A Review on 'Privacy Preservation Data Mining'

Shivangi Modhiya[1], **Jay Vala**[2], **Prem Balani**[3]

[1] *Student / M.Tech,Information Technology, GCET ,Gujarat, India*

[2] *Assistant Professor, Information Technology, GCET ,Gujarat, India*

[3] *Assistant Professor, Information Technology, GCET ,Gujarat, India*

ABSTRACT

Privacy Preserving Data Mining (PPDM) is used to mine the potential valuable knowledge without revealing the personal information of the individuals. Now a days due to data privacy has become a major concern as quasi and sensitive attribute in terms of privacy needs a lot of research existing researches on privacy of quasi and sensitive attribute are time consuming and also are not space efficient so by this research to trying different techniques used in various papers for establish that which method is more convenient. It is often highly valuable for organizations to have their data analyzed by external agents.. In the era of information society, sharing and publishing data has been a common practice for their wealth of opportunities. However, the process of data collection and data distribution may lead to disclosure of their privacy. Privacy is necessary to conceal private information before it is shared, exchanged or published. PPDM has thus has received a significant amount of attention in the research literature in the recent years. Various methods have been proposed to achieve the expected goal. In this paper we have given a brief discussion on different dimensions of classification of privacy preservation techniques.

Keyword : *Data Mining, Cryptography, Randomization, Quasi Attribute, Sensitive Attribute, perturbation data mining, privacy preservation, Slicing.*

1. INTRODUCTION

Now a days to preserve the privacy of the data is important aspect and to save them we need the research this paper has the comparison and conclusion of the various privacy preserving techniques for quasi and sensitive attribute. Data mining is the extraction of huge sensational examples or learning from huge measure of information. PPDP thinks about how to change crude Data into a form that is protected against protection assaults however that still backings powerful information mining assignments. Protection saving for the two Data mining (PPDM) and Data distributing (PPDP) has turned out to be progressively well known in light of the fact that it permits sharing of security delicate Data for examination purposes. Data mining aims to extract useful information from multiple sources, whereas privacy preservation in data mining aims to preserve these data against disclosure or loss. The main consideration of the privacy preserving data mining is two-fold. First, sensitive raw data like identifiers, name, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy [1]. The main objective of privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and the private knowledge remain private even after the mining process. In this paper, we provide a classification and description of the various techniques and methodologies that have been developed in the area of privacy preserving data mining. The goal of this paper is to give an review of the different dimension and classification of privacy preservation techniques used in privacy preserving data mining. Also aim is to give different data mining algorithms used in PPDM and related research in this field. [5].

2. DIFFERENT TECHNIQUES OF PRIVACY PRESERVING DATA MINING

Different types of privacy preservation techniques are used. They are mainly classified into following categories: Anonymization based, Randomized Response based, Condensation approach based, Perturbation based, Cryptography based and Elliptic Curve Cryptographic based.

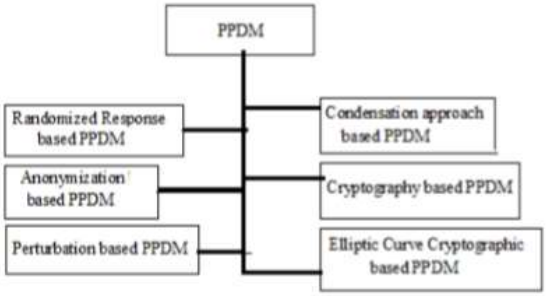

Fig -1: Techniques of PPDM

2.1 PAPER-1

Rail fence algorithm and Vigenere cipher algorithm have been implemented. These algorithms are initially tested using numerical data. Though there might not be a significant reduction in execution time by removing the noise generation operation for small number of records (say 100 records),. On comparison of the obtained results of both the algorithms, modified Vigenere Cipher has more complex procedures than modified Rail fence algorithm. On the contrary the encryption of records with the modified Vigenere cipher produces highly efficient results than the modified Rail fence algorithm. in the future this modified technique has to be tried with categorical data**.**

2.2 PAPER-2

Introduce techniques which publish more than one table for organizations preserving individual's privacy. One of this is (α, k) – anonymity using lossy-Join which releases two tables for publishing in such a way that the privacy protection for (α, k)-anonymity can be achieved with less distortion, and the other one is Anatomy technique which releases all the quasi-identifier and sensitive values directly in two separate tables. Future work can include defining a new privacy technique for multiple sensitive attributes and researchers will focus to publish attributes without suppression using generalization boundaries technique that used to achieve k-anonymity maintaining individual privacy without influence data utility.

2.3 PAPER-3

Paper present an extensive survey of PPDM techniques, their classification and give a preliminary implication of technique to be used under specific scenarios.

