A Hybrid Approach for Privacy Preserving using Randomization for Data mining

[1]*Halak P. Patel*
Department of computer science & engineering
Parul institute of Technology, Vadodara, Gujarat, India

[2]*Warish D. Patel*
Department of computer science & engineering
Parul institute of Technology, Vadodara, Gujarat, India

ABSTRACT

Many organizations large amount of data are collected. These data are further used by the organizations for the analysis purposes which help gaining useful knowledge. The data collected may contain private or sensitive information which should be protected. Privacy protection is an important issue if we release data for the mining or sharing purpose. Our technique protects the sensitive data with less information loss which increase data usability and also prevent the sensitive data for various types of attack. Data can also be reconstructed using our proposed technique. A novel hybrid method to achieve k-support anonymity based on statistical observations on the datasets. Our comprehensive experiments on real as well as synthetic datasets show that our techniques are effective and provide moderate privacy. A hybrid approach for used to improved security and accuracy to private data. Our novel hybrid approach towards privacy preserving k-anonymity and artificial neural network techniques are effective, scalable and no information loss.

Keywords: - *Data Mining, Privacy preserving, k-anonymity, artificial neural network.*

1. INTRODUCTION

Data mining and knowledge discovery is the process of getting important and valuable information that has been unknown in the raw data previously.it is a process of capturing knowledge with the help of computer. Data mining is exploration of large dataset to extract hidden and previously unknown patterns, relationships and knowledge which are not easy to detect with traditional statistics. Researchers are trying to obtain satisfactory result in reasonable time with help of searching techniques because many problem are difficult to be solved in feasible time by analytically[1] .The use of Varity of analysis tool to determine the relationship between data and database ,and used to same generate group of data. It is new technology, developing with database and artificial intelligence. Data mining is large number of incomplete, noise, fuzzy, random, practical application of the data in hidden, regulating, people not known in advance, but is potentially useful and ultimately untestable information and knowledge of novelty.

Many organizations like credit card companies, real estate companies, search engines, hospitals collects and hold large volume of data. The data are further used by the data miner for the analysis purpose which helps the organizations for gaining useful knowledge [2]. These data may contain sensitive or valuable information of any individuals, organizations such as hospitals contain medical records of the patients, and they provide these database or records to the researchers or data miner for the purpose of research. Data miner analyzes the medical records to gain useful global health statistics. However, in this process the data miner may able to obtain sensitive information and in combination with an external database may try to obtain personal attribute of an individual.

An interesting new direction in the field of data mining has been emerged known as privacy preserving in data mining (PPDM).privacy preserving is extraction of useful knowledge from the large amount of data. Privacy preserving data mining techniques divided into two broad area:
Data hiding - The data hiding are modification or removable of confidential information from the data before disclose to other.knowlege hiding - The knowledge hiding are hiding the sensitive knowledge which can be mined from the database using any data mining algorithms.

Privacy preservation one of the most important issues in data mining. The privacy preserving mining methods modify the original data in some other form, so that the privacy of the user data is preserved and at the same time the mining models can be reconstructed. The modified data with reasonably accuracy. Various approaches have been proposed in the existing literature for privacy-preserving data mining which modified with some other values. Several techniques used for the privacy preservation in data mining. Non-cryptographic techniques contain k-anonymity, I-diversity-closeness, and perturbation and association rule. The problems with non-cryptographic Method are information loss. On the other hand cryptographic method provides accurate results but it suffers from high computation and communication cost. In this paper, we focus on non-cryptographic techniques.

2. RELATED WORK

The literature reviews are to get depth knowledge of the basics of privacy preserving data mining. It is necessary to identity the various approaches and techniques that could be possibly used to preserve the sensitive data. The objective of literature review are to identity existing privacy preserving techniques, its advantages and disadvantages to find efficient approach for preserving private or sensitive data.

The data owner using a one-to-n mapping encrypts the original transaction database to be sent to the third-party miner. The data miner performs the data mining task on the encrypted database and hands over the association rules to the data owner. The data owner then extracts the original rules from the encrypted rules by decryption. These methodologies have been identified in by proposing a frequency analysis based method for breaking the encryption scheme [4]. The main flaw in the former work was that the fake items added in the encrypted database were independent of other items.

