Volume 5, Issue 11, November 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Quasi & Sensitive Attribute Based Perturbation Technique for Privacy Preservation

[1]Neha Patel[*], [2]Prof. Shrikant Lade, [3]Prof. Ravindra Kumar Gupta
[1]Research Scholar, [2]HOD,
[1, 2, 3] Department of Computer Science & Engineering, RKDF IST,
RGPV University, Bhopal, India

Abstract— Privacy preservation is an important application of data mining techniques to provide security and efficiency of data security. Data distortion is a major component for privacy preservation in security-related data mining applications. This article proposes a perturbation based PDM technique for data distortion, and compare it with the existing method. The experimental results show that the perturbation based PDM technique is better in comparison of existing technique.

Keywords— Privacy, Quasi, Perturbation, Sensitive, Privacy Preservation.

I. INTRODUCTION

Privacy preservation have an important role in the field of data mining because Huge amounts of data has been collected in many organizations for this data privacy preservation must be necessary. These collected data may be used by the many of the organizations for performing data mining tasks. However, private and sensitive information of collected data should be protected. If any organization release or share their data then preservation of privacy is an important issue. Data mining privacy preservation techniques allows publishing data for the mining purpose with preserving private information of the organization. There are various privacy preservation techniques available for sensitive and confidential data but they all faces many types of attacks .Privacy preserving data mining (PPDM) has become increasingly popular because it allows to privacy of Sensitive data for analysis purposes. Various data mining algorithms (DMA), used for achieving privacy preserving mechanisms, have been invented that allow one to extract relevant knowledge from bulk amount of data. [2].Privacy preserving Data Mining(PPDM) is a latest research field of data mining(DM) by this we can provides privacy to the organizations as well as others who needs privacy of data in sharing and communication for data mining data mining purpose. [3]

Privacy Preserving Data Mining (PPDM) is focus on how to develop an algorithm for preserving original data, so that the knowledge regarding data should be private after the process of mining [1]. In data mining (DM), users are provided with the data but not the association rules and they are free for using their own tools; So, it is mandatory to applied privacy constraints on the data before the mining phase. So we need to develop a Architecture that helps to implement privacy control systems. This system is helpful for converting existing database into new version of database. But the way in which the general rules should be mined from the existing database. Through sanitization process existing database can be convert into new database with concept of hiding of sensitive information.

Privacy preserving data mining (PPDM) has become very curtail problem in recent years, because of the large amount of internet users track their data by automated systems. Many of the customers have used the E-comm. feature of internet for this huge amount of storage required for storing transactional and personal information about the users and advance hardware technology also help to make it successful to track information about transaction of individual user in daily life.

II. PRIVACY PRESERVATION

The aims of Data mining (DM) are to take out useful information from multiple sources, whereas the aim of privacy preservation in data mining is to preserve these data without loss of private and sensitive information. Privacy preserving data mining (PPDM) is a new research area of data mining and statistical databases [3], Various types of enhanced data mining algorithms easily achieve data privacy. Through these two constraints privacy preservation can be possible.

1) Raw sensitive data information like identifiers, name, and addresses should be modified in original database.
2) By using data mining algorithm Sensitive knowledge may be mined from a database because knowledge has an important role for privacy of data.

The main goal of privacy preserving data mining (PPDM) is to develop algorithms for modifying the original data in some way, so that the private data and the private knowledge should stay private even after the mining process. Privacy preserving data mining algorithms (PPDMA) which allow consumers to collaborate in the extraction of knowledge, without any party having to reveal individual items or data.

