A SURVEY ON PRIVACY PRESERVING DATA MINING

1. Aneesha Shaik, Research Scholar, Dept. of Computer Science & Engineering, Acharya Nagarjuna University, Guntur.

2. Dr. Shaheda Akthar, Registrar $_{FAC}$, Dr. Abdul Haq Urdu University, Kurnool.

Abstract: Data mining is the process of extracting useful knowledge from the enormous amounts of data repositories. Finding the association, similarity and prediction are different techniques used in data mining. In general, data need to be shared with other parties for analysis purpose to help in effective decision making or to accomplish other goals such as research work. The data analyst who performs data analysis acquires data from different sources. Each data set may have different characteristics. This data can contain sensitive attributes which can uncover some touchy information of an individual or an organization. When the data is disclosed to outsiders it imposes a threat to its privacy. Before the data is released for analysis it is cardinal to protect the sensitive data elements. Hence, maximizing the data utilization and minimizing the data protection danger are the two significant issues to be considered during the information exchange. However, the most challenging task of data analyst is to apply data mining algorithms on the protected data. In this paper, we made a drill-down analysis of different privacy preserving data mining techniques.

I. Introduction:

Data mining is one of the finest areas of extracting the relevant and useful knowledge from the vast data sources. The data generated from different sources may exist in different forms such as structured, semi structured and un-structured data. Algorithms like association, classification and clustering are some techniques present in the data mining. These algorithms need different types of datasets to facilitate data mining. Each dataset is unique in its nature. Depending on the type of domain, data can also be differentiated. Each data from different domain exhibits different characteristics such as records and attributes. The data used in the process of data mining consists of sensitive and non-sensitive information. In general, datasets such as hospital data, voter identity and customer data consist of certain attributes which are sensitive and should be protected from the intruders or the outside world. If such confidential information is disclosed to the other party, privacy of the data has been compromised. The ultimate aim is to protect the privacy of the data. In this case, the data owner will send the original data to analyst by means of preserving the privacy of the data. The analyst has very crucial role to play in this because data which he holds in different form (not in original form). Data publishing and Data mining are two connected streams where one concentrates on security and privacy of data by conversion while the other is concerned with applying data mining algorithms on those modified data. Privacy of data is the crucial aspect of data mining which cannot be compromised at any cost. The word privacy is defined very clearly by **Clifton and Murat [19]** in their article. There are several data publishing algorithms like generalization, suppression and randomization etc. Privacy preserving data mining broadly classified into centralized and distributed. In centralized data mining, all the data resides at one place where original data has been modified to another form before applying data mining algorithm. In distributed method, the data is preserved by partitioning into several chunks sent to different locations. Further, data is partitioned in such a way that each node consists of all samples but few attributes is known to be vertical partitioned. On the other hand, each node containing all attributes but few samples is known to be horizontally partitioned. In classical way, all these data mining algorithms work very well on original data but not on the modified data. In the past decades, many algorithms were proposed based on different characteristics of data. **Lindell and Pinkas[9,10]** has

discussed about SMC (Secure multi-party computation) a cryptographic solution to distributed data either vertical or horizontal partitioned data. **Kargupta and Qi wang[8]** proposed a random matrix based spectral filtering techniques for random perturbation and retrieving the original values from that modified data. **Hung and Du[4]** has proposed a method of finding the correlation among the Data sets and reconstruction methods through Principle Component Analysis and Bayes estimates techniques. **Evfimievski and Srikant[3]** proposed an association based apriory algorithm on random perturbed data**. Agarwal and Srikant[1]** has proposed an ID3 classification tree based on Randomized data, in which they perturbed the original data with random noise generated from any distribution. **Wenliang and Mikhail[20]** proposed a protocol named privacy preserving cooperative least square which is used for sharing of data between multiple parties securely. **Stanley and Usmar[21]** proposed a new data transformation technique using object based similarity on vertically partitioned data and applied the clustering algorithm by reducing the data dimensionality. In this paper, we first discuss the privacy publishing concepts which help us understand how we protect our sensitive data from adverse. Later, we discuss the privacy preserving data mining with their taxonomy.