Techniques	Methods Employed	Scenarios	Data Mining Tasks			
			Classification	Clustering	Association Rule Mining	Outlier detection
Anonymization	Generalization Suppression, Permutation	Central Commodity	✓	✓	✓	✓
Condensation	Aggregation	Central Commodity	✓			
SMC	Cryptographic	Distributed	✓	✓	✓	✓
Pseudonymization	Cryptographic	Distributed	✓	✓	✓	✓
Perturbation	Adding Noise, Data Swapping, Global recoding, Microaggregation	Both	✓	✓	✓	✓
Randomization	Adding Noise, Scrambling, Resampling	Both	✓			
Fuzzy based	Clustering, Microaggregation, c-regression	Central Commodity	✓	✓	✓	
Neural Network Based	Bayesian Network, Probabilistic NN	Central Commodity	✓	✓	✓	✓

Fig -2 : Comparison of PPDM Techniques

2.4 PAPER-4

Extends work the efficiency of random perturbation techniques using additive noise for privacy-preserving data mining in continuous valued domain and presents new results in the discrete domain shows that the growing collection of random perturbation-based "privacy-preserving" data mining techniques may need a careful scrutiny in order to prevent privacy breaches through linear transformations. It showed that under certain conditions it is relatively easy to breach the privacy protection offered by the random perturbation based techniques. It provided extensive experimental results with different types of data and showed that this is really a concern that we must address.

2.5 PAPER-5

The results in this paper analysis are developing a realistic anonymization framework for health data sharing. anonymization must support both the complexity of health data and frequency of change in complex transactional systems. This paper tries to answer the question anonymization frameworks for high-dimensional streaming health data? Paper concluded that 1) Impossible the values of all quasi-identifiers in most high dimensional datasets.2) Clustering and micro-aggregation are good techniques for anonymizing high-dimensional datasets. But they result in excessive information loss when there are too many quasi identifiers. 3) multi-dimensional data in their natural forms suffer from dimensionality curse. One of the recommended solutions is to transform this data to fewer and lower dimensions before anonymization.4) Anonymization for health data sharing cannot be solved by technology alone. There is need to investigate regulatory frameworks that can complement technology to achieve the right mix of solution that allows health care organization to share data more efficiently and effectively.

2.6 PAPER-6

Sensitivity Based Tuple Anonymity Method. In this method first consider the sensitivity of values in sensitive attribute and then only tuples having sensitive values are generalized, and the other tuples can be directly published. Experiment results on the Adult Database show the proposed methods not only can improve the accuracy of the publishing data, but also can preserve privacy.

2.7 PAPER-7

Show that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. paper show how slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the ℓ-diversity requirement.in workload experiments confirm that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. these experiments also demonstrate that slicing can be used to prevent membership disclosure.

2.8 PAPER-8

The various models have been designed for privacy preserving data sharing. In this paper, Randomization Method, Anonymization Method, Encryption Method For Distributed Privacy Preserving Data Mining are describe.These three secure data sharing directions of privacy preserving approaches by analyzing the existing work and develop the new model in the future.1) The personalized privacy preservation will become the issue. 2) How to improve the efficiency of implementation and ensure availability of the result in order to meet the various requirements. 3) The analyzation consists of how to combine the advantage of the above approaches.

2.9 PAPER-9

Discussed about various approaches such as anonymization, ℓ-Diversity, t-closeness, perturbation, Randomized response, Distributed K -Anonymity framework (DKA) , Slicing and techniques used in privacy preservation of data mining. Most privacy attacks can be effectively destroyed by the advanced techniques and approaches. In distributed privacy preserving data mining areas, efficiency is an essential issue. Privacy and accuracy is the pair of contradiction.

2.10 PAPER-10

They describe different method of PPDM and conclude there does not exists a single privacy preserving data mining algorithm that out performs all other algorithms on all possible criteria like performance, utility, cost, complexity, tolerance against data mining algorithms.

2.11 PAPER-11

Proposes a multi attribute statistical anonymization method. This algorithm processes the data, before the microdata are published. After processing the data, it sends the anonymous data to service providers. Then service provider incorporates the supporting information to analyze the anonymous data, so that knowledge can be mined from the cloud. Paper shows Comparison of the query accurateness between different anonymization methods and Comparisons of association discovery among different anonymization methods.

2.12 PAPER-12

This paper proposes and shows how to use microaggregation to generate k-anonymous t-close data sets using three algorithms. Microaggregation is a family of perturbative methods for statistical disclosure control of microdata releases. This divide in two steps Partition and Aggregation.

2.13 PAPER-13

has highlighted and exploited several connections between k-anonymity, t-closeness and ε-differential privacy.t-closeness and ε-differential privacy are strongly related to one another when it comes to anonymizing data sets.in this paper show that k-Anonymity, t-Closeness, ε-Differential privacy are performed. From differential privacy to (stochastic) t-closeness is best result gives also ε-Differential privacy through t- closeness is perfomed.also shows that A bucketization construction to attain t-closeness.