III. PRIVACY PRESERVING TECHNIQUES

Privacy preserving data mining (PPDM) is focus on how to develop such kind of data mining methods without increasing the risk factor of mishandling of the data used to generate those methods. Many of the techniques use some form of alteration on the main data for achieving privacy preservation. This altered dataset is useful for mining and must meet privacy requirements without losing the benefit of mining. There are some privacy preservation techniques.[2]

A. Randomization
B. Anonymization
C. Secure multi-party computation
D. Sequential pattern hiding

A. Randomization

Randomization technique is a cheap and best approach for privacy preserving data mining (PPDM). Randomization process should be implemented for assuring the performance of data mining as well as preserving privacy. This approach protects the customers' data by letting them arbitrarily alter their records before sharing, taking away some correct information and introducing some noise.

B. Anonymization

For protection of individuals' identity at the time of releasing sensitive information, sometimes data holders encrypt or remove explicit identifiers, like names and unique security numbers. However, encrypted data provides no guarantee for anonymity.

C. Sequential pattern hiding

Sequential pattern hiding method is necessary in hiding sensitive patterns that can otherwise be extracted from published data, without critically affecting the data and the non sensitive interesting patterns. Sequential pattern hiding is a challenging problem, because sequences have more composite semantics than item sets, and calls for better solutions that gives high utility.

D. Secure multi-party computation

A substitute approach depends upon the multiparty computation. That offers each part of private data is validly known to more than one parties. Revealing private data to parties such as by whom the data is owned or the individual to whom the data refers to is not a condition of violating privacy. The problem arises when the private information is revealed to some other third parties.

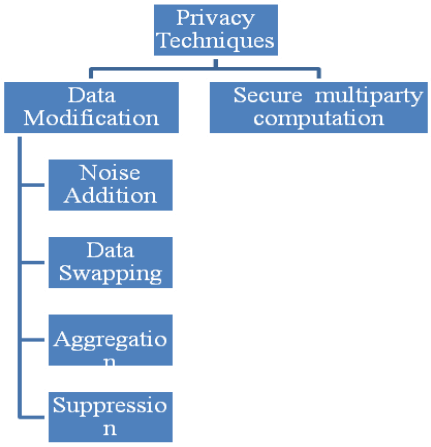


Fig 1 Classification of Privacy preserving techniques

IV. PRIVACY V/S SECURITY

Certain confusion exists in the industry people about security and privacy. Some people think that security and privacy is same, while others think that privacy means some information will be hidden for some once. Security is an important tool for privacy. Security and privacy both are slightly similar technologies; however, there are important differences are:

1) Developer needs to understand these two technologies at the time of developing new system.
2) Developer should have ability to understand what information is going and what is coming from data base.

A. The Relationship between Data Security and Data Privacy:

Companies have made data security policy for ensuring data privacy of their customer's information. Data privacy is very important aspect for the companies for this they provides security of their customers data because this data is the asset. A data security policy is simply the means to the desired end, which is data privacy. However, no data security policy can overcome the willing sell or soliciting of the consumer data that was entrusted to an organization.

B. Classification of various Privacy Preservation Techniques:

Here some Privacy preserving are listed with their advantages and disadvantages [8].

1) Perturbation Technique

This technique used for the privacy preservation in which data are perturbed But this cannot reconstruct the original data and also not good for the large data.

2) Condensation Technique

Instead of perturbed data, it works on the pseudo data. So it provide better privacy preservation than the techniques which use simply data modification on original data. But it does not give longer effect on data mining. Because it has the same format as the original data.

3) Cryptographic Technique:

It performs encryption of the sensitive data. There is also proper toolset for algorithm in the field of the data mining But this technique is difficult to scale when more parties are involved and also not good for large database.

4) Blocking Based Technique:

In this technique to provide privacy to the individual it replaces the unknown values to the sensitive transaction. Reconstruction of the original data is quite difficult

C. Combine strategy for the privacy preservation

In the combine strategy multiple strategies are used to obtain any privacy preserving methodology. In the given figure combine strategy is shown based on the data transformation and data encryption techniques. Due to the combining this various techniques robust security can be obtained. Sometime privacy preserving techniques may have some disadvantages or some limitations but that can be overcome in combine strategy. So security result would be more effective of combine strategy than a single privacy preserving techniques used [8]. In the given figure original data are transformed after that transformed data are encrypted. So here data transformation and data encryption methods are used in this combine strategy.

