International Journal of Linguistics and Computational Applications (IJLCA) ISSN 2394-6385 (Print)
Volume 3, Issue 1, January – March 2016 ISSN 2394-6393 (Online)

PPDM Methods and Techniques

Jyothi Mandala
Asstistant Professor, GMRIT, Rajam, Srikakulam. AP, INDIA
jyothirajb4u@gmail.com

Abstract**—** Data Mining is a process of discovering useful information in large data repositories. Data mining techniques have been used to enhance information retrieval systems. Privacy-preserving data mining (PPDM) refers to the area of data mining that is primarily concerned with protecting against disclosure of individual data records i.e., to safeguard sensitive information. To address about privacy researchers in data mining community have proposed various solutions. Privacy in data mining can be obtained by various techniques like Perturbation, Anonymization and Cryptographic.The main intension of this paper is to explore various PPDM techniques in literatures for handling privacy issues in data mining.

Keywords**—** Perturbation, PPDM, K-Anonymity, Cryptography

1. Introduction

DataMining research deals with the extraction of useful information from large collection of data.Most organizations own large information in databases. It is often highly valuable for organizations to have their data analyzed by external agents. Knowledge discovery techniques need to be applied on collection of databases of different organizations involved in the same field. Analysis of these databases is beneficial for the organizations. Thepaper[1] addresses two issues associated with electronic data gathering: confidentiality of the organization that supplies the database and authentication of the database provided.

Data may contain some sensitive individual information such as medical and financial information. This sensitive data may be exposed during the data mining process and it is possible to learn lot of information about individuals from public data. Privacy preserving Data Mining (PPDM) is an area where data mining algorithms can be applied on centralized or distributed data without compromising with the privacy of the sensitive data. In paper [14] PPDM is defined as "getting valid data mining results without learning the underlying data values". PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results.

PPDM can be one of the three approaches: 1. Data hiding, in which sensitive raw data like identifiers, name, addresses, etc were altered, blocked or trimmed out from the original database in order for the users of the data not to be able to compromise another person's privacy. 2. Rule hiding, in which sensitive data extracted from the data mining process is excluded for use, because confidential information may be derived from the released knowledge and 3. Secure multiparty computation (SMC), where distributed data are encrypted before released or shared for computations, so that no party knows anything except its own inputs and the results.

Privacy preserving Data Mining Algorithms can be applied on centralized database or distributed database. In centralized database environment data are all stored in a single database, where as in distributed database environment data is distributed in different locations. . 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.

2. Perturbation technique

In this technique the original data are not open, and users can only access perturbed data. The data mining is done on the perturbed data to extract patterns about the original data. The estimation of original values is not possible using this technique.Agarwal and srikantpropsed a random-value perturbation method[2] attempts to preserve privacy of the data by modifying values of the sensitive attributes using a randomized process.Data perturbation can be done by adding noise to the original data or by multiplication of some noise value to the original data to prevent the identification of confidential information relating to a particular individual. [3]Proposed following method for data perturbation technique. This option is only for numeric values. Add any value in attribute's values (input from the user) suppose any attribute have value 12, 14, 11, 15, 9 etc. user give input 5, so add 5 to each value and output will be 17, 19, 16, 20, 14 etc. This option is only for non-numeric values. Change the non-numeric value of selected attribute by any other non-numeric value. (Suppose values is car1 so replace by selected value suppose p1 or other) (Used ASCII in programming)

- Select non numeric attribute.
- Find distinct values of selected attribute.
- Generate distinct values mapping to each value of distinct values of selected attribute.

4. Replace old distinct values with generated/new distinct values. Consider the following dataset

Table1: Microdata

Name	age	gender	Salary	Car
Alice	25	F	25000	car1
Bob	22	M	24000	car3
Peter	24	M	24500	car1
James	28	M	23700	car2

From the above table non-numeric attribute is Gender. So it contain only two distinct value in whole column that is M and F. so create to random value for this two value.

Suppose For M is P and for F is Q. then replace value M by P and F by Q.

- This option is only for numeric and non-numeric values

Interchange the values of the same attribute (by randomly choose value only from that attribute)

1. Select Attribute from the data file.
2. Loop through all instances, I=0 TonumberOfInstance
   a. Randomly select instance/row and get Value of selected attribute of that instance/row.
   b. Set randomly selected value to selected attribute of instance 'I'.

