International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
255 – 259

An approach of Privacy Preserving Data mining using Perturbation & Cryptography Technique

Akash Siddhpura
Computer Engineering
Noble Group of Institutions
Junagadh, India
aakash.siddhpura@gmail.com

Prof. Daxa V. Vekariya
Computer Engineering
Noble Group of Institutions
Junagadh, India
daxa.vekariya@ngivbt.edu.in

***Abstract* —** Due to the wide deployment of information technology, privacy concern has been major issue in data mining. So for that new path is identified which is known as Privacy Preserving Data Mining (PDDM). Available PDDM techniques are Perturbation, Generalization, Anonymization, Randomization and Cryptography. All of them have some advantages as well as disadvantages also. If apply only cryptography PDDM using symmetric key encryption algorithm, then there will chances of losing data, because if anyone knows the key then data is available to anyone. If we apply perturbation PDDM only then it will not give you accurate result. So if we will use cryptography and perturbation then it will achieve security as well as very less chances of losing data after applying the privacy preserving.

Keywords- *Data Mining, Privacy Preserving Data Mining; Privacy Preserving; Cryptography, Encryption; Privacy Preserving using Perturbation; Privacy Preserving using Cryptography; Cryptography and Perturbation*

*****

I. INTRODUCTION

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. 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. [1] The need of using the data mining techniques for Privacy Preservation is increasing rapidly. A lot of research has been done in the field of anonymization and cryptographic techniques. Anonymization and perturbation techniques can be considered better when compared to cryptographic techniques in terms of complexity and efficiency for large number of users. When compared on the basis of information loss and privacy achieved anonymization suffers a significantly high information loss. The proposed hybrid technique can successfully achieve the goal of privacy preservation without any information loss as the using the algorithm the distorted values can be reverted to its original values successfully [2]. There is no single technique reliable in all domains is also analyzed. While the proposed methods are, only near to our goal of privacy preservation. The study is now concluded with the fact that Cryptography and Random Data Perturbation methods perform superior than the other existing methods and considered as one of the best technique for encryption of sensitive data [3]. 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 [4]. Data perturbation way is very efficient used in data mining alone in centralized environment, but it will produce some problems in distributed environment where there are many different data sources and data have the disagreement problem [5]. In secret key cryptography, a single key is used for both encryption and decryption. Public or asymmetric key cryptography involves the use of key pairs: one private key and one public key. Both are required to encrypt and decrypt a message or transmission [6].The main idea of Perturbation- Based technique involves increasing a noise in the raw data in order to perturb the original data distribution and to preserve the content of hidden raw data [7].

II. LITERATURE SURVEY

A. A Survey: Privacy Preservation Techniques in Data Mining [1]

Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. Privacy preserving data mining algorithms are measured in terms of its performance, data utility, level of uncertainty or resistance to data mining algorithms etc. The Extracted information can be a patterns, rules, clusters or classification models. Sensitive individual information such as medical and financial information, often get exposed to several parties including collectors, owners, users and miners. 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. The framework for PPDM works level wise; the first level is raw data or databases where transactions exist in. The second level is data mining algorithms and techniques that ensure privacy. Privacy can be ensured by different processes like blocking, suppression, perturbation, modification, generalization, sampling etc. The third level is the output of different data mining algorithms and methods. In this paper it has also focused on 2 different techniques, first Perturbation based PPDM: In this technique different attributes are preserved

 ISSN: 2454-4248
04 – 2018

independently but Original data values cannot be regenerated and Loss of information can be occur. Another one is Cryptography based PPDM: Here Transformed data are exact and protected but Better privacy compare to randomized approach and This approach is especially difficult to scale multiple parties are involved.

