IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 4, April 2015.
www.ijiset.com

ISSN 2348 – 7968

Privacy Preserving Data Mining On Relational Streaming Data

Ashish Mane, Prof. Pankaj Agarkar

[1] Department Of Computer Engineering, Savitribai Phule Pune Universty, Dr. D.Y. Patil School Of Engineering
Pune, Maharashtara, India.

[2] Department Of Computer Engineering, Savitribai Phule Pune University, Dr. D. Y. Patil School Of Engineering,
Pune, Maharashtra, India.

Abstract

Today every sector, whether it is business, educational, Medical, Military etc, required to store and handle large amount of information. Organization publishes the required data for data miner. Privacy of this information has become an important issue. This paper focuses on the privacy of sensitive attribute data. Here relational streaming data is used. Previous data mining applications were used to prefer the single level trust approach, in which data owner trust all data miner at single trust level, and only single perturbed copy was generated. In this paper Multilevel Trust approach is used, in which multiple perturbed copies of original sensitive attribute data is generated. Additive data perturbation approach is used to add the noise to sensitive attributes and group generation algorithm is used to generate the perturbed copies of original sensitive attribute data. In addition, when the records of database are updated, immediately new perturbed copies for that updated records are generated.

Keywords: *Privacy Preserving Data Mining (PPDM), MLT-PPDM, Additive Perturbation, Group Generation algorithm.*

1. Introduction

Various organization stores large amount of data and carries out different activities on this data. So, to maintain the privacy of this information have became an important issue. There are the data miners which are malicious and try to breach the privacy. Malicious data miner follows the legal protocols of the data mining applications, but their intention is to breach the privacy of sensitive information.

There is one approach, which is known as Data perturbation [1], [3], [4], [5], [6], [7]. This approach includes various techniques to perturb the data. This approach assumes single level trust on data miner. In the single level

Trust scenario data miner generates only one perturbed copy of the original data. Data owner trusts the data miner at same trust level. Today there are numerous amount of data mining applications exists where this single level trust approach cannot be used due to its high number of limitations. There is also uncertainty about the perturbed copy generated under single level trust approach.

So, multilevel trust approach [1] is proposed instead of single level trust approach. The proposed system uses the multilevel trust approach. In this, data owner trusts the data miner at different trust level, the proposed system introduce three trust levels namely high, low and medium and more than one number of perturbed copies are generated for each trust level. For high trust level more percentage of noise is added to original data and perturbed copies are generated. For low trust level less number of noise are added.

Additive perturbation technique [1], [3], [4], [5], [7], [8] is used for the addition of noise to the attribute data. This is very popular technique in which noise can be added to each record separately. It is of low cost.

In the proposed system, for the generation of perturbed copies sequential Group generation algorithm is used.

Malicious data miner may try to reconstruct the original data by combining the various copies at each trust level. This activity is known as diversity attack. It is important to prevent the sensitive information from diversity attack. The noise to the attributes of original data are added in such a way that the data miner finds it difficult to reconstruct the original information.

In proposed system, relational streaming data is used which was not used by any previous data mining applications. Data is updated periodically. All data mining related activities are to be carried out on relational streaming data. The main contribution is that, when records are updated, at same time new perturbed copies for the updated record are generated. Data miner need not have to worry about the generated records.

The rest of the paper is organized as follows, in section II

2. Literature Survey

There are various techniques for privacy preserving data mining which are proposed by various authors. Various researchers are working for various techniques on maintaining the privacy of sensitive data. Privacy preservation data mining was first proposed by the authors R. Agarwal and R. Shrikant [5]. In this, authors have discussed the technique to build the decision tree classifier from the training data in which individual record values are perturbed.

Authors in [10] and [11] discussed the technique of secure multiparty computation (SMC). The SMC provides strongest level of privacy. In this technique, the data miner of one party will not be able to know the others information except its own input and output results. Algorithms used by this technique are more complex and impractical for actual use. Also the input consists of large amount of datasets. Due to this reasons another approaches were preferred for preserving the privacy of original information.