Name	Age	Address	Social status number	constituency
Mohan	34	Delhi	376689	West Delhi
Ravi	55	Hyderabad	999977	Uppal
Geeta	33	madras	786756	South madras

1. Table of electoral data

Name	Age	Address	Social status number	Bill No	Disease
Mohan	34	Delhi	376689	22	Lung cancer
Ravi	55	Hyderabad	999977	33	Head cancer
Geeta	33	Madras	786756	44	Jon disease

2. Table of patient diagnosis information of a Hospital

II. Privacy Preserving data publishing:

Data publishing is the process of unfolding the data table to the outside world. Before data is disclosed, it can be analyzed thoroughly to ensure that sensitive data is protected. Data linkage is the process in which sensitive knowledge can be extracted by linking the different data tables from different resources. Different techniques and methods are available through which sensitive attributes can be protected during the information change. The following are some of the methods used in data publishing. From the table 1 and table 2 even after removing the sensitive information

1. K anonymity: [sweeney 2002a][16]

It is the process in which quasi identifiers are protected from other people by using different methods. This k anonymity gives the information about how many attributes can be protected before it can be released to the outside world

1.1. Generalization

It is the process in which quasi attribute values can be generalized. For example, from table 1, 'Social status number' is generalized with a range of value and the attribute 'age' is also generalized

Age	Social status number	constituency
30-40	376679-376699	West Delhi
50-60	999975-999988	Uppal
20-30	786750-786760	South madras

3. Published Table after Generalization

1.2 Suppression

In this method, the quasi identifiers are replaced with certain symbols like "*". From table 1, the 'Social Status number' and 'Age' attributes are replaced with certain symbols.

Age	Social status number	constituency
*	*	West Delhi
*	*	Uppal
*	*	South madras

4. Published Table after Suppression

2. ℓ diversity [Wong 2006][17]

This method makes the probability of recognizing the record value in the given table by $1/\ell$. From the table 5 it is observed that last two records are very close to each other, in this case adversary cannot distinguish between two records.

Age	Social status number	Bill No	Disease
34	376679-376699	22	Lung cancer
55	999975-999988	33	Head cancer
*	786750-786760	*	Jon disease
*	998654-998665	*	Jon disease

5. Published Table for ℓ diversity

3. Perturbation: In this method, privacy preserving can be achieved by changing the original values in different ways. [18]

1. Adding Noise (Randomization): A Random value is added to the original attribute value to protect from adversary. Generally the added noise is selected from certain probability distribution. For example $x_1, x_2, x_3 \ldots\ldots x_n$ are original attribute values from the privacy table and $y_1, y_2, y_3 \ldots\ldots y_n$ are probability values from certain distribution, which are added to the original values to get$z_1, z_2, z_3 \ldots\ldots z_n$.

$$x_1 + y_1 = z_1, x_2 + y_2 = z_2, \ldots . x_n + y_n = z_n \qquad (1)$$

From table 1 quasi attribute 'age' is randomized for privacy preservation.

Age	Address	Social status number	constituency
77	Delhi	376689	West Delhi
98	Hyderabad	999977	Uppal
44	Madras	786756	South madras

6. Data table after randomization

4. Cryptographic Distributed techniques [18]

The data is distributed across multiple sites and sharing of data is achieved through cryptographic protocols.

4.1. Horizontal partition

In this, the data is partitioned and distributed across many sites, where a part consists of certain number of records by retaining total number of attributes. From table 2, data is distributed between two sites 1 and site 2.

Name	Age	Address	Social status number	Bill No	Disease
mohan	34	delhi	376689	22	Lung cancer
Ravi	55	hyderabad	999977	33	Head cancer

7. Horizontal Data partition at Site 1 from the table 2

Name	Age	Address	Social status number	Bill No	Disease
geeta	33	madras	786756	44	Jon disease

8. Horizontal Data partition at Site 2 from the table 2

4.2 Vertical partition

In this, data is partitioned and distributed across many sites by holding few attributes and total records. From table 2, data is distributed between two sites 1 and site 2.

Name	Age	Address
mohan	34	delhi
Ravi	55	hyderabad
geeta	33	madras

9. Vertical Data partition at Site 1 from the table 2

Social status number	Bill No	Disease
376689	22	Lung cancer
999977	33	Head cancer
786756	44	Jon disease

10. Vertical Data partition at Site 2 from the table 2

III. Privacy Preserving Data mining Algorithms

PPDM classifications	Data is secured by	Data Mining Algorithms (Associations, Classifications, Clustering)
Centralized	Perturbation., Randomization, K-anonymity, L-diversity	[1][8]
Distributed	1. Horizontally Partitioned Data	[2][3][5][6][7] [9][10][11][12]
	2. Vertically Partitioned Data	

1. Privacy preserving on centralized data

1.1 Privacy Preserving ID3 based on Perturbation data **[Agrawal and Srikant][1]**

In this, data on which classification algorithm is to be applied is perturbed with random noise. An ID3 classification algorithm is a decision tree based algorithm. It has many steps to perform to form the tree. During the construction of the tree, best splitting nodes or attributes are identified based on split criterion entropy , gain, and gain ratio.