From the above table suppose selected attribute is Gender so randomly select any value from Gender attribute and replace first values by that. Again randomly select new value and replace second values by that selected value. Continue for all value of selected row.

- This option is only for numeric values

Find mean of numeric value of any particular row's numeric attribute and replace chosen attribute value by this answer.

1. Select non numeric attribute.
2. Loop through all instances.
   a. Find mean of numeric attribute of all each instance/row.
   b. Set mean to selected attribute.

Consider the above table ,there are 2 numeric attributes age and salary and 3 non numeric attributes
suppose the numeric attribute selected is salary
Numeric attributes are 2. For the first row add all these ex. 25+25000=25025. Mean =25025/2=12512.5
So replace salary attribute 25000 by 12512.5
So after completing all the rows the dataset will be looking like this,

Table 2: Modified Microdata

Name	**age**	**gender**	**Salary**	**Car**
Alice	**25**	**F**	**12512.5**	**car1**
Bob	**22**	**M**	**12011**	**car3**
Peter	**24**	**M**	**12262**	**car1**
James	**28**	**M**	**11864**	**car2**

In paper[4], LiLiu, Lantaricoglu, bhavani propose an individually adapted perturbation model, which enables the individuals to choose their own privacy levels. Their proposed model is two-phase perturbation model.

3. K-Anonumity

In paper [5] Samarati and Sweeney address the problem of releasing person-specific data while, at the same time, safeguarding the anonymity of the individuals to whom the data refer.To achieve the *k*-anonymity requirement, they used both generalization and suppression for data anonymization.

For example, consider hospital dataset, Table 3, which contains patients diagnosis records. This dataset can be used by the researchers to study the characteristics of various diseases. The raw data(micro data) contains the identities (e.g. names) of individuals, which are not released to protect their privacy. However, there may exist other attributes that can be used, in combination with an external database, to recover the personal identities.

Table 3: Assume the following data table which is published by a hospital

ID	Attributes			
	Age	Sex	Zip code	Disesase
1	26	M	83661	Headache
2	24	M	83634	Headache
3	31	M	83967	Toothach
4	39	F	83949	Cough

The above table does not explicitly indicate the names of patients. However, if an adversary has access to the voter registration list in Table4, he can easily discover the identities of all patients by joining the two tables on {Age, Sex, Zipcode}. These three attributes are, therefore, the quasi-identifier (QI) attributes.

Table 4: Voter Information

ID	Attributes			
	Name	Age	Sex	Zip code
1	Alice	26	M	83661
2	Bob	24	M	83634
3	Peter	31	M	83967
4	James	39	F	83949

A table is *k*-anonymous if the QI values of each tuple are identical to those of at least *k*-1 other tuples. Below Table shows an example of 2-anonymous generalization for Table3.

Table 5: 2-Anonymous

ID	Attributes			
	Age	Sex	Zip code	Disesase
1	2*	M	836**	Headache
2	2*	M	836**	Headache
3	3*	M	839**	Toothach
4	3*	F	839**	Cough

Even with the voter registration list, an adversary can discover the real disease of Alice only with probability 50%. In general, *k*anonymity guarantees that an individual can be associated with his real tuple with a probability at most 1/*k* .In paper[6], Yan Zhu and Lin Peng formulated a modified entropy 1- Diversity model which was extension of basic k-anonymity.

4. Cryptographic Techniques

This technique is used if two or more parties want to perform data mining task on combined datasets. This problem is referred as Secure Multi-party Computation (SMC) problem[7].By using cryptographic techniques we can perform privacy preserving classification [8], privacy preserving association rule mining [9] and privacy preserving clustering [10].

Secure multi-party computation has two models: A semi-honest participant will not deviate from the protocol but will only try to extract some extra information from the messages On the other hand; a malicious adversary can arbitrarily deviate from the protocol.

4.1 Public-key cryptosystems (asymmetric ciphers)

A cipher is an algorithm that is used to encrypt plaintext into cipher text (encryption) and cipher text to plain text (decryption).Ciphers are said to be divided into two categories: private key and public key.Private-key (symmetric key) algorithms require a sender to encrypt a plaintext with the key and the receiver to decrypt the cipher text with the same key. A problem with this method is that both parties must have an identical key, and somehow the key must be delivered to the receiving party.Example algorithms are DES, AES.