Lokesh Patel and Prof. Ravindra Gupta in [3] have discussed various privacy preservation techniques. Various techniques to perturb the original data are discussed in this paper.

The next category is data perturbation approach [1], [3], [4], [5], [6], [7], [8]. In this approach there are various techniques such additive perturbation [1], [4], [5], [7], [8]. In this technique noise are added to individual records separately. This is very popular technique for perturbation of original data. The proposed system uses this technique to perturb the attribute values of the relational streaming data.

Data swapping technique is discussed in [2], it transforms the sensitive data of dataset by replacing the values of sensitive records.

The authors in [6], [9] have proposed Matrix multiplicative approach, in this technique inter record distances are preserved approximately and ultimately transformed records can be used with many types of data mining applications. Multiplicative perturbation approaches are not safe from malicious attacks.

The authors in [12], [13], [14], [15], [16], [17] proposed the technique of K-anonymity. This K-anonymity technique has two methods Generalization and Suppression. In generalization attribute values are generalized into the range so that the granularity of the representation is reduced. For example, date of joining of an employee in an organization can be generalized to year of joining, whereas in the suppression method the attribute value is removed completely.

There are also number of another techniques for privacy preservation data mining, Here only required important techniques are discussed.

3. BASIC CONCEPTS

A. Gaussian Noise

In this paper, original data is perturbed by addition of additive Gaussian noise [1], the noise that will be added are jointly Gaussian. Suppose K_1 to K_L are L random Gaussian variables. They are said to be jointly Gaussian, if and only if linear combination of them is also a Gaussian random variables.

B. Additive Perturbation

Data perturbation has one of the best applicable and most popularly used technique for perturbing the data, that technique is called additive perturbation technique [1],[2], [3], [5]. According to this technique some noise is added to the extracted original data.

Let Q be the original extracted attribute data, to which some random noise R is to be added and the perturbed copy P is to be generated. The idea is given below

$$P = Q + R \text{ - - - - - - - - } (1)$$

The original dataset follows the mean and covariance matrix. The covariance matrix KR is given as,

$$K_Q = [(Q - \mu_Q)(Q - \mu_Q)^T] \text{ - - - - - } (2)$$

Here noise R is considered to be independent of P. Also noise R is jointly Gaussian vector, mean and covariance matrix and the mean value is zero. The covariance matrix is given as

$$K_R = [RR^T] \text{ - - - - - - - } (3)$$

From the above equation the mean value of perturbed copy P is μ_Q, and its covariance matrix is K_P is given as,

$$K_P = K_Q + K_R \text{ - - - - - - - } (4)$$

The malicious data miner may try to achieve its goal by removing the noise from perturbed copy. To restrict this, the added noise in the perturbed copies should have same correlation, so that malicious data miner find it difficult to reconstruct the original data. This can be achieved by selecting $K_R = \sigma^2 K_Q$, where σ^2 is a perturbation magnitude.

If the perturbation magnitude is high, then perturbation is more, if it is low, then perturbation is less.

Additive perturbation has various benefits, they are illustrated below,

i) Noise can be added separately to each record.

ii) This technique is easy to use and has low cost.

iii) This technique can be used at web and corporate sector.

3. Proposed System

The following Figure illustrates the proposed System Architecture.