Gianotti et al [5] adopt a frequency-based attack model where the adversary knows the exact set of items along with the item support. Even in this work, the approach of k-support anonymity named differently by the authors as k-privacy is undertaken. The work considers adding fake transactions containing of real items for achieving k-privacy.

Providing privacy-preserving for Internet data is a longstanding goal of the computer research community. It is has received considerable attention with the development of data mining and network Technology. Cryptograph based privacy-preserving method can provide a better guarantee of the privacy when different institute want to cooperate in a common goal[4].Noise addition using random perturbation technique has been explored in [2]. However since it uses random perturbation technique therefore even though privacy is obtained, data mining are affected since noise are added randomly without identifying the data characteristics.

T- Closeness model [1] uses the k-anonymity and 1 diversity approach but in addition ensures that that the distances between the distributions of the sensitive attribute within an anonymized group should not be different from the global distribution by more than a threshold t. Compared to the previous methods. This model provides better privacy but there is information gain for the attacker and data characteristics are also lost.

In [7] R. Agrawal and R.Srikant started the work towards PPDM. They categorized the methods as perturbation and secure multi-party computation. Many variation of algorithm has been suggested like database extension geometric data.

X. Xiao, Tao, and M. Chen introduces the multiple version of the dataset and anonymized data set at different privacy levels.G. Wang. z.zhu, W. DU, and Z.Teng proposed maximum Entropy Principle to perform interface analtsis for disguised dataset[8].

Lea Kissner and Dawn Song propose efficient techniques for privacy-preserving operations on multi sets using cryptosystem. In [8], the authors proposed a security protocol for the IVC applications based on group signature and ID-based signature schemes.

3. PROPOSED METHODOGY

From the literature review it is concluded that many algorithms that are already work for increase the accuracy and security. The proposed work privacy preservation KNN (k-anonymity) and ANN (artificial neural network) using MAES (modified advance encryption standards) to improve security and accuracy. Using artificial neural network try to improve get high quality of dataset and result. Using modified advance encryption standards improve security using shift row method compare to advance encryption standards. The base learning and testing approach is required to also improve the quality. The proposed worked improved quality and security both. Modified-AES algorithm is a fast encryption algorithm for security of multimedia data. The ANN with crab classifier. The crab classifier depend characteristics of the sensitive attribute. The crab classifier improved performance of sensitive attribute. The main purpose is to protect sensitive information. Apply hybrid approach makes difficult for the attacker to identify various type of attack.

The Neural networks have proven themselves as proficient classifiers and are particularly well suited for addressing non-linear problems. Given the non-linear nature of real world phenomena, like crab classification, neural networks is certainly a good candidate for solving the problem. The physical characterstics will act as inputs to a neural network and the sex of the crab will be target. Given an input, which constitutes the observed values for the physical characterstics of a crab, the neural network is expected to identify if the crab is male or female. This is achieved by presenting previously recorded inputs to a neural network and then tuning it to produce the desired target outputs. This process is called neural network training.

The Neural network is ready to be trained. The samples are automatically divided into training, validation and test sets. The training set is used to teach the network. Training continues as long as the network continues improving on the validation set. The test set provides a completely independent measure of network accuracy.

The following Step of proposed work:-

- Read dataset to select quasi identifier, key attribute and sensitive attribute from given dataset. For prefix dataset select identifies and attributes, which is used for set prefix data as an input data. Remove key attribute.
- Generate and select probability matrix for desire data set. To choose particular matrix for given dataset. The modified advance encryption standard using shift row method.
- After shifting apply k-anonymity for quasi identifier. The k-anonymity suppressed and generalization to probability matrix.
- Then apply artificial neural network techniques to improved accuracy and efficiency for the given dataset. The artificial neural network is based on learning and testing dataset. Using ANN with crab classifier is used to sensitive attribute. The Crab classifiers are comparing to actual data set. Apply this operation modified the original dataset value.
- Now, recombine & rearrange element for desired output and integrate both value and matrix using classifier. Combine element and finally get output.

The proposed flow diagram as shown in fig..