A public-key (asymmetric key) algorithm uses two separate keys: a public key and a private key. The public key is used to encrypt the data and only the private key can decrypt the data. A form of this type of encryption is called RSA.A classical example for PPDM is Yao's millionaire's problem: two millionaires want to find out who is richer without revealing toeach other how many millions they each own. In [11] a solution to the Yao's millionaire problem is given.Ashraf B. El-Sisi and Hamdy M. Mousa[12] proposed a cryptographic approach for PPDM which uses a semi-honest model. This employs a public-key cryptosystem algorithm on horizontally partitioned data among three or more parties. The approach is as follows:

Consider three parties A, B and C

- A generates the public key KPA. This KPA is known to B and C.
- Now A, B and C encrypts their dataset DB_i with KP_A key. Encryption is applied on each row of the dataset. This encryption is denoted as $KP_A(DB_i)$ as shown in Fig 1. Only A can perform decryption on these datasets as A only knows his private key.

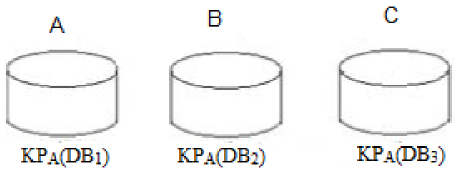

Fig 1: A, B and C encrypts their data sets

- A passes his encrypted dataset i.e. $KP_A(DB_1)$ to B.
- Now B performs random shuffle of $KP_A(DB_1)$ and $KP_A(DB_2)$ and forwards the resultant dataset to C as shown in Fig 2.

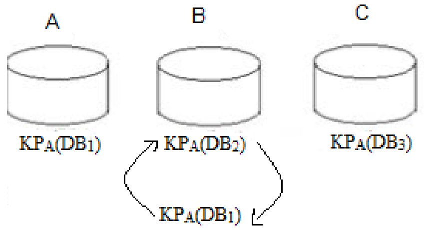

Fig 2: B shuffles the data sets transactions

- C adds and shuffles hid dataset transactions $KP_A(DB_3)$ to the transactions received from B as shown in Fig 3

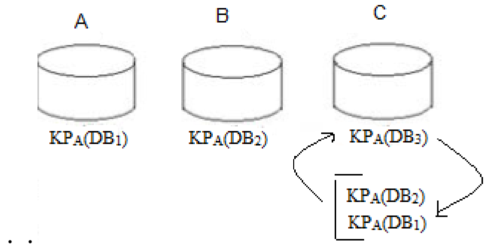

Fig 3: C shuffles the datasets transactions

- C forwards these transactions back to A.
- A decrypts the entire dataset with his secret private key as shown in Fig 4. A can identify his own transactions. However, A is unable to link transactions with their owners because transactions are shuffled.
- Finally A publishes the transactions to all other parties.

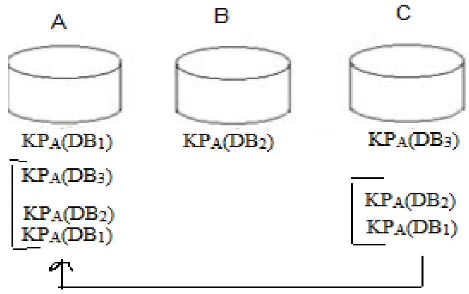

Fig 4: A perform the decryption.

Using the above approach the information that is hidden is what data records where in the possession of which party. Murat Kantarcioglu and Chris cliftion [13] proposed another approach for PPDM using cryptographic techniques. This approach uses commutative encryption for privacy preserving association rule mining on horizontally distributed data. Commutative encryption means the order of encryption does not matter. If a plaintext message is encrypted by two different keys in a different order, it will be mapped to the same cipher text. Formally, commutatively ensures that Ek1 (Ek2(x)) = Ek2 (Ek1(x)). To determine global candidate itemsets the approach is as follows:

Each party encrypts its own frequent itemsets along with enough "fake" itemsets. The encrypted itemsets are the passes to other parties until all parties have encrypted all itemsets. These are passed to a common party to eliminate duplicates and to begin decryption. This set is then passed to each party and each party decrypts each itemset. The final result is the common itemsets. Fig 5 shows an example of this approach where ABC and ABD are common itemsets.