A. System Architecture

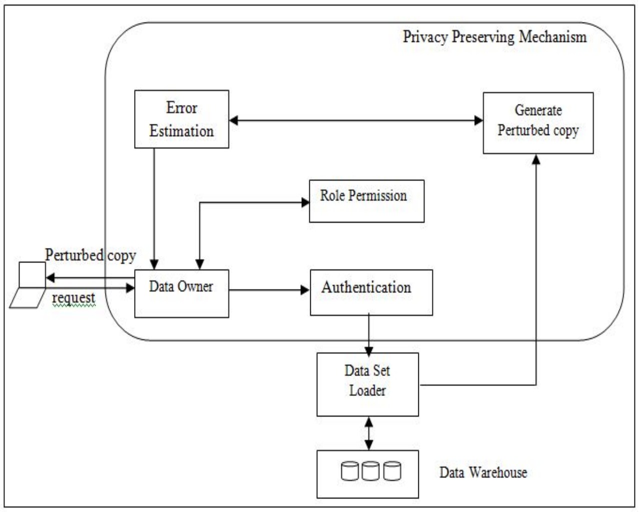


Fig. 1 Proposed System Architecture

Here data used is relational Streaming Data, it also means that the records in the database will be updated simultaneously. Activities are to be carried out on the relational streaming data. There is an data warehouse which contains the student, medical and banking dataset. The proposed System works according to the following stages.

Authentication: It is assumed that data miner has send the necessary request to data miner for required data. Data miner which is a program which performs various important activities required for data mining. In the first stage it perform the registration and login of data miner who wants the data from the system.

Role Determination: When the necessary login is done, at the same time data miners role is determined. In the role determination, two things are achieved-

i) Based on the type of of request, it is determined that which type of data to be given to data miner.

ii) The trust level of the data miner is also determined , it is most important step for the privacy preserving data mining. Here multilevel trust approach is used to determine the trust level of data miner. Three trust levels are defined namely High, Medium and low, based on the trust level the noise will be added to the original attribute data. If the trust level is high then low amount of noise is added, if the trust level is low then percentage of noise will be high.

Generation Of Perturbed Copy: After the determination of trust level of data miner, the necessary dataset is loaded from the data warehouse, here data warehouse contains student, medical and banking dataset. After loading the required dataset, required sensitive attribute data is extracted from dataset and is forwarded to perturbation phase.

Additive perturbation method is used for adding the noise to extracted attribute data. It is very effective perturbation approach and noise can be added to each record independently at any position. Which noise to be added is based on the type of attribute data, according to that the noise is added to attribute data. The noise is added in such a way that, if the data miner malicious, it would find it difficult to reconstruct the original data.

Sequential Group generation algorithm is used for the generation of multiple perturbed copies of the extracted original data. Here multiple perturbed copies of original extracted attribute data are generated based on the trust level of data miner. After generation of perturbed copies, these copies are forwarded to error estimation phase.

Error Estimation: In this phase the percentage of noise of the generated perturbed copies are checked, percentage of added noise is given by perturbation magnitude ‘σ’. Noise difference is verified properly. Here it is checked that the added noise are relevant according to the trust level of data miner or not, if they are perfect, the perturbed copied are forwarded to data owner, else they are sent back to perturbed copy generation phase.

After successful error estimation of the perturbed copies, these copies are sent to data owner and finally data owner publishes these copies to data miner. When the records in the database are updated, new perturbed copied for the updated records are also generated which is the new contribution to the proposed system.

4. Implementation

A. Authentication:

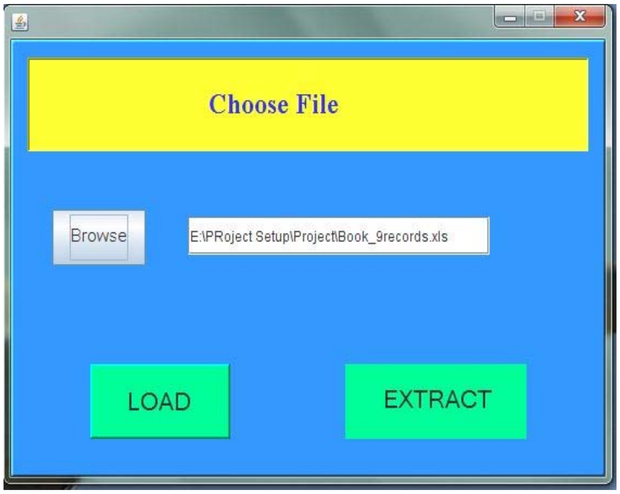
In the first stage Data miner sends the request for required information. The data owner performs the necessary registration and login activities of the data miner.

