Multiparty Privacy Preserving Data Mining for Vertically Partitioned Data

Rashmi Wandile[1], Neha Unavane[2], Yogesh Sangekar[3], Dhanraj Kachole[4]

[1] *Information Technology, Jayawantrao Sawant College of Engineering, Pune,Maharashtra, India*
[2] *Information Technology, Jayawantrao Sawant College of Engineering, Pune,Maharashtra, India*
[3] *Information Technology, Jayawantrao Sawant College of Engineering, Pune,Maharashtra, India*
[4] *Information Technology, Jayawantrao Sawant College of Engineering, Pune,Maharashtra, India*

ABSTRACT

The field of privacy pursues rapid advances in recent years because of the increases in the ability to store data. One of the most important topics in research community is Privacy preserving data mining (PPDM). Privacy preserving data mining has become increasingly popular because it allows sharing of privacy sensitive data for analysis purposes. People today have become well aware of the privacy intrusions of their sensitive data and are very reluctant to share their information. The major area of concern is that non-sensitive data even may deliver sensitive information, including personal information, facts or patterns. In this paper we demonstrate how the different departments of same organization combine their data without harming the privacy of the client. Then we use this data for making effective decisions in efficient and accurate manner. Data is said to be vertically partitioned when several organizations own different attributes of information for the same set of entities.

Keyword : - *Privacy preservation, Data mining, Randomization, Chaos Algorithm, K-means*

1. INTRODUCTION

Data mining is defined as a process used to extract usable data from a larger set of any raw data. Data can be mined whether it is stored in flat files, spreadsheets, database tables, or some other storage format. The important criteria for the data is not the storage format, but its applicability to the problem to be solved. Proper data cleansing and preparation are very important for data mining, and a data warehouse can facilitate these activities. However, a data warehouse will be of no use if it does not contain the data you need to solve your problem. Information about us that we feel is personal, confidential or private should not be unnecessarily distributed or publicly known. The problem with data mining is that with the availability of non-sensitive information, one is able to infer sensitive information that is not to be disclosed. Thus privacy is becoming an increasingly important issue in many data mining applications. This has led to the development of privacy preserving data mining. Data is said to be vertically partitioned when several organizations own different attributes of information for the same set of entities. Thus, vertical partitioning of data can formally be defined as follows: First, define a dataset D in terms of the entities for whom the data are collected and the information that is collected for each entity. Thus, $D \equiv (E, I)$, where E is the entity set for whom information is collected and I is the feature set that is collected. Assume that there are k different sites, $P_1,...,P_k$ collecting datasets $D_1 \equiv (E_1, I_1),...,D_k \equiv (E_k,I_k)$ respectively. Therefore, data is said to be vertically partitioned if $E = \bigcap_i E_i = E_1 \bigcap ... \bigcap E_k$, and $I = \bigcup_i I_i = I_1 \bigcup ... \bigcup I_k$. In general, distributed data can be arbitrarily partitioned. Vertical partitioning can also be defined as a special case of arbitrary partitioning, where all of the partitions consist of information about the same set of entities.

2. EXISTING SYSTEM

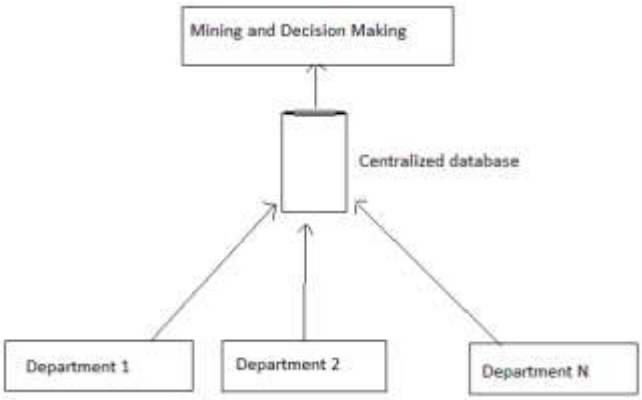

Fig 1: System environment

In multiparty privacy preserving data model different kinds of parties are participating. And for exposing the common knowledge of data attributes and mining of data is required. That environment can be understood using the fig.1. According to the given diagram the different departments are submitting their data in a centralized server. The server incorporates the data in same place with the associate classes. The centralized server uses the data for mining and extraction of patterns by which the decision making is performed. For example for an engineering student who studied five subjects namely data structure, digital electronics, network analysis and synthesis, computational algorithms, and computer graphics. Thus the student's marks and assessments are provided by three different departments of engineering electrical engineering, electronics engineering and computer science.