B. A proposed hybrid approach for privacy preserving data mining [2]

Types of data may be image, video, text or images etc. Privacy preserving techniques are generally classified in to main five categories 1. Distribution of data 2. Data Modification 3. Data mining algorithms 4. Data Hiding 5. Privacy preserving technique. There are many kind of the perturbation techniques. In the random perturbation data will be considered as matrix and numeric fields are represented in the form of intervals. But this technique generates a huge information loss. After that another methods are there which are, translation and rotation based perturbation. In translation based method single value as a noise will be taken and then that value will be added to the data or subtracted from the data, so original data will be changed. In rotation based method data is rotated by the angle Ɵ. Fast perturbation for large size databases and it will focus on minimizing the utility of the data to hide the original values of the data set by using a tree structure. Now in cryptography technique, it will hide the sensitive information from the public by encryption algorithms. Privacy preserving data mining techniques are measure on the basis of Information loss, Privacy preserved, Computational time, Complexity, Dependency on the size of data. Each and every technique for privacy preservation data mining had some advantage as well as disadvantages. If we use anonymization and perturbation techniques then there will be a huge amount of information loss. Whereas if we use cryptography based technique then it will increase the complexity and computational time. At the end we can say that, compare to the cryptography, anonymization and perturbation techniques are easy to apply PDDM. So the main goal of the privacy preservation without information loss as the using the algorithm the changed values can be reverted to its original values successfully.

.

C. Privacy preserving data mining- A state of art[3]

There is no single technique available in all domains is analyzed. PDDM algorithms are classified in many ways, one of them is Data Distribution. Here data can be categorized as Centralized and Distributed. As in the centralized database, all the data is stored in a single database. Where in distributed different databases may be at different location are used to store the data. In distributed databases data owners may not prepared to release their data. Another method for PDDM algorithm is Data hiding and Rule hiding. Here sensitive data were either changed, blocked or trimmed out. Data Modification, it is used to modify or change the original values released to ensure high-level privacy protection. Perturbation, Blocking, Aggregation, Swapping, Sampling are the example of data modification. Cryptography technique does not work with large databases and this technique is difficult to implement among few parties. Data perturbation, there may be chances of adversarial attacks and after adding the noise still we can identify the pattern. Anonymization technique do not preserve sensitivity of an attribute. In the Blocking based technique original values of the dataset will not be reconstructed. So in this way we can say that, all proposed methods are, only near to privacy preservation. At the end we can conclude that Cryptography and Data perturbation methods are better than the other existing methods.

D. A survey of cryptographic and Non-cryptographic techniques for privacy preservation[4]

There are many advantages of cryptographic technique are Robust, Anonymity, Fairness, Accountability. Well there are also some disadvantages like takes a long time to figure out the code, takes long to create the code, cryptography is long process also. There are two type of privacy preserving techniques. One is Cryptographic, in which it decrease the granularity of arranging so that it decrease the privacy, because of that there will be information loss. In Non cryptographic- randomization technique, it implements data distortion technique for adding little noise in the actual data. Where as in Non cryptographic-anonymization method implements generalization and suppression method to generate individual record in distinguishable within the group record. So there are these much amount of methods which are working for privacy preservation in data mining. To provide accurate output many privacy preservation techniques are provided, but no such techniques took place which works perfect. This paper provides the knowledge about the cryptographic and non-cryptographic methods with centralized and distributed database.

E. Research on privacy preserving technology of data mining[5]

Data perturbation is classified into two kinds: random perturbation (RP) and randomized response (RR). The order relationship of exchanging number in RP achieves privacy-preserving goal by hiding the corresponding relation between numbers and their objects; RR is to add appropriate amount of random noise into data under the condition of no change in raw data distribution. Data can be centralized data and distributed data or horizontal partition and vertical partition. At present centralized privacy-preserving method mainly adopts RR and RP to realize. Distributed privacy preserving data mining algorithm is realized through the employment of SMC. Data perturbation way is very efficient used in data mining alone in centralized environment, but it will produce some problems in distributed environment where there are many different data sources.