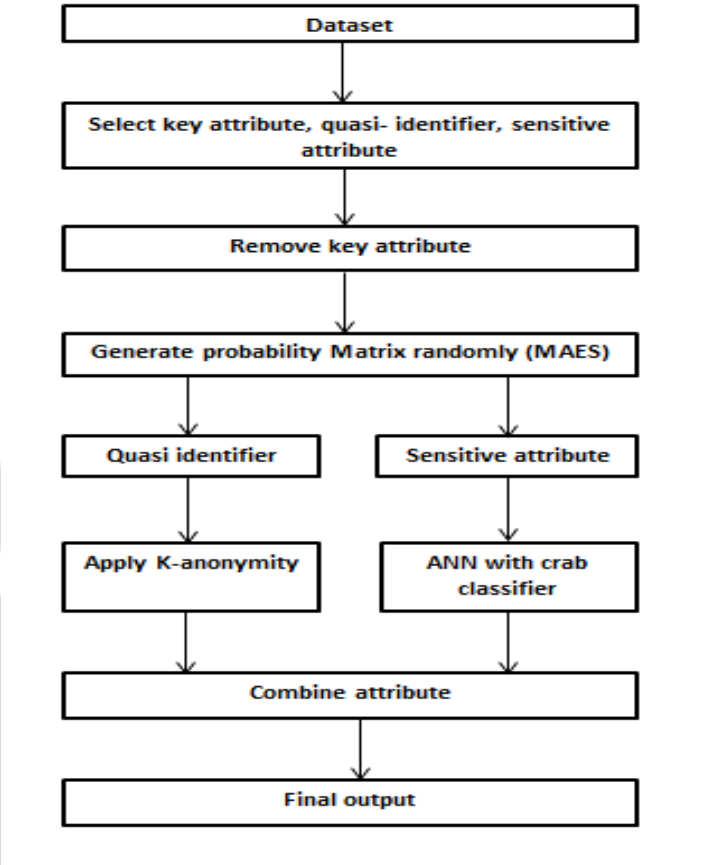

Fig .1 flow chart of proposed system

4. EXPRIMENTAL RESULT

The proposed experimental result are implemented in metlab.The attribute value are generating identified and compare value. The matrix values are shifting to modified advance encryption techniques. The generalization and suppression is matrix based on k-anonymity. The proposed hybrid approaches are scalable and better protect private information to proposed techniques.