B. Role Determination:

When the login of data miner is performed, at the same time, which type of attribute data is given to the data miner and trust level of the data miner is determined.

C. Generation Of Perturbed Copies:

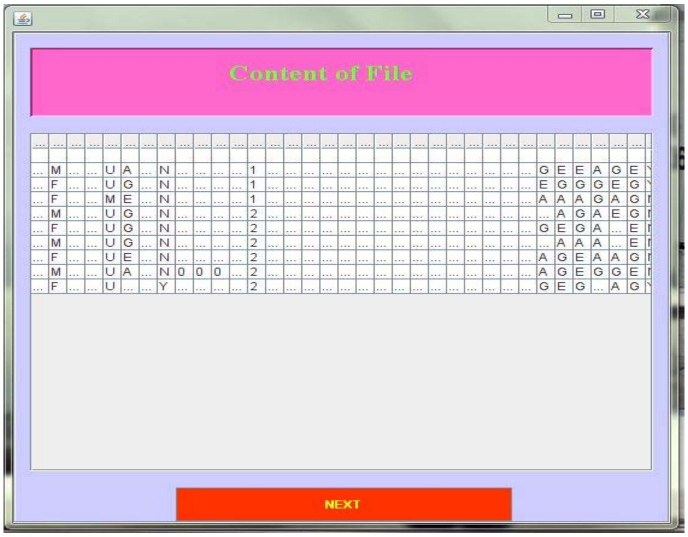
When trust level of data miner is determined, the loading of necessary dataset from the data warehouse is performed as shown below, in this system data warehouse contains student, Medical and banking dataset, there are more than thousand records maintained in each dataset, the records can be updated periodically.


Choose File

E:\PRoject Setup\Project\Book_9records.xls

Fig. 2 Loading Of Dataset

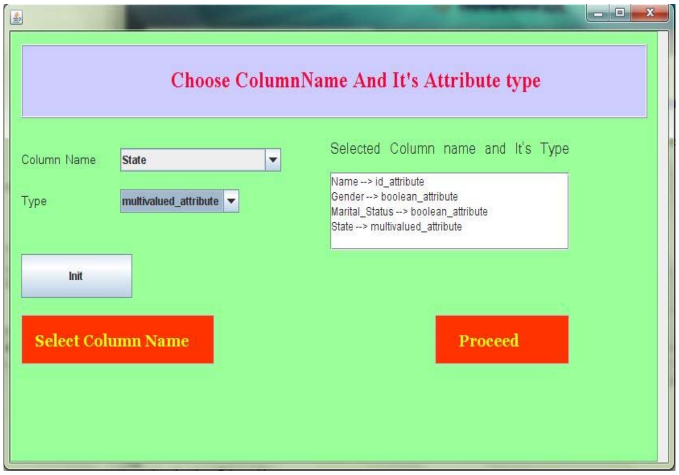
When required dataset is loaded successfully, the contents of datasets are displayed as below, dataset contains multiple attribute values. Here contents of student dataset are displayed,


Content of File

...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	M	...	...	U	A	...	N	...	...	...	...	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	G	E	E	A	G	E	Y
...	F	...	...	U	G	...	N	...	...	...	...	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	E	G	G	G	E	G	Y
...	F	...	...	M	E	...	N	...	...	...	...	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	A	A	A	G	A	G	N
...	M	...	...	U	G	...	N	...	...	...	...	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	A	G	A	E	G	N
...	F	...	...	U	G	...	N	...	...	...	...	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	G	E	G	A	...	E	N
...	M	...	...	U	G	...	N	...	...	...	...	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	A	A	A	...	E	N
...	F	...	...	U	E	...	N	...	...	...	...	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	A	G	E	A	A	G	N
...	M	...	...	U	A	...	N	0	0	0	...	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	A	G	E	G	G	E	N
...	F	...	...	U	...	...	Y	...	...	...	...	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	G	E	G	...	A	G	Y

Fig. 3 Contents of dataset

Then the necessary attribute data are extracted from the dataset for perturbation purpose as shown in the following fig, the required column name and their type are selected for perturbation, For example if the attribute 'name' is selected as a column name and its type ' id- attribute' is selected, then selected column name and its type is displayed in the adjacent window.