F. A symmetric key cryptographic algorithm[6]

There are mainly two type of cryptography. One is symmetric key and another is asymmetric key. Symmetric key algorithms are fast and quick. In that only one key will be used

 ISSN: 2454-4248
04 – 2018

for the encryption as well as decryption also. Where as in asymmetric key is little bit slower than symmetric key algorithm, because there will 2 keys, one key for the encryption and another is for the decryption. While applying the cryptography there are some parameters which we are going to achieve which are. Confidentiality, Data integrity, Authentication. While now in symmetric key cryptography, can be categorized in to two categories. One is Stream cipher, in which it is operate on a single bit at a time, and implement some form of feedback mechanism so that the key is continually changing. Second is Block cipher, in which it encrypts one block of a data at a time using the same key on each block. If we apply block cipher then on the same block of text, key will the same, so it will return same encrypted block of code, while as in the stream cipher on the same block of text, key will be different, so it will return different block of text for the same block of text. At the end we can conclude that asymmetric key algorithm is more secure then the symmetric key algorithm.

G. Preserving privacy using data perturbation in data stream[7]

The proposed hybrid algorithms for data perturbation that is the data perturbation for privacy preserving in data stream clustering. Perturbation techniques are often evaluated with two basic metrics: level of privacy guarantee and level of model-specific data utility preserved. The main idea of Perturbation- Based technique involves increasing a noise in the raw data in order to perturb the original data distribution and to preserve the content of hidden raw data. In this paper data perturbation algorithms have been proposed for data set perturbation. Also included permutation techniques like Translation Based Perturbation and Rotation Based Perturbation.

H. Attribute based encryption with privacy preserving in clouds[8]

In cloud computing, Security and privacy both are very significant matter. A decentralized access control technique with anonymous authentication, which provides user revocation and prevents replay attacks, is achieved. Asymmetric key algorithm uses different key for both encryption and decryption. As per paper, The Paillier crypto system is a probabilistic asymmetric algorithm for public key cryptography. Pailier algorithm use for Creation of access policy, file accessing and file restoring process. There are two classes of ABEs. First is Key-policy ABE and second Cipher text-policy.

III. PROPOSED WORK

In this system, we are going to apply privacy preservation on the datasets. First of all we will convert all data in to their respective asci values. Now after that we will apply perturbation techniqes means we are going to add noise on the data. Next we will apply cryptography technique on the data. In the cryptography technique we will apply ECB algorithm for the next procedure. Now data perturbation is applied successfully. Now to achieve original data back we have to perform the reverse process again.

Algorithm

Step 1 – Start
Step 2 – Take dataset
Step 3 – Import dataset in to SQL Server
Step 4 – Perform Data Cleaning Process
Step 5 – Fetch data in to system
Step 6 – Want to perform privacy preservation? , Goto 12
Step 7 – Convert data in to respective ASCII value
Step 8 – Perform Perturbation (Add Noise in to the Data)
Step 9 – Perform Encryption (ECB Algorithm)
Step 10 – More records are there, Goto 17
Step 11 – Goto 7
Step 12 – Perform Decryption (ECB Algorithm)
Step 13 – Perform Perturbation (Remove Noise from the Data)
Step 14 – Convert ASCII values in to respective data or values
Step 15 – More records are there, Goto 17
Step 16 – Goto 12
Step 17 – Stop

```
Step 1 – Start
Step 2 – Take dataset
Step 3 – Import dataset in to SQL Server
Step 4 – Perform Data Cleaning Process
Step 5 – Fetch data in to system
```

```
Step 6 – Want to perform privacy preservation?, Goto 12
Step 7 – Convert data in to respective ASCII value,
Repeat the steps until i=no. of rows
        Repeat the steps until j=no. of columns
        Step 7.1 – Celldata ← Cell[i][j]
        Step 7.2 – Convert Celldata's value in to their
respective asci values
        Step 7.3 – rowdata ← Celldata
```

```
Step 8 – Perform Perturbation (Add Noise in to the
Data), Repeat the steps until i=no. of rows
        Repeat the steps until j=no. of columns
        Step 8.1 – length ← Find the length of Cell[i][j]
        Step 8.2 – value ← Cell[i][j]
        Step 8.3 – TempValue ← value + length
        Step 8.4 – UpdatedValue ← TempValue *
length
        Step 8.5 – Cell[i][j] ← UpdatedValue
```