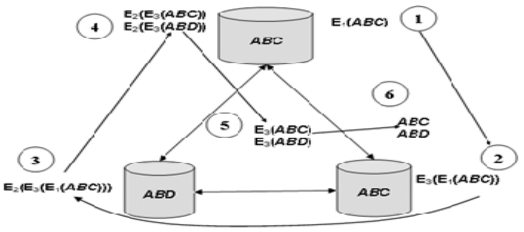

Fig 5: Determining global candidate item sets

The paper [15] addressed a privacy-preserving protocol for filling missing values using decision-tree classification algorithm for data that is horizontally partitioned between two parties.

5. Conclusion

With the development for need of data analysis of data and also the privacy disclosure problem about individual or company is identified when releasing or sharing data to mine. To solve this new research field on privacy preserving data mining is evolved. The main intension of this paper to through various PPDM techniques in literatures for handling privacy issues in data mining. To provide accurate results in data mining, many PPDMtechniques are task based. There is no such technique which overcomes all privacy issues. For a new comer, this paper provides a brief review about existing privacy preserving techniques.

References

[1] Privacy Preserving Electronic data gathering, E. C. Laskari, G. C, Melitieou, D. K. Tasoulis, M.N. vrahatis, 2005 elsevier

[2] Agrawal R, Srikant R (2000) Privacy-preserving data mining. In: Proceeding of the ACM SIGMOD conference on management of data. ACM Press, Dallas, TX, pp 439–450

[3] Kiran Patel, Hitesh Patel, Parin Patel, "Privacy Preserving in Data stream classification using different proposed Perturbation Methods", © 2014 IJEDR | Volume 2, Issue 2 | ISSN: 2321-9939

[4] Li Liu, Murat Kantarcioglu and BhavaniThuraisingham, "The applicability of the perturbation based privacy preserving data mining for real-world data", Data & Knowledge Engineering 65 (2008) 5–21.

[5] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression", In Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[6] Yan ZHU and Lin PENG, "Study on K-anonymity Models of Sharing Medical Information", 1-4244-0885- 7/07/$20.00 ©2007 IEEE.

[7] W. Du and M. J.Atallah. Secure Multi-party Computation Problem and their Applications: A Review and Open Problems. In Proc. of Data Warehousing and Knowledge Discovery DaWak-99, Florence, Italy, August, 1999.

[8] M. Kantarcioglu and J. Vaidya. Privacy Preserving Naïve Bayes Classifier for Horizontally Partitioned Data. In Proc. of the IEEE ICDM Workshop on Privacy Preserving Data Mining, Pages 3-9, Melbourne, FL, USA, November 2003.

[9] J. Vaidya and C. Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In Proc. of the 8th ACM SIGKDD Intl. Cong. On Knowledge Discovery and Data Mining, Pages 639-644, Edmonton, AB, Canada, July 2002.

[10] J. Vaidya and C. Clifton. Privacy Preserving K-Means Clustering Over Vertically Partitioned Data. In Proc. of the 9th ACM SIGKDD Intl. Cong. On Knowledge Discovery and Data Mining, Pages 206-215, Washington, DC, USA, August 2003.

[11] O. Goldreich, "Secure multi-party computation",(working draft). [online]. Available: http://www.wisdom.weizmann. ac.il/ oded/ pp. html

[12] Ashraf B. El-Sisi and Hamdy M. ousaEvaluation of Encryption algorithms for privacy preserving Association Rules Mining",International Journal of Network Security,vol 14, No.5,sep 2012.

[13] Murat kantarcioglu and chris Clifton," Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", IEEE Transactions on Knowledge and Data Engineering , vol. 16 No. 9, sep 2004

[14] C. Clifton, M. Kantarcioglu, and J. Vaidya. Defining Privacy For Data Mining. In Proc. of the National Science Foundation Workshop on Next Generation Data Mining, pages 126-133, Baltimore, MD, USA, November 2002.

[15] GeethaJagannathan, Rebecca N. Wright, "Privacy-Preserving Imputation of Missing Data", Data & Knowledge Engineering, 2008 Elsevier

M.Jyothi received B.Tech degree in Computer Science and Information Technology from AITAM College, Tekkali, India and M.Tech degree in Computer Science and Engineering from JNTUK, Kakinada, India. She is pursuing her Ph.D from Acharya Nagarjuna University Guntur, in the area of Data Mining. Currently she is working as Assistant professor in Information Technology Department at GMRIT, Rajam.