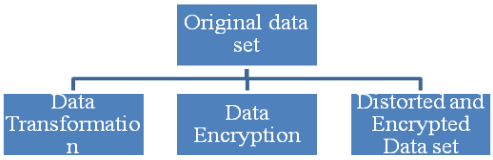


Fig 2: PPDM Methodology

V. LITERATURE REVIEW / RELATED WORK

Privacy preserving data mining techniques provides how to preserve private and sensitive information when anyone publishing their data for mining purpose. While at the same time preserve the private information of the individuals. Many of the privacy preservation techniques have been proposed for mining purpose but they fights from various types security attacks .These are harmful for information. Paper [1] proposed an efficient approach for privacy preservation in data mining. 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. The Base paper approach uses the combined techniques of randomization and k-anonymization. Base paper approach contains three main advantages:

1) It protects private data with low information loss.
2) Utility of data is increased.
3) Data can also be reconstructed.

Mainly Base paper approach is describe two algorithms for achieving privacy preservation algorithm I randomization is performed on dataset using attribute transitional probability matrix and in algorithm II anonymity is performed on randomized dataset which is result.

In paper [2] there are many future research directions for privacy preserving data mining. First, present studies tend to use different terminology to describe similar or related practice. For instance, people used data modification, data perturbation, data sanitation, data hiding, and pre-processing as possible methods for preserving privacy; however, all are in fact related to the use of some types of technique to modify original data so that private data and knowledge remain private even after the mining process. Lacking a common language for discussions will cause misunderstanding and slow down the research breakthrough. Therefore, there is an emerging need of standardizing the terminology and PPDM practice.

Second, most prior PPDM Algorithms were developed for use with data stored in a centralized database. However, in today's global digital environment, data is often stored in different sites. With recent advances in information and communication technologies, the distributed PPDM methodology may have a wider application, especially in medical, health care, banking, military and supply chain scenarios. Third, data hiding techniques have been the dominated methods for protecting privacy of individual mining results, which may lead to sensitive rules leakages. While some algorithms are designed for preserving the rule such as with sensitive information, it may degrade the accuracy of other non-sensitive rules.

Paper [8] explores the possibility of using multiplicative random projection matrices for privacy preserving distributed data mining. It specifically considers the problem of computing statistical aggregates like the inner product matrix, correlation coefficient matrix, and Euclidean distance matrix from distributed privacy sensitive data possibly owned by multiple parties. This class of problems is directly related to many other data-mining problems such as clustering, principal component analysis, and classification. Paper [8] makes primary contributions on two different grounds.

First, it explores Independent Component Analysis as a possible tool for breaching privacy in deterministic multiplicative perturbation-based models such as random orthogonal transformation and random rotation. Then, it proposes an approximate random projection-based technique to improve the level of privacy protection while still preserving certain statistical characteristics of the data. Paper[10] proposed the concept of Fast Fourier Transform (FFT) based data distortion method and compare its performance with Singular Value Decomposition (SVD) based distortion method. Result analysis of this paper shows that FFT based method is similar to SVD based method in preserving privacy and keeping utility of dataset. However, the computational time used by the FFT based method is much less than the SVD based method. With the help of result analysis of paper[9] proves that the FFT based method is a very efficient data distortion method.

Paper [10] shows that general and efficient distributed privacy preserving knowledge discovery is truly feasible. Paper[10] considered the security and privacy aspects when dealing with distributed data that is partitioned either horizontally or vertically across multiple sites, and the challenges of performing data mining tasks on such data. Since RDTs can be used to generate equivalent, accurate and sometimes better models with much smaller cost, this paper proposed distributed privacy-preserving RDTs and its approach leverages the fact that randomness in structure can provide strong privacy with less computation. The results of paper [10] show that the privacy preserving version of the RDT algorithm scales linearly with dataset size, and requires significantly less time than alternative cryptographic approaches.

