an important issue in suppression by minimizing the number of values suppressed [9][10].

Aggregation is also known as generalization or global recording. It is used for protecting an individual privacy in a released data set by perturbing the original data set before its releasing. Aggregation change k no. of records of a data by representative records. The value of an attribute in such a representative record is generally derived by taking the average of all values, for the attributes, belonging to the records that are replaced. Another method of aggregation or generalization is transformation of attribute values. For ex- an exact birth date can be changed by the year of birth. Such a generalization makes an attribute value less informatics. For ex- if exact birth date is changed by the century of birth then the released data can became useless to data miners [11].

3. DIFFERENT TECHNIQUES OF PEIVACY PRESERVING DATA MINING

Different types of privacy preservation techniques are used. They are mainly classified into following categories: [31] *Anonymization based*, *Randomized Response based*, *Condensation approach based*, *Perturbation based*, *Cryptography based* and *Elliptic Curve Cryptographic based*.

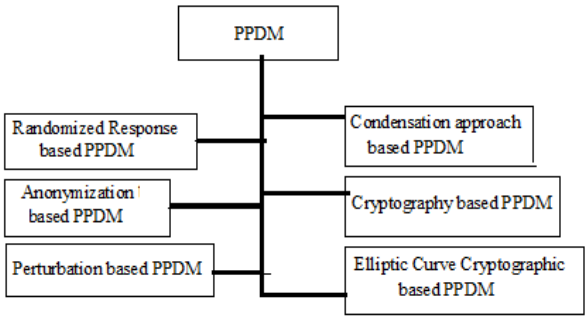


Fig 1: Techniques of PPDM

3.1 Anonymization Based PPDM

Anonymization based PPDM: Anonymization method aim is to make the individual record be indistinguishable among a group of records with using techniques of generalization and suppression [12][13][31]. Replacing a value with less specific but semantically consistent value is called as generalization and suppression involves blocking the values. K-anonymity is used to represent anonymization method. The anonymization method is ensured that after getting transformation data is true but there is some information loss in some extent. A database is *k*-anonymous with respect to quasi-identifier attributes (a set of attributes that can be used with certain external

information to identify a specific individual) if there exist at least *k* transactions in the database having the same values according to the quasi-identifier attributes

E_ID	Name	Age	Disease
101	X	45	Cancer
112	Y	43	Cancer
123	Z	44	Fever

Fig: 2 (a) Original Data

E_ID	Name	Age	Disease
1**	X	4*	Cancer
1**	Y	4*	Cancer
1**	Z	4*	Fever

Fig: 2 (a) (b) K-Anonymous data

Advantages: [14]

- This method is protects identity disclosure when it is releasing sensitive information.

Disadvantages:

- It is prone to homogeneity attack and the background knowledge attack.
- Does not protect attribute disclosure to sufficient extent
- It has the limitation of *k*-anonymity model which fails in real scenario when the attackers try other methods.

3.2 Randomized Response Based PPDM

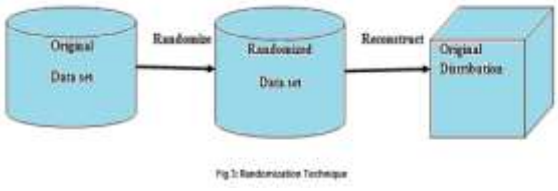
In Randomized response [14][31], the data is muddled in such a way that the central place cannot let know with probabilities better than a pre-defined threshold, whether the data from a customer contains truthful information or false information. The information received from each individual user is scrambled and if the number of users is significantly large, the aggregate information of these users can be predictable with good amount of accuracy. One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. The process of data collection in randomization method encompasses two steps [14]. In the first step, the data providers transmit the randomized data to the data receiver. In second step, reconstruction of the original distribution of the data is done by the data receiver by employing a distribution reconstruction algorithm

Advantages:

- It is a simple technique which can be easily implemented at data collection time.
- It is more efficient. However, it results in high information loss.

Disadvantages:

- It is not required for multiple attribute databases
- It results in high information loss



3.3 Condensation based PPDM

In a condensation approach, [15][31] constrained clusters are constructed in the data set, and pseudo-data from the condensed statistics of these clusters are generated. The constraints on the clusters are defined in terms of the sizes of the clusters which are chosen in such a way to preserve *k* anonymity. Some of the advantages and disadvantage of this method is [14]:

Advantages:

- This approach works with pseudo-data rather than with modifications of original data
- It is a better preservation of privacy compared to the techniques which simply use modifications of the original data.