Before applying the algorithm from the randomized data w, distribution of noise Y is Estimated, which is added to the original data X and is identified through Baye's rule.

$$F'_X(a) = \int_{-\infty}^{a} f_X(z|X+Y=w)dz \qquad (2)$$

From the equation 1.

$$F'_X(a) = \frac{1}{n}\sum_{i=1}^{n} F'_{X_i} = \frac{1}{n}\sum_{i=1}^{n} \frac{\int_{-\infty}^{a} f_Y(w_i-z)f_X(z)dz}{\int_{-\infty}^{\infty} f_Y(w_i-z)f_X(z)dz} \qquad (3)$$

After differentiating with dz the equation (3)

$$f'_X(a) = \frac{1}{n}\sum_{i=1}^{n} \frac{f_Y(w_i-a)f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i-z)f_X(z)dz} \qquad (4)$$

1.2The privacy preserving Properties of Random Data Perturbation Techniques **[Kargupta , Qi Wang, Siva Kumar][8]**

In this, the author has covered the drawback of traditional random perturbation technique which makes it difficult to recover the sensitive data. They proposed a random matrix based perturbation technique to preserve the privacy in the original data. They separated successfully original data from the perturbed data.

2. Privacy preserving on distributed data

In this method, theoriginal data is partitioned and distributed to many sites. Data is partitioned in such a way that all the sites will retain all the attributes and subsets of records known to be horizontal partitioned data. When the sites will have all records but retain subsets of attributed is known to be vertically partitioned.

2.1 Privacy Preserving on Horizontally partitioned distributed data**.[Lindell and Pinkas][9,10]**

The data is partitioned horizontally at different nodes. Each node consists of subsets of samples with all attributes. Basically ID3 is a decision tree classifier based on construction of decision tree and on node splitting criterion. In order to split the tree, it takes into account of tree splitting criterion like information gain and gene index. In this situation data is partitioned into different node, where data mining algorithm is to be applied. Each node contributes its subsets of values for computation of classification but conditioned that no node can be able to see the subsets of another node. For that reason, a secure mechanism is applied over here through SMC (Secure multi party computation), Oblivious Polynomial Evaluation and Homomorphic Substitution. By using these each party can perform securely the data mining computation of classification. Similarly other data mining techniques like basyian classifier, SVD, KNN , associations and clustering can be performed.

2.2 Privacy Preserving on vertically partitioned distributed data **[Du and Zhan][2]**

Here, the data is vertically partitioned such that each node holds total samples with few subset of attributes. In this, privacy preserving data mining with classification algorithm ID3, has two parties which hold the data as s_a and s_b on vertically partitioned data. Each node at first perform the information gain , entropy and gini index on s_a and s_b on the same class to find the best split among the different attributes. Once the best split node is identified the process has been repeated for remaining number of nodes.

2.3 Privacy-preserving clustering by object similarity-based representation and dimensionality reduction transformation **[S. R. M. Oliveira and O. R. Zaiane][11]**

In this paper, the author has proposed new transformation method called object based similarity and dissimilarity on vertically partitioned data. Also proposed pattern recognition based dimensionality reduction techniques. In this they discussed about complexity of object based representation technique on vertically partitioned data of clustering algorithm. They implemented dimensionality reduction technique on both centralized and vertically partitioned distributed data.

2.4 Privacy-preserving distributed /c-means clustering over arbitrarily partitioned data**. [G. Jagannathan and R. N. Wright][5]**

Unlike the traditional way of data partitioning i.e. partitioning data horizontally and vertically, in this method data is partitioned arbitrarily where each sample and attribute has been distributed arbitrarily. The k-means clustering on arbitrarily partitioned data was investigated for different nodes.

2.5 Privacy-preserving distributed mining of association rules on horizontally partitioned data**. [M. Kantarcioglu and C. Clifton][7]**

Here the secure multi-party computation was used to apply the data mining association algorithm. At first, each node calculates the local association among its attributes and later shares its computed results through secure multiparty computation with the other nodes without revealing its original identity.

2.6 Privacy Preserving Naive Bayes classifier for horizontally partitioned data**. [M. Kantarcioglu and J. Vaidya.][6]**

The data is partitioned horizontally among different sites. Naive Bayes is the basic classification used among different classifiers. In this technique, several procedures for nominal and numerical data were proposed. Different nodes share their data for secure multi computation to preserve the security of sensitive data during computation of data mining classification algorithm.