ISSN: 2454-4248
04 – 2018

Step 9 – Perform Encryption (ECB Algorithm), Repeat the steps until i=no. of rows
Repeat the steps until j=no. of columns
Step 9.1 – $plaintext_{ij}$ – Cell[i][j]
Step 9.2 – CipherText ← F($plaintext_{ij,}$ key)
Step 9.3 – Cell[i][j] = CipherText

Step 10 – More records are there, Goto 17
Step 11 – Goto 7
Step 12 – Perform Decryption (ECB Algorithm), Repeat the steps until i=no. of rows
Repeat the steps until j=no. of columns
Step 12.1 – $cipherext_{ij}$ – Cell[i][j]
Step 12.2 – PlainText ← F($ciphertext_{ij,}$ key)
Step 12.3 – Cell[i][j] = PlainText

Step 13 – Perform Perturbation (Remove Noise from the Data), Repeat the steps until i=no. of rows
Repeat the steps until j=no. of columns
Step 13.1 – length ← Find the length of the Cell[i][j]
Step 13.2 – value ← Cell[i][j]
Step 13.3 – TempValue ← value / length
Step 13.4 – OriginalValue ← TempValue - length
Step 13.5 – Cell[i][j] ← OriginalValue

Step 14 – Convert ASCII values in to respective data or values, Repeat the steps until i=no. of rows
Repeat the steps until j=no. of columns
Step 14.1 – tempdata ←Convert asci values' in to their respective characters/digits
Step 14.2 – Cell[i][j] ← tempdata
Step 15 – More records are there, Goto 17
Step 16 – Goto 12
Step 17 – Stop

IV. RESULT

Here we had applied data perturbation on the different datasets. All the datasets are downloaded from archive.ics.uci.edu (Machine Learning, UCI).We had taken Indian Liver Patient Dataset [13], Balance Scale Dataset [14], Ablon Dataset [15], and Bank Marketing Dataset [16]. We had applied our proposed work on all above listed datasets. In the following figure we had shown the No. of records v/s Time (In Seconds).

Sr. No.	Dataset Name	No. of Records	Time (In Seconds)
1	Indian Liver Patient	583	22
2	Balance Scale	625	12
3	Ablon	4117	32
4	Bank Marketing	45211	800

Table 1. Dataset Details and Time taken for final output

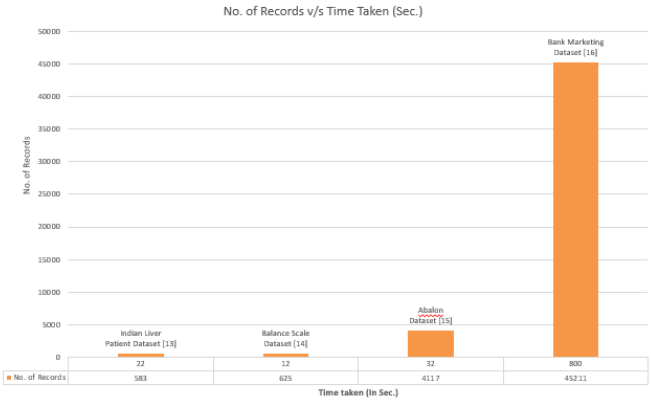

Fig 1. No. of records v/s Time (In Seconds)

We had also generate graph of different techniques v/s different datasets.