- Combine data on server
- Secure key generation
- Encryption and encoding
- Mining and decisions

3. PROPOSED SYSTEM

Fig2: Architecture of Privacy Preserving Data Mining

An integrated architecture takes a systematic view of the problems,implementing established protocols for data collection and information sharing. Multiple departments can upload their data with the security that one department could not learn the other departments data. And encrypted data would be uploaded on server,so that the security from the third party is also maintain. To encrypt the data we used different techniques like Randomization techniques and Random Perturbation. In randomization technique we use Chaos algorithm.

On the admin side admin have to decrypt the data first to analysis purpose, because we need the original data to do the analysis. K-means clustering when different sites contain different attributes for a common set of entities. Each site learns the cluster of each entity, but learns nothing about the attributes at other sites. And then we do the analysis and display the analysis in the form of graph.

4. PRIVACY PRESERVATION TECHNIQUES

4.1 Random perturbation

The random perturbation method attempts to preserve privacy of data by modifying values of the sensitive attributes using randomize techniques. They attempt to hide the sensitive data by randomly modifying the data values. In this approach, the owner of the dataset returns a value.

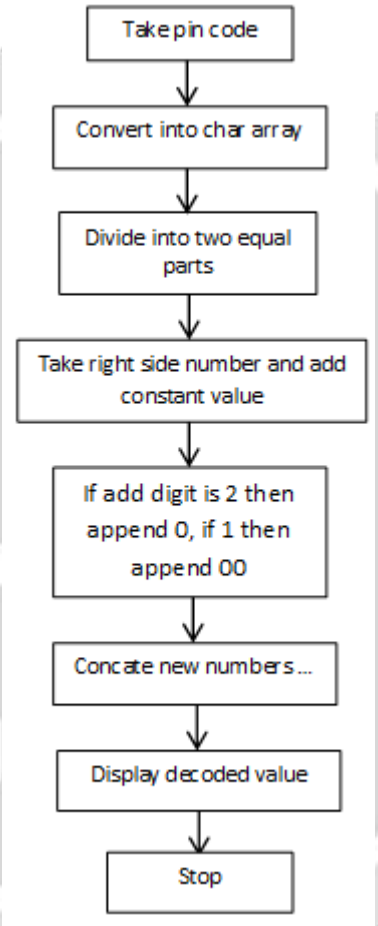

Fig3: flowchart of Random Perturbation

- **Steps for Encoding**

1. Take input data.
2. Convert it into char array.
3. Divide into two equal parts.
4. Take right side number & add constant value.
5. If new formed digit is 2 then append 0, if 1 then append ’00’ at start .
6. Concate new numbers by ***.
7. Display encoded value.
8. STOP.

- **Steps for Decoding**

1. Take input.
2. Convert it into char array.
3. Divide into two equal parts.
4. Take right side number & substract constant value.
5. If new formed digit is 2 then append 0, if 1 then append '00' at start which result in new number .
6. Concate key with new numbers .
7. Display encoded value
8. STOP

4.1 Chaos Algorithm

Chaos theory is an area of deterministic dynamics proposing that seemingly random events can result from normal equations because of the complexity of the systems involved. Chaos system like logistic mapping and Lorenz mapping designed for image encryption and researchers presented different encryption schemes based on chaos system]. Chaos system process has various features like high sensitivity to initial state, certainty, ergodic and etc. chaos sequence which are random sequences are generated by chaos mapping. These structures are very complex and their analysis and prediction is too difficult for encryption. In this system chaos theory is applied on numeric data. Example age column, we shuffled the age values using some method so after encryption we can't see the actual age of respective person.

- **Steps for chaos algorithm**

1. Retrieve column from corresponding table
2. Initialize variables

   MAX_ITER=3,a=1.4,b=0.3

   y0=0,y1,x0=0.1,x1=0
3. For x=0 to maxlenght
4. x1 = 1 + y0 - (a * x0 * x0);

   y1 = b * x0;

   shuffle[i] = (int) ((x1 + 1.3) * maxLength);
5. if (shuffle[i] < 0) set shuffle[i] = 0;

   if (shuffle[i] >= maxLength set shuffle[i] = maxLength - 1;
6. update for next iteration

   x0 = x1; y0 = y1;

   and goto step 3
7. for k=0 to MAX_Iter-1

   for i=0 to maxlenght-1

   shufflePixelsEncryption(imageMathObject.addIndex(shuffle[i], 0, maxLength), (new ImageMath()).addIndex(shuffle[i], i + maxLength / 2 + 1, maxLength));

   shufflePixelsEncryption(imageMathObject.addIndex(shuffle[i], 1, maxLength), (new ImageMath()).addIndex(shuffle[i], i + maxLength / 4 + 1, maxLength));

   shufflePixelsEncryption(imageMathObject.addIndex(shuffle[i], 2, maxLength), (new ImageMath()).addIndex(shuffle[i] , i + (3 * maxLength / 4) + 1, maxLength));

8.End.