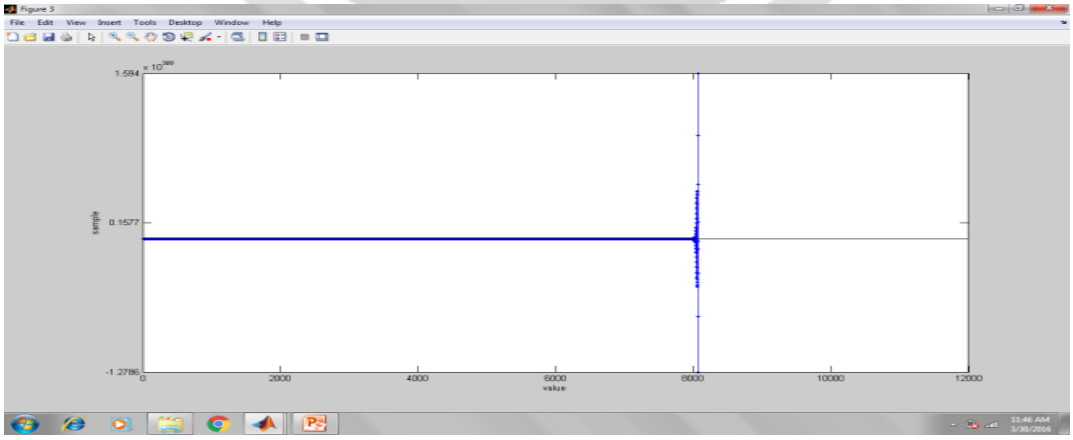

Fig.2 compares and identified attribute

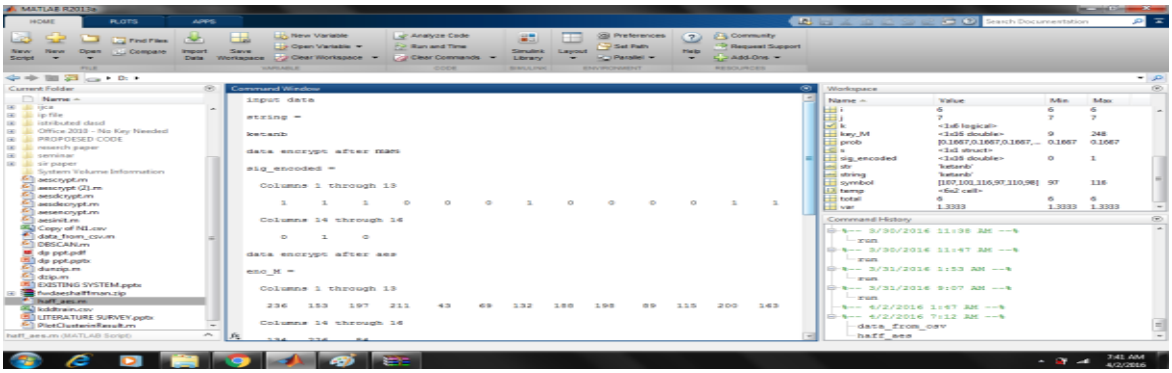

Fig.3 Modified value of matrix

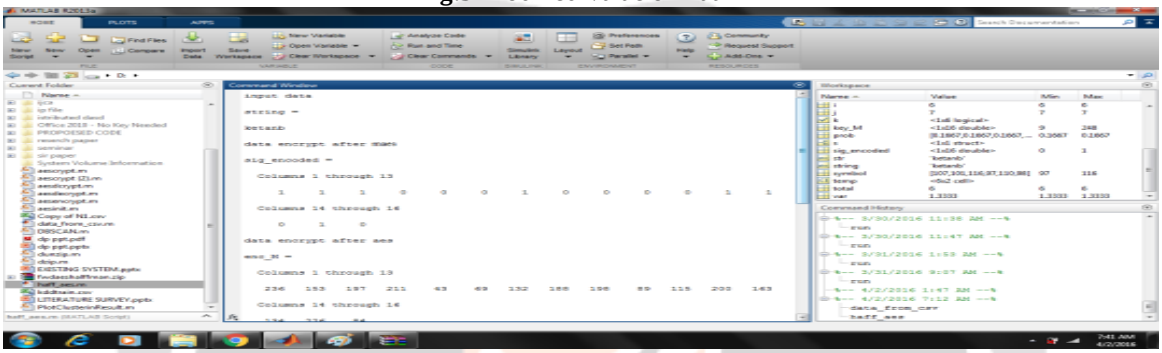

Fig.4 Modified value of matrix

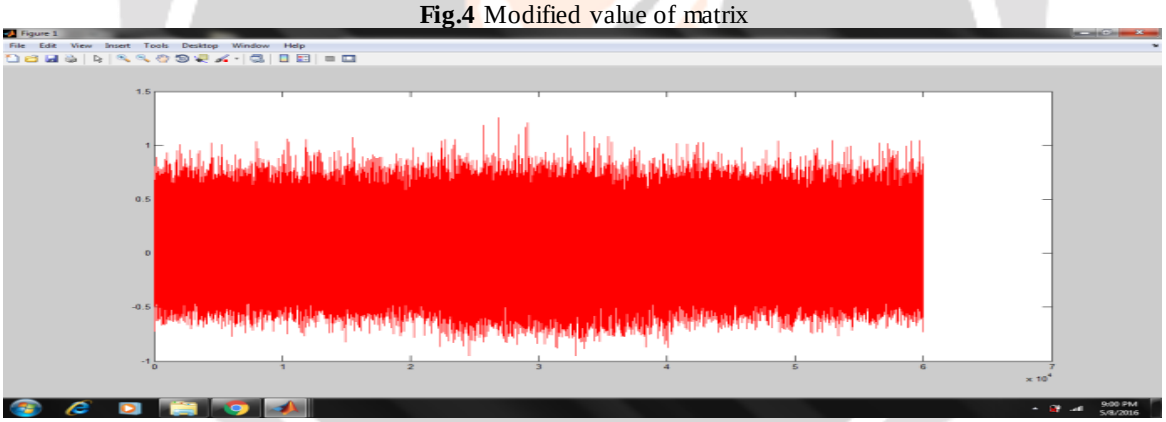

Fig.5 Modified value of attribute

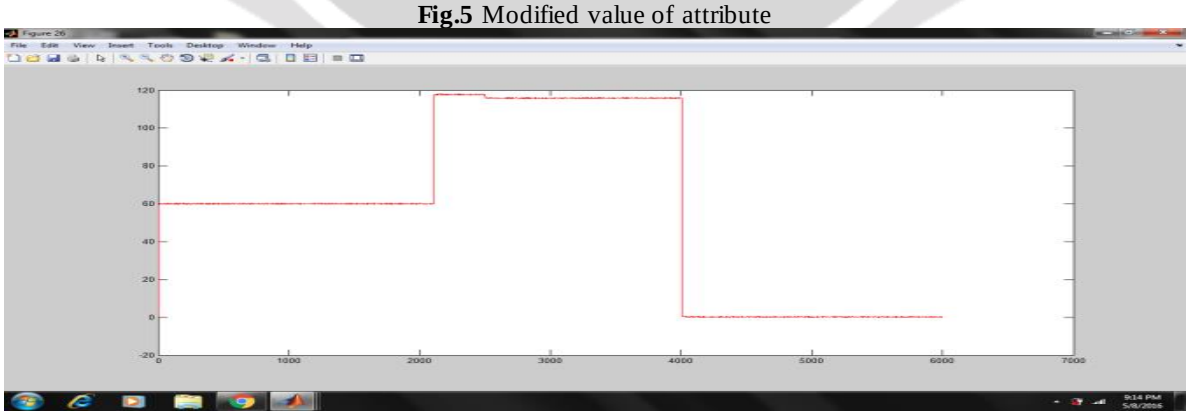

Fig.6 Combine Element

CONCLUSION

In this paper, privacy preservation search scheme to enable accurate, efficient and secure search over encrypted private data. By using Randomization technique attacker cannot identify a pattern of data. K-anonymity method has shortcoming of homogeneity and background attack. Using hybrid approach It makes difficult for the attacker to identify background and homogeneity attack. Apart from that it protects private data with better accuracy and gives no loss of information which increases data utility. Data can also be reconstructed by our proposed approach. using artificial neural network more efficient for the dataset.

REFERENCES

[1]"An Efficient Approach for Privacy Preserving in Data" Mining"Manish Shannal Atul Chaudhar/ Manish Mathuria3 Shalini Chaudhar/ Santosh Kumar5, IEEE , 2014
[2] "Privacy Preservation Algorithm in Data Mining for CRM Systems"Shashidhar Virupaksha, G Sahoo, Ananthasayanam Vasudevan,2014, IEEE