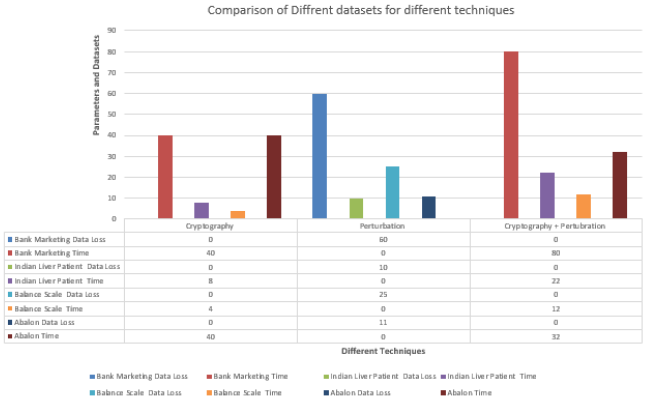

Fig 2. Different techniques v/s Different Dataset

The result and graph shows that our proposed technique is more efficient than the other available techniques.

V. FUTURE WORK

In ECB, it will generate the same output for the same cipher. So pattern can be recognized. We can also work with the video and audio data.

VI. CONCLUSION

Privacy preserving can be achieved by using two techniques, by adding the noise and using the cryptography we can protect the data. Here Data loss will be 0%. But it will some time while performing encryption as well as decryption. There are no chances of data loss. While if we apply only Perturbation technique then there will be chances of data loss. If we apply only cryptography technique then quality of data will not that much good, we had improved quality of the data here also.

REFERENCES

[1] Hina Vaghashia, Amit Ganatra, PhD, "A Survey: Privacy Preservation Techniques in Data Mining" © International Journal of Computer Applications (0975-8887), Volume 119 – No 4, June 2015.
[2] Arshweer Kaur, Sanjeev Sofat, "A proposed hybrid approach for privacy preserving data mining" © Inventive Computation Technologies (ICICT), Internation Conference On.
[3] Mamta Narwaria, Suchita Arya, "Privacy preserving data mining- A state of the art", © Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference On.
[4] Bhawani Singh Rathore, Anju Singh, Divakar Singh, "A Survey of cryptographic and Non-cryptographic techniques for privacy preservation" © International Journal of Computer Application (0975-8887), Volume 130, No 13, November 2013.
[5] Anu Thomas, Jimesh Rana, "A Review on privacy preserving data mining approaches" © National Conference on Recent Research in Engineering and Technology (NCRRET – 2015)
[6] Ayushi, "A Symmetric key cryptographic algorithm", © 2010, International Journal of Computer Application (0975-8887), Volume I, No 15.
[7] Neha Gupta, IndrJeet Rajput, "Preserving Privacy using data perturbation in data stream", © International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 2, No 5, May 2013. ISSN : 2278 – 1323
[8] M. Suriyapriya, A.Joicy, "Attribute based encryption with privacy preserving in clouds", © International Journal on Recent and Innovation Trends in Computing and Communication, Volume:2, Issue: 2, ISSN: 2321-8169, 231-236
[9] Preet Chandar Kaur, Tushar Ghorpade, Vanita Mane, "Analysis of data security by using anonyization techniques", © Cloud System and Big Engineering, 2016 6th International Conference.
[10] Nisha Mattas, Smarika, Deepti Mehrotra, "Comparing Data Mining techniques for mining patents", © Advanced Computing & Communication Technologies, 2015 5th International Conference On
[11] Zakaria Gheid, Yacine Challal, "Efficient and Privacy Preserving k-means clustering for Big Data mining", © Trustcom/BigData/ISPA, 2016IEEE
[12] M.Prakash, Dr.G.Singarawel, "A new model for privacy preserving sensitive data mining", © Computing Communication & Networking Technologies (ICCCNT), 2012 3rd International Conference On
[13] Indian Liver Patient Dataset, Machine Learning, UCI. https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)
[14] Balance Scale Dataset, Machine Learning, UCI. http://archive.ics.uci.edu/ml/datasets/balance+scale
[15] Abalon Dataset, Machine Learning, UCI. https://archive.ics.uci.edu/ml/datasets/Abalone
[16] Bank Marketing Dataset, Machine Learning, UCI. https://archive.ics.uci.edu/ml/datasets/bank+marketing