A Survey: Privacy Preserving Techniques in Data Stream Mining

Ankit Jasoliya[1], **Tejal Patel**[2]
[1]Department of Information & Technology PIET, Parul University, Vadodara, India.
[2]Assistant Professor ,Department of Information & Technology PIET, Parul University, Vadodara, India.
Email: ankitjess@gmail.com, Tejal.Patel@paruluniversity.ac.in

Abstract: *Data mining gets valuable knowledge from huge amounts of data. In latest, data streams are new type of data, which are completely different from existing static data. The characteristics of data streams are: Data has timing preference; data distribution changes constantly with time; the amount of data is large; Data flows in and out quickly; and immediate reply is necessary. Existing algorithm is designed for the static database. If the data changes, it would be compulsory to rescan the whole dataset, which takes to more computation time and providing late respond to the user. The problem of privacy-preserving data mining has widely been studied and many techniques have been find. However, existing techniques for privacy-preserving data mining are designed for static databases and are not suitable for dynamic data. When need to perform computation at that time to providing privacy also. So the privacy preservation problem of data streams mining is very big issue. The success of privacy preserving data stream mining algorithms is measured in terms of its accuracy, performance, data utility, level of uncertainty or resistance to data mining algorithms etc. However no privacy preserving algorithm exists that outperforms all others on all possible criteria. Rather, an algorithm may perform better than another on one specific criterion. So, the aim of this paper is to present current scenario of privacy preserving data stream mining framework and techniques.*

Keywords: *Anonymization, Condensation, Cryptography, Distributed Data Mining, Perturbation, Privacy Preserving Data Mining (PPDM), Randomized Response.*

1. INTRODUCTION:

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyse this huge amount of data and extract useful information from it. Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation [11].

Information is today probably the most important and demanded resource. We live in an internet worked society that relies on the dissemination and sharing of information in the private as well as in the public and governmental sectors. Governmental, public, and private institutions are increasingly required to make their data electronically available. So here need to protect the privacy of the respondents (individuals, organizations, associations, business establishments, and so on) [7].

A data stream is a sequence of unbounded, real time data items with a very high data rate that can only read once by an application . Imagine a satellite-mounted remote sensor that is constantly generating data. The data are massive (e.g. terabytes in volume), temporally ordered, fast changing, and potentially infinite. These features cause challenging problems in data streams field. Data Stream mining refers to informational structure extraction as models and patterns from continuous data streams. Data Streams have different challenges in many aspects, such as computational, storage, querying and mining [1].

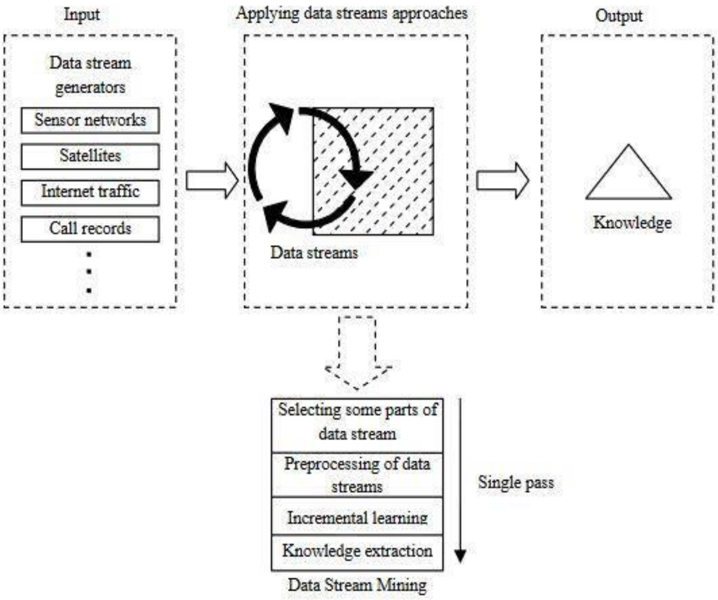

Figure 1 : General Process of Data Stream Mining

2. PPDM TECHNIQUES:

2.1 Anonymization based PPDM

The basic form of the data in a table consists of following four types of attributes:

(i) Explicit Identifiers is a set of attributes containing information that identifies a record owner explicitly such as name, SS number etc.

(ii) Quasi Identifiers is a set of attributes that could potentially identify a record owner when combined with publicly available data.

(iii) Sensitive Attributes is a set of attributes that contains sensitive person specific information such as disease, salary etc.

(iv) Non-Sensitive Attributes is a set of attributes that creates no problem if revealed even to untrustworthy parties [3].

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden. It even assumes that sensitive data should be retained for analysis. It's obvious that explicit identifiers should be removed but still there is a danger of privacy intrusion when quasi identifiers are linked to publicly available data. Such attacks are called as linking attacks. For example attributes such as DOB, Sex, Race, and Zip are available in public records such as voter list.