Disadvantages:

- The pseudo-data have the same format as the original data.
- So, it is no longer necessitates the redesign of data mining algorithms

3.4 Perturbation Based PPDM

The perturbation approach the data service is not allowed to learn or recover precise records. This restriction naturally leads to some challenges. This method does not reconstruct the original data. But it can do only distributions. So, new algorithms need to be developed which use these reconstructed distributions in order to perform mining of the underlying data. This means that for each individual data problem, a new distribution based data mining algorithm needs to be developed [3]. Some of the advantages and disadvantage of this method is [14]:

Advantages:

- It is very simple technique.
- Different attributes are treated independently.

Disadvantages:

- Does not reconstruct the original vale rather than only distortion
- The perturbation approach does not provide a clear understanding of the level of indistinguishability of different records

3.5 Cryptography Based PPDM

In many cases, multiple parties may require to share private data. They want to share information without leakage at their end. For example, different branches in an educational institute wish to share their sensitive sales data to co-ordinate themselves without leaking privacy. This requires secure and cryptographic protocols for sharing the information across the different parties. Cryptography [31], in the presence of a intruder extends from the traditional tasks of encryption and authentication .In an ideal situation, in addition to the original parties there is also a third party called "trusted party". All parties send their inputs to the trusted party, who then computes the function and sends the appropriate results to the other parties. The protocol that is run in order to compute the function does not leak any unnecessary information. Sometimes there are limited leaks of information that are not dangerous. This process requires high level of trust. Some of the advantages and disadvantage of this method is [14]:

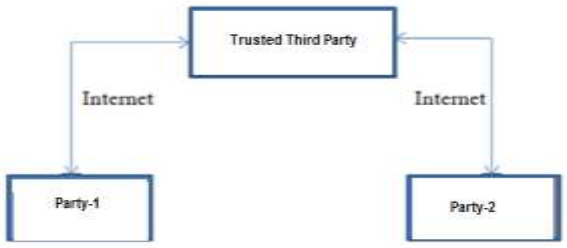


FIG 4: System using Semi Trusted Third Party

Advantages:

- Cryptography offers a well-defined model for privacy for proving and quantifying it.
- There exit a vast range of cryptographic algorithms

Disadvantages:

- It is difficult to scale when more than a few parties are involved
- It does not guarantee that the disclosure of the final data mining result may not violate the privacy of individual records.

3.6 Elliptical Curve Cryptographic Based PPDM

Elliptic Curve Cryptography (ECC) is an smart alternative to conservative public key cryptography, such as RSA. ECC are useful in the implementation on constrained devices where the major computational resources such as speed, memory is limited and low-power wireless communication protocols are used. That is because it attains the same security levels with traditional cryptosystems using smaller parameter sizes as discussed in [16][31]. Some of the advantages and disadvantage of this method is:

Advantages:

- Very few attributes are required compared to traditional cryptographic approaches

Disadvantages:

- Implementation is a complex task.

4. PRIVACY PRESERVING DATA MINING ALGORITHM

The followings are some of the data mining algorithms that have been used for privacy preservation:

4.1 Association Rule Mining

The association rule mining problem can be formally stated as follows [17]: Let $I = \{i_1, i_2, ---, i_m\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if $X \subseteq T$. An association rule is an implication of the form, $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varphi$. The rule $X \Rightarrow Y$ holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y . The rule $X \Rightarrow Y$ has support s in the transaction set D if s% of transactions in D contains X U Y. To find out if a particular itemset is frequent, it count the number of records where the values for all the attributes in the itemset are 1.

4.2 Clustering

Clustering [18] is a data mining method that has not taken its real part in the works already quoted although, the most important algorithm of this method was very studied in the context of privacy preserving, which is k-means algorithm [19]. Surveying privacy preserving k-means clustering approaches apart from other privacy preserving data mining ones is important due to the use of this algorithm in important other areas, like image and signal processing where the problem of security is strongly posed [20]. Most of works in privacy preserving clustering are developed on the k-means algorithm by applying the model of secure multi-party computation on different distributions (vertically, horizontally and arbitrary partitioned data). Among the formulations of Partition clustering based on the minimization of an objective function, k-means algorithm is the most widely used and studied. Given a dataset D of n entities (objects, data points, items,…) in real p-dimension space R^p and an integer k. The k-means clustering algorithm partitions the dataset D of entities into k-disjoint subsets, called clusters. Each cluster is represented by its center which is the centroid of all entities in that subset. The need to preserve privacy in k-means algorithm occurs when it is applied on distributed data over several sites, so called "parties" and that it wishes to do clustering on the union of their datasets. The aim is to prevent a party to see or deduce the data of another party during the execution of the algorithm. This is achieved by using secure multi-party computation that provides a formal model to preserve privacy of data.