Choose ColumnName And It's Attribute type

Column Name: State

Type: multivalued_attribute

Selected Column name and It's Type

Name --> id_attribute
Gender --> boolean_attribute
Marital_Status --> boolean_attribute
State --> multivalued_attribute

Fig. 4 Selection Of Attributes For Perturbation

For perturbing the required attribute data, Group generation algorithm is used, this algorithm is explained below,

Group Generation Algorithm:

This algorithm requires that, the trust level of the data miner should be predefined, according to the requirement of this algorithm, the trust level is determined at first glance during the role determination phase, then the algorithm generates N perturbed copies in one group. Basically this algorithm has two types, one parallel generation which generates the noise R_1 to R_M in parallel and another one is sequential generation, which generates noise sequentially.

Algorithm 1: Parallel Generation:

```
{
  Input :  Q, K
  Output :  P
  Construct KR  ←  KQ
  Generate R    ←  KR
  Generate P    ←  HQ + R
  Output        ←  P
}
```

The above algorithm then generates PP as HQ + RR and outputs the result. Here this algorithm acts as the base algorithm for the next algorithm. i.e Sequential Algorithm.

Algorithm 2: Sequential Generation

The parallel generation algorithm generates the noise, but there is one drawback with the parallel generation algorithm that it requires large amount of memory. So, one will prefer to use the algorithm which achieves task of memory efficiency. Instead of parallel generation algorithm one can sequentially start generating noise R1 to RM each of which is a Gaussian vector of N dimension. The validity of alternative procedure is based on insight by using the sequential process.

Sequential Generation:

```
{
 Input: Q, K
 Output: Pm
 Construct R←  N (0…KQ)
 Generate Pi←  Q + R1
 For i=2 to M do
 Create Noise
 Generate P←  Pi 1 + E
 Output ←  Pi
}
```

The perturbed copies are generated successfully by applying the sequential generation algorithm, for addition of noise additive perturbation is used, the generated perturbed copies are displayed below,

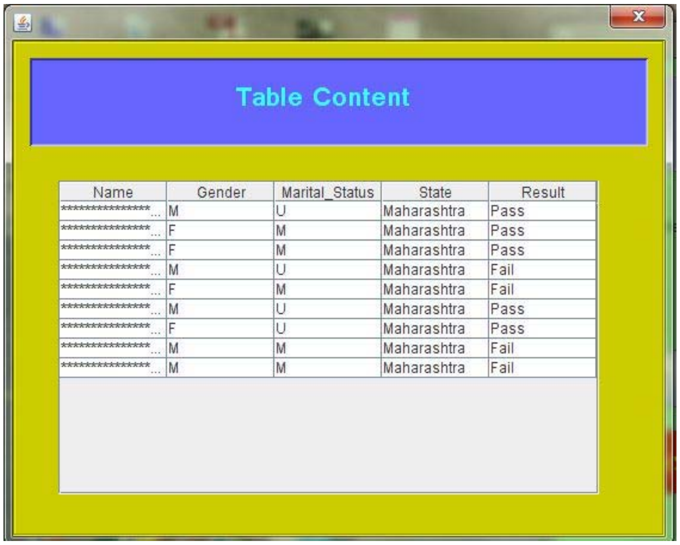


Table Content

Name	Gender	Marital_Status	State	Result
****************...	M	U	Maharashtra	Pass
****************...	F	M	Maharashtra	Pass
****************...	F	M	Maharashtra	Pass
****************...	M	U	Maharashtra	Fail
****************...	F	M	Maharashtra	Fail
****************...	M	U	Maharashtra	Pass
****************...	F	U	Maharashtra	Pass
****************...	M	M	Maharashtra	Fail
****************...	M	M	Maharashtra	Fail

Fig. 4 Generated Perturbed copies

Error Estimation:

After successfully generation of perturbed copies, these perturbed copies are forwarded to error estimation phase and then noise difference of this copies are checked in this phase. If conditions are satisfied then these copies are forwarded to data owner and finally to data miner.