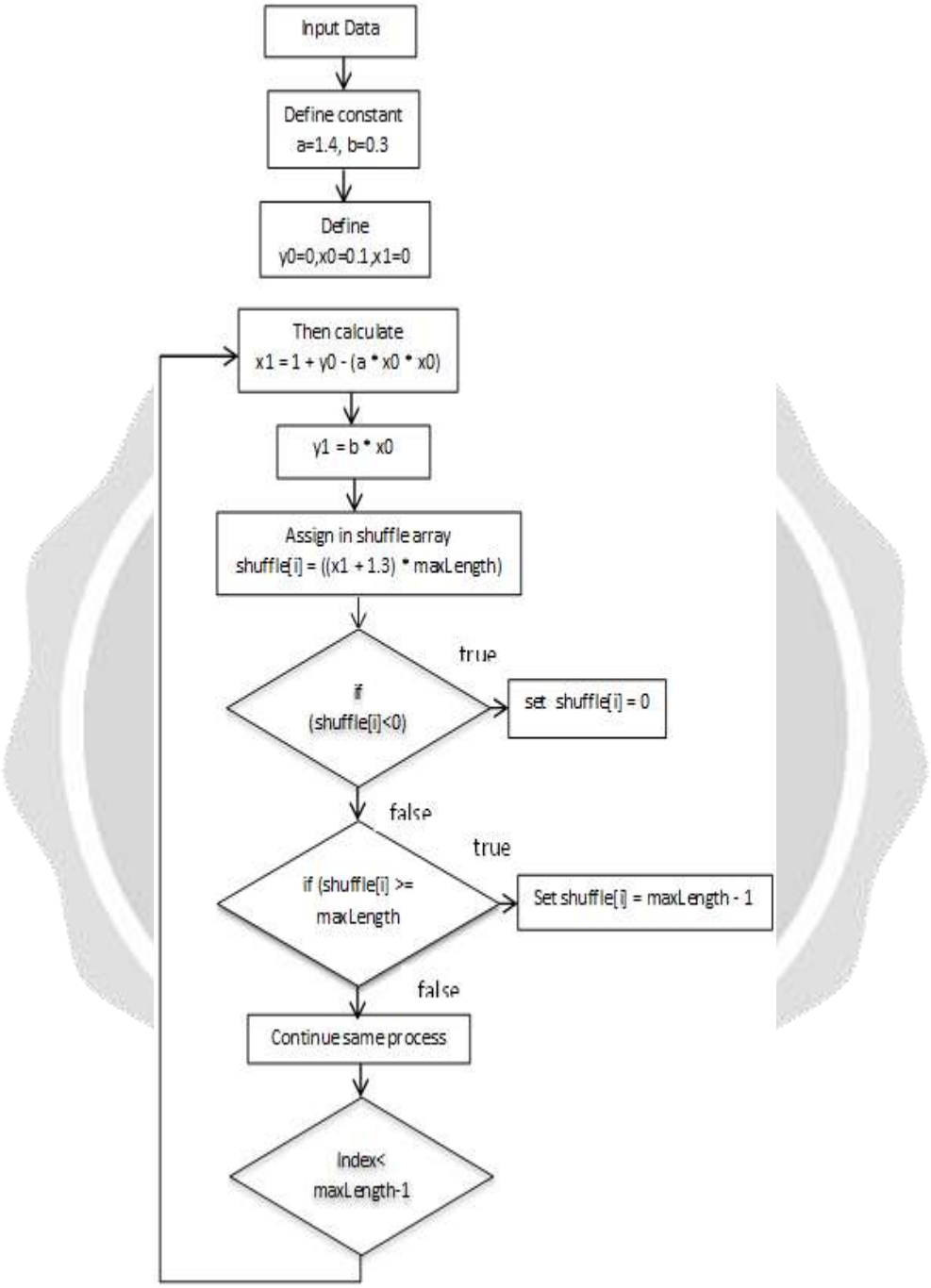

Fig4 : flowchart of chaos algorithm

4.3 K-means Algorithm

The purpose of the k-means algorithm is to cluster the data. K-means algorithm is one of the simplest partitions clustering methods. K-Means is the unsupervised learning algorithm for clusters. Grouping of pixels is done according to the same characteristics. In the k-means algorithm initially, we have to define the number of clusters k. Then k-cluster center is chosen randomly. The distance between the each pixel to each cluster centers is calculated. The distance may be of simple Euclidean function. A single pixel is compared to all cluster centers using the

distance formula. The pixel is moved to the particular cluster which has the shortest distance among all. Then the centroid is re-estimated. Again each pixel is compare to all centroids. The process continuous until the center converges.