VI. PROPOSED WORK

This article is all about maintaining the privacy of the data. This work gives emphasis on the perturbation of the data. While maintaining the privacy of the data, propose work divided the data into two categories. One is called Quasi and another is calling Sensitive data.

The Architecture of the proposed work is shown in figure 1.

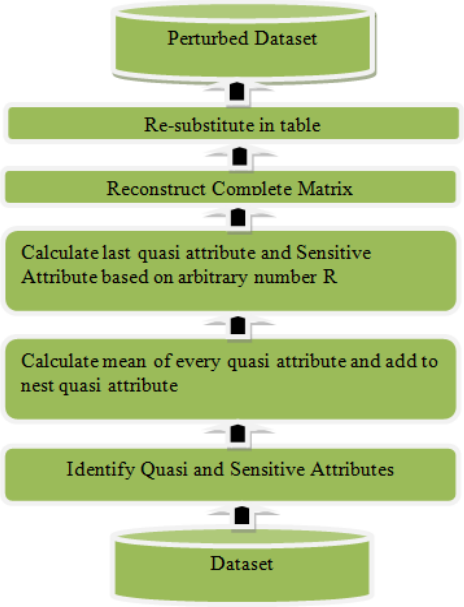


Fig 3: Architecture of proposed work

Fig 4 are shows the algorithm of the proposed work below:

INPUT - Original dataset T.
OUTPUT - Converted dataset P.
METHOD –
Step 1: Select the Quasi column Qij and Sensitive column Sij.
Step 2: Preprocessing Data.
Step 3: (A)
For (j=1; j<n; j ++)
Qj + Mj+1.
Here j: - column, Mj+1:- Mean of next column and n:- Length of Quasi table.
(B)
Qn + R.
Sij+ R.
Here R: - A Random value and On: - Last column of Quasi table.
Step 4: Recombine T matrix.
Step 5:Re-substitute in table.
Step 6: Stop.

Fig 4: Proposed Algorithm

VII. RESULT ANALYSIS

The system used for execution of the data perturbation based Privacy Preservation method with existing methods is as follows:

All the experiments were conducted on a PC, Dual-Core (2.20 GHz), with 2GB of RAM, running a Windows7 operating system and 32-bits operating system.

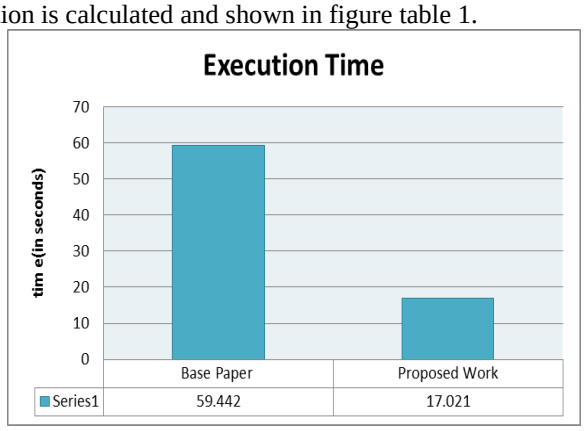
Time complexity

The effectiveness is measured in terms of the time requirement in execution of the existing work with proposed work. It compares between execution timing of existing method with proposed method.

Table 1: Execution time comparison between proposed works with existing work (in seconds)

BASE PAPER	PROPOSED WORK
59.44200	17.02100

Average of ten times of execution is calculated and shown in figure table 1.



Graph 1: Execution time comparison between existing and Proposed Work.