5. Experimental Results

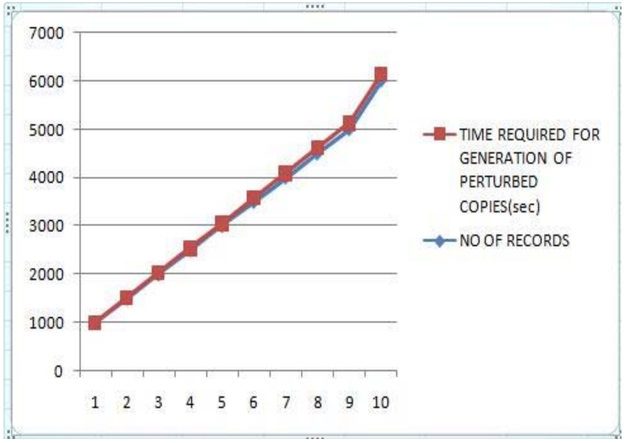
A. Time Required For Execution

The following Table I and Line chart I represents the time required for execution, as the number of records will increase, the time required for the generation of perturbed copies will also increased. In this experiment, first 1000 records are selected and the time required for thousand records is calculated which is 10 seconds.

TABLE I TIME REQUIRED FOR EXECUTION

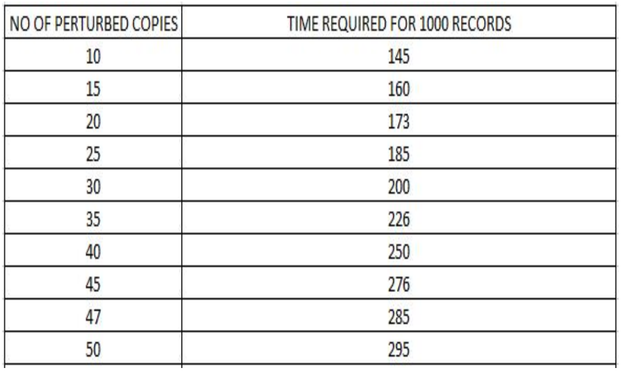
NO OF RECORDS	TIME REQUIRED FOR GENERATION OF PERTURBED COPIES(sec)
1000	10
1500	18
2000	35
2500	45
3000	60
3500	84
4000	95
4500	120
5000	135
6000	145

LINE CHART I



B. Time Required For Perturbed Copy Generation

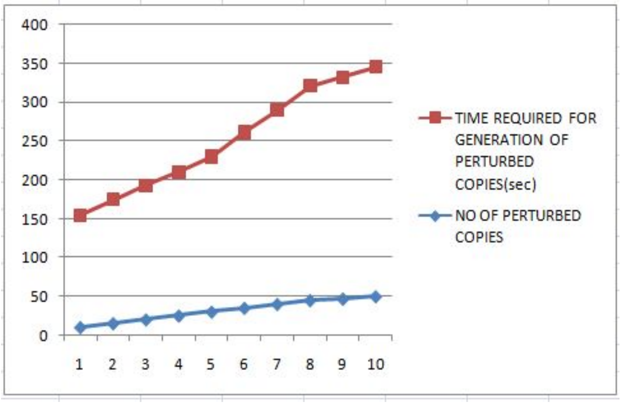
The Table II and Line chart II represents the time required for generation of perturbed copies



NO OF PERTURBED COPIES	TIME REQUIRED FOR 1000 RECORDS
10	145
15	160
20	173
25	185
30	200
35	226
40	250
45	276
47	285
50	295

LINE CHART II

TIME REQUIRED FOR PERTURBED COPY GENERATION



6. Acknowledgment

I wish to express my sincere thanks to guide Prof. Pankaj Agarkar and Head of Department Prof. Soumitra Das, as well as our principal Dr. S.S. Sonavne and at last but not least our Departmental staff members for their support.