4.3 Classification Data Mining

Classification is one of the most common applications found in the real world. The goal of classification is to build a model which can predict the value of one variable, based on the values of the other variables. For example, based on financial, criminal and travel data, one may want to classify passengers as security risks. In the financial sector, categorizing the credit risk of customers, as well as detecting fraudulent transactions is classification problems. Decision tree classification is one of the best known solution approaches. The decision tree in ID3 [21] is built top-down in a recursive fashion. In the first iteration it finds the attribute which best classifies the data considering the target class attribute. Once the attribute is identified in the given set of attributes algorithm creates a branch for each value. This process is continued until all the attributes are considered. In order to calculate which attribute is the best to classify the data set information gain is used. Information gain is defined as the expected reduction in entropy. Another most actively developed methodology in data mining is the Support Vector Machine (SVM) classification [22]. SVM has proven to be effective in many real-world applications [23]. Like other classifiers, the accuracy of an SVM classifier crucially depends on having access to the correct set of data. Data collected from different sites is useful in most cases, since it provides a better estimation of the population than the data collected at a single site.

4.4 Bayesian Data Mining

Bayesian networks are a powerful data mining tool. A Bayesian network consists of two parts: the network structure and the network parameters. Bayesian networks can be used for many tasks, such as hypothesis testing and automated scientific discovery. A Bayesian network (BN) is a graphical model that encodes probabilistic relationships among variables of interest [24].

Formally, a Bayesian network for a set V of m variables is a pair (Bs, Bp). The network structure Bs = (V,E) is a directed acyclic graph whose nodes are the set of variables. The parameters Bp describe local probability distributions associated with each variable. The graph Bs represents conditional independence assertions about variables in V: An edge between two nodes denotes direct probabilistic relationships between the corresponding variables. Together, Bs and Bp define the joint probability distribution for V.

5. RELATED WORKS

R. Agrawal, T. Imielinski, A. N. Swami [17] present a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions.

Kiran.P, Kavya N. P. [25] propose a SW-SDF based privacy preserving data classification technique. They uses sensitive weight to differentiate between sensitive attribute values.

Kiran.P, Kavya N. P. [26] proposed a method SW-SDF personal privacy for k means clustering. This algorithm groups objects in to k-clusters. Each item is placed in to the closest cluster based on the distance measures computed. In this method they propose an algorithm in such a way that the resultant clusters are almost equal to the original cluster and the privacy is retained.

J. Vaidya, C.W. Clifton, [27] proposed an association rule mining algorithm based on the Apriori algorithm. The Apriori algorithm was selected to extract the candidate set.

Vaidya J., Clifton C. [28] proposed a Privacy-Preserving k-means clustering over vertically partitioned Data based on the k-means algorithms. The k-means algorithm was selected for partitions of the clusters based on their similarity.

Vaidya J., Clifton C. [29] provides solution for privacy-preserving decision trees over vertically partitioned data.

Yu H., Vaidya J., Jiang X. [22] provides solution for privacy-preserving SVM classification on vertically partitioned data (PP-SVMV). It securely computes the global SVM model, without disclosing the data or classification information of each party to the others (*i.e.*, keeping the *model privacy* as well as the *data privacy*)

Vaidya J., Clifton C. [30] provides solution for two parties owning confidential databases to learn the Bayesian network on the combination of their databases without revealing anything else about their data to each other.

6. CONCLUSION

Classical data mining algorithms implicitly assume complete access to all data. However, privacy and security concerns often prevent sharing of data, thus devastating data mining projects. Recently, researchers have gaining more interested on finding solutions to this problem. Several algorithms have been proposed to do knowledge discovery, while providing guarantees on the non-disclosure of data. In this paper we have given a brief discussion on different dimensions of classification of privacy preservation techniques. We have also discussed different privacy preservation techniques and their advantages and disadvantages. We also discuss some of the popular data mining algorithms like association rule mining, clustering, decision tree, Bayesian network etc. used to privacy preservation technique. We also presented few related works in this field.

7. REFERENCES

[1] Chris Clifton and Donald Marks, “Security and privacy implications of data mining”, In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.

[2] Daniel E. O’Leary, “Knowledge Discovery as a Threat to Database Security”, In Proceedings of the 1st International Conference on Knowledge Discovery and Databases (1991), 107–516.

[3] R. Agrawal and R. Srikant, “Privacy-preserving data mining”, In ACM SIGMOD, pages 439–450, May 2000

[4] Y. Lindell and B. Pinkas, “Privacy preserving data mining”, J. Cryptology, 15(3):177–206, 2002.