[3] "Dataless Data Mining: Association Rules-based Distributed Privacy-preserving Data Mining" Vikas G. Ashok, IEEE 2015
[4] "Privacy-Preserving Frequent Itemset Mining in Outsourced Transaction Databases" Iyer Chandrasekharan P.K. Baruah , 2015, IEEE
[5] **"**SYMMETRIC-KEY BASED PRIVACYPRESERVING SCHEME FOR MINING SUPPORT COUNTS", Yu Li1 AND SHENG ZHONG, IEEE, 2013
[6] **"**k-anonymity security preserving crime data publishing in resource constrain environment" , Mark-Johan burke, Anne V.D.M Kayem,IEEE,2014
[7] "Data Anonymization using Augmented Rotation of sub-cluster for Privacy preservation in data mining"V.Rajalakshmi, G.S.Anandha Mala,IEEE,2013
[8] "A Combine Random noise perturbation approach for multi-level privacy preservation in data mining" Mr.S.Chidambaram, Research Scholar/NEC,Dr.k.g Srinivasagam/CSE/NEC,IEEE,2014
[9] "Analysis of privacy preserving k-anonymity method and techniques", S.Vijayarani, A.Tamilarasi,M. Sampoorna, ICCCI, 2010
[10] "Enabling multilevel trust in privacy preserving data mining," Yaping li,Minghua Chen,Qiwei Li and Wei Zhang ,IEEE ,YoI.24,No.9,Sep 2012.
[11] "An Efficient Conjunctive Query scheme over encrypted multidimensional data in smart grid"(GLOBECOM), IEEE, 2013