- **Steps for K- means Algorithm:**

1. Give the no of cluster value as 'k'.
2. Randomly choose the 'k' cluster centers
3. Calculate mean or center of the cluster
4. Calculate the distance between each pixel to each cluster center
5. If the distance is close enough to the center then move to that cluster.
6. Otherwise, move to next cluster.
7. Re-estimate the center.

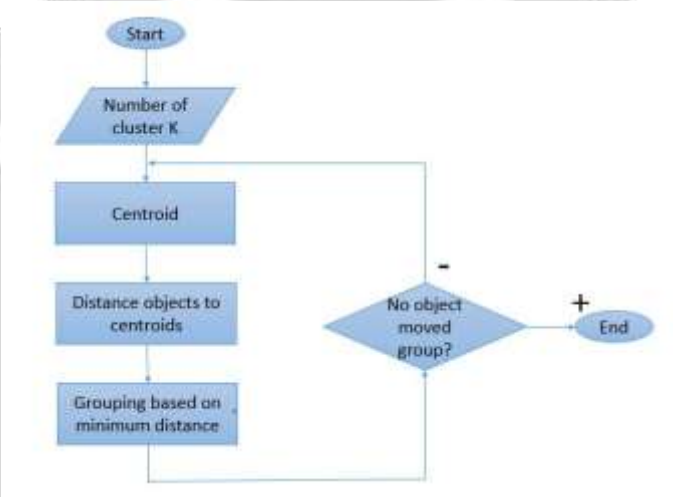

Fig 5 : flowchart of K- means algorithm

5. APPLICATIONS

- **Banking domain**

Data mining can contribute to solving business problems in banking and finance by finding patterns causalities and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts.

- **Educational**

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicating students future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning.

- **E-commerce**

Many E-commerce companies use Data Mining and Business intelligence to offer cross-sells and up-sells through the websites. One of the most famous of these is of course Amazon, who use sophisticated mining techniques to drive their, "People who viewed that product, also liked this" functionality.

- **Processing sales and marketing data**

Data mining is used for market analysis to provide information on what product combinations were purchased together when they were bought and in what sequence. This information helps businesses promote their most profitable products and maximize the profit.

6. LITURATURE SURVEY

We have studied some papers those paper focus on various approaches implement by the miners for preserving of information. A detail description with limitation and strength of different techniques of privacy preserving is explained. (Data perturbation, noise addition, data swapping, aggregation, suppression).

The models of privacy preserving will be discussed. Trust Third Party Model, Semi-honest Model, Malicious Model. Also discuss the survey of privacy preserving techniques such as Randomization method, Anonymization method and Encryption method. Also presents a brief survey on various standard techniques for privacy preserving data mining was presented namely: Classification, Clustering and Associated rule mining. Some paper reviews main PPDM techniques based on a PPDM framework. We compare the advantages and disadvantages of different PPDM techniques and discuss open issues and future research trends in PPDM. It proposed an efficient approach for privacy preservation in data mining. This technique protects the sensitive data with less information loss which increase data usability and also prevent the sensitive data for various types of attack.

7. CONCLUSIONS

The proposed work is intended to find the solution for privacy preserving data mining technique in efficient and accurate manner. Therefore a number of research articles are studied and a multiparty privacy preserving data mining technique is proposed. The proposed technique is able to combine multiparty vertically partitioned data securely. Then this data is encrypted to maintain the security. And analysis is done after the applying decoding of on respective data. The implementation of the proposed technique is performed using the JAVA technology and their performance in terms of accuracy, error rate, time consumption and memory used.

8. REFERENCES

[1] Dhanalakshmi, M., and E. Siva Sankari. "Privacy preserving data mining techniques-survey."Information Communication and Embedded Systems (ICICES), 2014 International Conference on. IEEE, 2014.

[2] K.Saranya, K.Premalatha, S.S.Rajasekar, . " A Survey on Privacy Preserving Data Mining." International Journal of Innovations & Advancement in Computer Science 2015,IEEE,2015.

[3] Manish Sharma,Atul Chaudhary,Manish Mathuria,Shalini Chaudhary,Santosh Kumar. “An Efficient Approach for Privacy Preserving in Data Mining" 2014 International Conference on. IEEE, 2014.

[4] Xueyun Li,Zheng Yan,Peng Zhang “A Review on Privacy-Preserving Data Mining" International Conference on Computer & Information Technology 2014,IEEE,2014.

[5] Bhavna Vishwakarma,Huma Gupta, Manish Manoria “A Survey on Privacy Preserving Mining Implementing Techniques“ Symposium on Colossal Data Analysis & Networking(CDAN), 2016 ,IEEE, 2014.