2.14 PAPER-14

Proposed new technique called shearing based composite transformation. Data owner transforms the original data into distorted data by shearing based composite data transformation. Only this distorted data is given to the clients. For clustering used k-means algorithm and from experiments found that the total number of elements in the clusters is same with the original and distorted data.

2.15 PAPER-15

Presented a survey of the different techniques of privacy-preservation in data publishing. Paper discussed a variety of data modification techniques such as randomization and k-anonymity based techniques. Also, methods for distributed privacy-preserving mining, and the methods for handling horizontally and vertically partitioned data. Also discuss the number of diverse application domains for which privacy- preserving data mining methods are useful.

2.16 PAPER-16

Proposes a perturbation based PDM technique for data distortion. Existing work 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.

2.17 PAPER-17

Homogeneous anonymization that anonymizes quasi attributes by choosing a single sensitive attribute. This approach causes high information loss and reduces the data utility. To overcome these issues in the existing system, sensitive attribute based non-homogeneous anonymization system is proposed. Based on the sensitive attribute, on-homogeneous anonymization technique (generalization and suppression) is applied to the identified quasi attributes and the non-sensitive attributes are directly published. Thus the proposed system achieves high degree of data utility, reduces information loss and also achieves high degree of Data Integrity.

3. CONCLUSIONS

Several algorithms have been proposed to do knowledge discovery, while providing guarantees on the non-disclosure of data. In this paper we have given a brief discussion on different dimensions of classification of

privacy preservation techniques. We have also discussed different privacy preservation techniques We also discuss some of the popular data mining algorithms used to privacy preservation technique.

4. REFERENCES

1. N.Abitha, G Sarada,Manikandan, Sairam.N. "A Cryptographic Approach for Achieving Privacy in Data Mining." IEEE, 2015: 0-1.
2. al, Abou_el_ela Abdo Hussein et. "Multiple-Published Tables Privacy-Preserving Data Mining: A Survey for Multiple-Published Tables Techniques." International Journal of Advanced Computer Science and Applications, 2015: 80-85.
3. Alpa Shah, Ravi Gulati. "Privacy Preserving Data Mining: Techniques, Classification and Implications - A Survey." International Journal of Computer Applications, 2016: 40-46.
4. Haimonti Dutta, Hillol Kargupta,Souptik Datta. "Analysis Of Privacy Preserving Random Perturbation Techniques: Further Explorations." 2003: 31-38.
5. Benjamin Ezea, Liam Peytona. "Systematic Literature Review on the Anonymization of High Dimensional Streaming Datasets for Health Data Sharing." Information and Communication Technologies in Healthcare, 2015: 1.
6. Bhavana Abad, Kinariwala S.A. "A Novel approach for Privacy Preserving in Medical Data Mining using Sensitivity based anonymity." International Journal of Computer Applications , 2012: 13-16.
7. CH. Srikanth Reddy, MD.John Saida. "A Novel Approach to Privacy Preserving Data Publishing Using Slicing Technique." International Journal of Research Studies in Science, Engineering and Technology, 2014: 55-63.
8. Ms. Dhanalakshmi.M, Mrs.Siva Sankari.E. "Privacy Preserving Data Mining Techniques-Survey." IEEE, 2014: 1.
9. Dhivakar K, Mohana S. "A Survey on Privacy Preservation Recent Approaches and Techniques." International Journal of Innovative Research in Computer and Communication Engineering, 2014: 6559-6566.
10. Hina Vaghashia, Amit Ganatra. "A Survey: Privacy Preservation Techniques in Data Mining." International Journal of Computer Applications, 2015: 20-26.
11. V. K. Saxena, Shashank Pushkar. "Protecting Sensitive Knowledge with Effective Privacy Preservation." International Journal of Applied Engineering Research, 2016: 4049-4052.
12. Jordi Soria-Comas, Josep Domingo-Ferrer,David S´anchez,Sergio Mart´ınez. "t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation." IEEE, 2015: 1.
13. Josep Domingo-Ferrer, Jordi Soria-Comas. "From t-Closeness to Differential Privacy and vice versa in Data Anonymization." 2015: 2-20.
14. Manikandan.G, Dr.N.Sairam. "Shearing Based Data Transformation Approach For Privacy Preserving Clustering." IEEE, 2012: 1.
15. Nargis Nasir Sheikh, Prof. S. A. Bhavsar. "Techniques of Privacy Preservation in data publishing." International Journal of Emerging Technology and Advanced Engineering, 2014: 188-193.
16. Neha Patel, Shrikant Lade,Ravindra Kumar Gupta. "Quasi & Sensitive Attribute Based Perturbation Technique for Privacy Preservation." International Journal of Advanced Research in Computer Science and Software Engineering, 2015: 450-455.
17. S.Sathishkumar, P.Usha R.Shriram. "Sensitive Attribute based Non-Homogeneous Anonymization for Privacy Preserving Data Mining." IEEE, 2014: 1.