7. Conclusion

This paper proposed the privacy preserving mechanism of sensitive data by generating the perturbed copies of original data. This paper gives the idea of additive perturbation technique and Group generation algorithm for generation of perturbed copies. Here different activities are to be performed on relational streaming data which is the new concept of this paper.

The main challenge is preventing the data miner from combining the perturbed copies of different trust levels to reconstruct the original data. This challenge is addressed by properly correlating the noise across copies at different trust levels. Also when the records are updated, at the same instance new perturbed copies for updated records are generated.

References

[1]Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang , “Enabling Multilevel trust in Privacy preserving data mining”, IEEE TRANSACTIONS ON KNOWLEDGE AND ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012.

[2] Samparthi V.S.Kumar, S.Sateesh Kumar, Nagaram Phani Kumar,” Joint Perturbed copies verification using Data Mining Techniques(Correlation)”, IJCST Vol. 4, Iss ue Spl - 4, Oct - Dec 2013.

[3] Lokesh Patel, Prof. Ravindra Gupta,” A Survey of Perturbation Technique For Privacy-Preserving of Data”, International Journal of Emerging Technology and Advanced Engineering, ijetae Volume 3, Issue 6, June 2013.

[4] D. Agrawal and C.C. Aggarwal, “On the Design and Quantification of Privacy Preserving Data Mining Algorithms,” Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS ’01), pp. 247-255, May 2001.

ISSN 2348 – 7968

[5] R. Agrawal and R. Srikant, “Privacy Preserving Data Mining,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’00),2000.

[6] K. Chen and L. Liu, “Privacy Preserving Data Classification with Rotation Perturbation,” Proc. IEEE Fifth Int’l Conf. Data Mining, 2005.

[7]Z. Huang, W. Du, and B. Chen, “Deriving Private Information From Randomized Data,” Proc. ACM SIGMOD Int’l Conf. Management of ata (SIGMOD), 2005.

[8] F. Li, J. Sun, S. Papadimitriou, G. Mihaila, and I. Stanoi, “Hiding in the Crowd: Privacy Preservation on Evolving Streams Through Correlation Tracking,” Proc. IEEE 23rd Int’l Conf. Data Eng. (ICDE), 2007.

[9] K. Liu, H. Kargupta, and J. Ryan, “Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 1, pp. 92-106, Jan. 2006.

[10]Y. Lindell and B. Pinkas, “Privacy Preserving Data Mining,” Proc. Int’l Cryptology Conf. (CRYPTO), 2000.

[11] O. Goldreich, “Secure Multi-Party Computation”, Final incomplete draft, version 1.4, 2002.

[12] C.C. Aggarwal and P.S. Yu, “A Condensation Approach to Privacy Preserving Data Mining”, Proc.Int’l Conf. Extending Database Technology (EDBT), 2004.

[13]E. Bertino, B.C. Ooi, Y. Yang, and R.H. Deng, “Privacy and Ownership Preserving of Outsourced Medical Data,” Proc. 21st Int’l Conf. Data Eng. (ICDE), 2005.

[14] D. Kifer and J.E. Gehrke, “Injecting Utility Into Anonymized Datasets,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 2006.

[15] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “L-Diversity: Privacy Beyond K-Anonymity,” Proc. Int’l Conf. Data Eng., 2006.

[16] L. Sweeney, “K-Anonymity: A Model for Protecting Privacy,” Int’l J. Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS), vol. 10, pp. 557-570, 2002

[17] Zahid Pervaiz, Walid G. Aref, Senior Member, IEEE, Arif Ghafoor, Fellow, IEEE, and Nagabhushana Prabhu, “Accuracy-Constrained Pivacy-Preserving Access Control Mechanism for Relational Data”

First Author Ashish .E. Mane, I did B.E.(CSE) Degree at T.K.I.E.T, Warananagar, Kolhapur. I am also pursuing my M.E(Computer Engg.) P.G. degree from Dr. D.Y. Patil School Of Engineering, Lohegaon, Pune.