[5] M. Sharma, A. Chaudhary, M. Mathuria and S. Chaudhary, “A Review Study on the Privacy Preserving Data Mining Techniques and Approaches”, International **J**ournal of **C**omputer **S**cience and **T**elecommunications, ISSN 2047-3338, Vol.4, Issue. 9, September 2013, pp: 42-46

[6] J. Han and M. Kamber. “Data Mining Concepts and Techniques”. Morgan Kaufmann Publishers, San Diego, CA 92101-4495, USA, 2001.

[7] Xinjing Ge and Jianming Zhu (2011). “Privacy Preserving Data Mining, New Fundamental Technologies in Data Mining”, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-547-1, InTech, DOI: 10.5772/13364. Available from: http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/privacy-preserving-data-mining

[8] Divya Sharma,” A Survey on Maintaining Privacy in Data Mining”, International Journal of Engineering Research and Technology (IJERT), Vol. 1 Issue 2, April – 2012,p.26.

[9] S. Rizvi and J.R Hartisa. “Maintaining data privacy in association rule mining”. In Proc. of the 28th VLDB Conference, pages 682-693, Hong-Kong, China, 2002.

[10] Y. Saygin, V. S. Verykios and A. K. Elmagarmid. “Privacy preserving association rule mining”. In RIDE, pages 151-158, 2002.

[11] V. S. Iyenger. “Transforming data to satisfy privacy constraints”. In Proc. Of SIGKDD’02, Edmonton, Alberta, Canada, 2002.

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

[13] Sweeney L, "k-Anonymity: A model for protecting privacy" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 557-570, 2002.

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

[15] Charu C. Aggarwal and Philip S. Yu,(2004) “A condensation approach to privacy preserving data mining”, In EDBT, pp. 183–199.

[16] Ioannis Chatzigiannakis, Apostolos Pyrgelis, Paul G. Spirakis, Yannis C. Stamatiou “Elliptic Curve Based Zero Knowledge Proofs and Their Applicability on Resource Constrained Devices” University of Patras Greece, arXiv: 1107.1626v1 [cs.CR] 8 Jul 2011

[17] R. Agrawal, T. Imielinski, and A. N. Swami. “Mining association rules between sets of items in large database”s. In P. Buneman and S. Jajodia, editors, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207{216, Washington, D.C., May 26{28 1993}.

[18] Jain A., Murty M., and Flynn P.” Data Clustering: A Review”, ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[19] MacQueen J., "Some Methods for Classification and Analysis of Multivariate Observations," in Proceedings of 5th Berkley Symposium Math. Statistics and Probability, California, USA, pp. 281-296, 1967.

[20] Erkin Z., Piva A., Katzenbeisser S., Lagendijk R., Shokrollahi J., Neven G., and Barni M., "Protection and Retrieval of Encrypted Multimedia Content: When Cryptography meets Signal Processing," EURASIP Journal of Information Security, vol. 7, no. 17, pp. 1-20, 2007.

[21] Lindell Y. , Pinkas B., "Privacy Preserving Data mining*", International Journal of Cryptology, Citesheer, 2000

[22] Yu H., Vaidya J., Jiang X.: "Privacy-Preserving SVM Classification on Vertically Partitioned Data", PAKDD Conference, 2006.

[23] V. N. Vapnik, "Statistical Learning Theory", John Wiley and Sons, 1998.

[24] G. Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data," Machine Learning, vol. 9, no. 4, pp. 309-347, 1992.

[25] Kiran.P, Kavya N. P.,"SW-SDF based privacy preserving data classification", International Journal of Computers & Technology, Volume 4 No. 3, March-April, 2013.

[26] Kiran.P, Kavya N. P.,"SW-SDF based privacy preserving for k-means clustering", International Journal of Scientific & Engineering Research, Volume 4, issue 6,pp. 563-566, ISSN 2229-5518, June, 2013.

[27] J. Vaidya, C.W. Clifton, "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. URL citeseer.nj.nec.com/492031. html.

[28] Vaidya J., Clifton C.: "Privacy-Preserving k-means clustering over vertically partitioned Data. ACM KDD Conference, 2003.

[29] Vaidya J., Clifton C.: "Privacy-Preserving Decision Trees over vertically partitioned data". Lecture Notes in Computer Science, Vol 3654, 2005.

[30] Vaidya J., Clifton C. "Privacy-Preserving Naive Bayes Classifier over vertically partitioned data". SIAM Conference, 2004.

[31] Shrivastava A., Dutta U.: "An Emblematic Study of Different Techniques in PPDM". International Journal of Advanced Research in Computer science and Software Engineering (IJARCSSE), Vol.3, Issue.8, pp.443-447, 2013.