2.7 Privacy preserving regression modeling via distributed computation**. [A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter.][12]**

The sensitive information at all nodes is perturbed with linear regression equation. Regression coefficients are building blocks which are shared among different sites after computation. These regression coefficients are useful to assess the effect on dependent variable by independent variable and coefficient of determination is useful to determine the strength of relationship between each variable.

2.8 Privacy preserving mining of association rules **[A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke][3]**

The data is distributed among different sites (clients) perturbed with random noise. In this system, the server collects the data from different sites and applies the association algorithm for data mining. Each client sends its perturbed data to the server without revealing its identity. They use recovery technique to get back the original data from randomized data. Usually Apriori based association rule mining on categorical data is used.

IV. Conclusion:

Privacy preserving data is one of the important concepts in Data Security and publishing. Data can be made secured before it is delivered outside. Privacy preserving data mining applies the mining algorithms on preserved data. One of the important challenges of Data Mining is to apply these algorithms on perturbed data. In this paper we made a study and analysis on Privacy preserving Data Mining in terms of centralized and distributed views of the data.

References

1. Agrawal, R. and R. Srikant. Privacy-preserving data mining. In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, Dallas, TX, 2000, pp. 439–450.

2. W. Du and Z. Zhan. Building decision tree classifier on private data. In C. Clifton and V. Estivill-Castro, editors, *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, volume 14, pages 1-8, Maebashi City, Japan, Dec. 9 2002. AustraUan Computer Society.

3. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217-228, Edmonton, Alberta, Canada, July 23-26 2002.

4. Z. Huang, W. Du, and B. Chen. Deriving private information from randomized data. In *Proceedings of the 2005 ACM SICMOD International Conference on Management of Data*, Baltimore, MD, June 13-16 2005.

5. G. Jagannathan and R. N. Wright. Privacy-preserving distributed /c-means clustering over arbitrarily partitioned data. In *Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 593-599, Chicago, IL, Aug. 21-24 2005.

6. M. Kantarcioglu and J. Vaidya. Privacy preserving naive bayes classifier for horizontally partitioned data. In *Workshop on Privacy Preserving Data Mining held in association with The Third IEEE International Conference on Data Mining,* Melbourne, FL, Nov.19-22 2003

7. M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In *The A CM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD^02),* pages 24-31, Madison, Wisconsin, June 2 2002.

8. H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03),* Melbourne, Florida, Nov. 19-22 2003.

9. Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Advances in Cryptology - CRYPTO 2000,* pages 36-54. Springer-Verlag, Aug. 20-24 2000.

10. Y. Lindell and B. Pinkas. Privacy preserving data mining. *Journal of Cryptology,* 15(3):177-206, 2002.

11. S. R. M. Oliveira and O. R. Zaiane. Privacy-preserving clustering by object similarity-based representation and dimensionality reduction transformation. In *Workshop on Privacy and Security Aspects of Data Mining (PSDM'04) in conjunction with the Fourth IEEE International Conference on Data Mining (ICDM'04),* pages 21-30, Brighton, UK, Nov. 1 2004.

12. A. P. Sanil, A. F. Karr, X. Lin, and J. P. Reiter. Privacy preserving regression modelling via distributed computation. In *KDD '04- Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,* pages 677-682, New York, NY, USA, 2004. ACM Press.

13. J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pages 639-644, Edmonton, Alberta, Canada, July 23-26 2002.

14. J. Vaidya and C. Clifton. Privacy-preserving /c-means clustering over vertically partitioned data. In *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pages 206-215, Washington, DC, Aug. 24-27 2003.

15. J. Vaidya and C. Clifton. Privacy preserving naiVe bayes classifier for vertically partitioned data. In *2004 SLAM International Conference on Data Mining,* pages 522-526, Lake Buena Vista, Florida, Apr. 22-24 2004.

16. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and knowldege based systems, 10(5):571–588, 2002a.

17. R.Wong, J. Li, A. Fu, and K.Wang. (alpha, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In Proc. 12th ACMSIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pages 754–759, 2006.

18. Charu C. Aggarwal, Philip S. Yu (auth.), Charu C. Aggarwal, Philip S. Yu (eds.) "Privacy-Preserving Data Mining: Models and Algorithms" Springer 2008.

19. Clifton, C., Kantarcioğlu, M., Doan, A., Schadow, G., Vaidya, J., Elmagarmid, A., & Suciu, D. (2004). *Privacy-preserving data integration and sharing. Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '04.*

20. Wenliang Du and Mikhail .J. Atatttah "Privacy-Preserving Cooperative Scientific Computation"

21. Stanley R.M. Oliveora and USMAR R. Zaione Achieve.