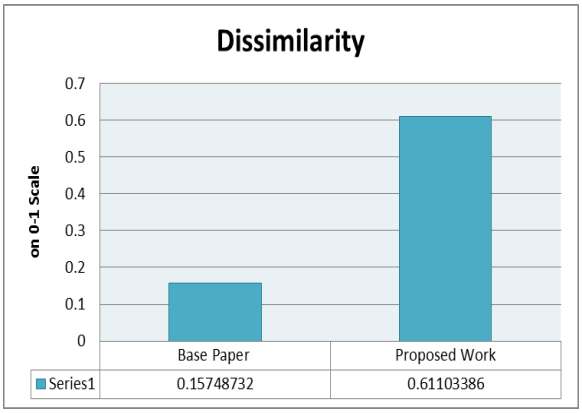
Dissimilarity-Privacy Preservation

The effectiveness is measured in terms of the Dissimilarity of both datasets with reference to existing work with proposed work. It compares between dissimilarity of dataset of existing method with proposed method.

Table 2: Dissimilarity comparison between proposed works with existing work

BASE PAPER	PROPOSED WORK
0.15748732	0.61103386

Average of ten times of execution is calculated and shown in figure table 2.



Graph 2: Dissimilarity-Privacy Preservation comparison between existing and Proposed Work.

VIII. CONCLUSION

In present scenario privacy preserving in data mining is very important topic of research. Literature review of this paper clears that there are many privacy preserving techniques available in data mining but still they have some disadvantages. Anonymity technique gives privacy protection and usability of data but it suffers from homogeneity and background attack. Blocking method suffers after analysis of Table 1 and 2, which is output of the proposed method for execution time and dissimilarity is father better.

By applying various operations on perturbed dataset the improved result of proposed method over base paper method. These improvements are in following direction:

1) Reduce Time Complexity
2) Improve Dissimilarity Result.

REFERENCES

[1] Manish Sharma, AtulChaudhary, Manish Mathuria, ShaliniChaudhary and Santosh Kumar, *An Efficient Approach for Privacy Preserving in Data Mining*, International Conference on Signal Propagation and Computer Technology (ICSPCT) IEEE 2014, pp 244-249.

[2] Tamanna Kachwala and Sweta Parmar, *An Approach for Preserving Privacy in Data Mining*, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)2014 pp. 370-373.

[3] Anbazhagan, Dr. R. Sugumar, M. Mahendran and R. Natarajan, *An Efficient Approach for Statistical Anonymization Techniques for Privacy Preserving Data Mining*, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 7, September 2012, pp. 482-485.

[4] M. Mahendran, Dr. R. Sugumar, K. Anbazhagan and R. Natarajan, *An Efficient Algorithm for Privacy Preserving Data MiningUsing Heuristic Approach*, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 9, November 2012, pp. 737-744.

[5] Jaideep Vaidya and Chris Clifton, *Privacy Preserving Association Rule Mining in Vertically Partitioned Data*, SIGKDD '02 Edmonton, Alberta, Canada 2002, ACM 158113567X.

[6] Murat Kantarcioglu and Chris Clifton, *Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data*, IEEE Transactions on Knowledge and Data Engineering, 2003, pp 01-21.

[7] Sarvaiya Sukhdev and Hemant Vasava, *Privacy Preserving Data Mining With Classification And Encryption Methods,* International Journal of Innovative and Emerging Research in Engineering Volume 2, Issue 5, 2015 pp. 19-23.

[8] Kun Liu, Hillol Kargupta, and Jessica Ryan, *Random Projection-Based Multiplicative Data Perturbation for Privacy PreservingDistributed Data Mining*, IEEE transactions on knowledge and data engineering, vol. 18, no. 1, January 2006, pp. 92-106.

[9] Shuting Xu, and Shuhua Lai, *Fast Fourier Transform Based Data Perturbation Method for Privacy Protection*, 1-4244-1330-3/072007 IEEE, pp. 222-225.

[10] Jaideep Vaidya, Basit Shafiq, Wei Fan, Danish Mehmood, and David Lorenzi, *A Random Decision Tree Framework forPrivacy-preserving Data Mining*, Journal of latex class files, vol. 6, no. 1, January 2007, pp. 1-14.