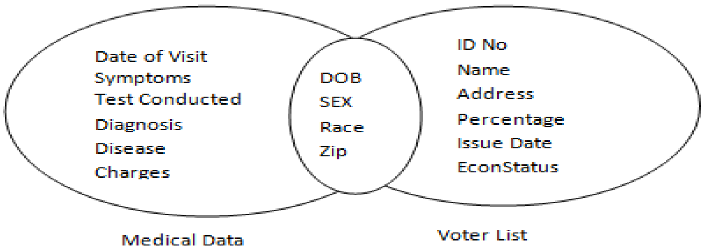

Figure 2: Linking Attack

Such records are available in medical records also, when linked, can be used to infer the identity of the corresponding individual with high probability as shown in Figure.2.

Sensitive data in medical record is disease or even medication prescribed. The quasi-identifiers like DOB, Sex, Race, Zip etc. are available in medical records and also in voter list that is publicly available. The explicit identifiers like Name, SS number etc. have been removed from the medical records.

. Still, identity of individual can be predicted with higher probability. Sweeney [8] proposed k-anonymity model using generalization and suppression to achieve k-anonymity i.e. any individual is distinguishable from at least k-1 other ones with respect to quasi-identifier attribute in the anonymized dataset. In other words, we can outline a table as k-anonymous if the Q1 values of each raw are equivalent to those of at least k- 1 other rows. Replacing a value with less specific but semantically consistent value is called as generalization and suppression involves blocking the values. Releasing such data for mining reduces the risk of identification when combined with publically available data. But, at the same time, accuracy of the applications on the transformed data is reduced. A number of algorithms have been proposed to implement k-anonymity using generalization and suppression in recent years.

Although the anonymization method ensures that the transformed data is true but suffers heavy information loss. Moreover it is not immune to homogeneity attack and background knowledge attack practically [9]. Limitations of the k-anonymity model stem from the two conventions. First, it may be very tough for the owner of a database to decide which of the attributes are available or which are not available in external tables. The second limitation is that the k-anonymity model adopts a certain method of attack, while in real situations; there is no reason why the attacker should not try other methods. However, as a research direction, k-anonymity in combination with other privacy preserving methods needs to be investigated for detecting and even blocking k-anonymity violations.

2.2 Perturbation Based PPDM

Perturbation being used in statistical disclosure control as it has an intrinsic property of simplicity, efficiency and ability to reserve statistical information. In perturbation the original values are changed with some synthetic data values so that the statistical information computed from the perturbed data does not differ from the statistical information computed from the original data to a larger extent. The perturbed data records do not agree to real-world record holders, so the attacker cannot perform the thoughtful linkages or recover sensitive knowledge from the available data. Perturbation can be done by using additive noise or data swapping or synthetic data generation.

In the perturbation approach any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. Relevant information for data mining algorithms such as classification remains hidden in inter-attribute correlations. This is because the perturbation approach treats different attributes independently. Hence the distribution based data mining algorithms have an intrinsic disadvantage of loss of hidden information available in multidimensional records. Another branch of privacy preserving data mining that manages the disadvantages of perturbation approach is cryptographic techniques.

2.3 Randomized Response Based PPDM

Basically, randomized response is statistical technique introduced by Warner to solve a survey problem. In Randomized response, the data is twisted in such a way that the central place cannot say with chances better than a predefined threshold, whether the data from a customer contains correct information or incorrect information. The information received by each single user is twisted and if the number of users is large, the aggregate information of these users can be estimated with good quantity of accuracy. This is very valuable for decision-tree classification. It is based on combined values of a dataset, somewhat individual data items. The data collection process in randomization method is carried out using two steps [8]. During first step, the data providers randomize their data and transfer the randomized data to the data receiver. In second step, the data receiver rebuilds the original distribution of the data by using a distribution reconstruction algorithm. The randomization response model is shown in Figure.3.

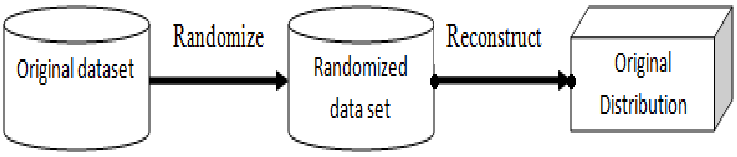

Figure 3: Randomization Response Model

Randomization method is relatively very simple and does not require knowledge of the distribution of other records in the data. Hence, the randomization method can be implemented at data collection time. It does not require a trusted server to contain the entire original records in order to perform the anonymization process [10]. The weakness of a randomization response based PPDM technique is that it treats all the records equal irrespective of their local density. These indicate to a problem where the outlier records become more subject to oppositional attacks as compared to records in more compressed regions in the data [5]. One key to this is to be uselessly adding noise to all the records in the data. But, it reduces the utility of the data for mining purposes as the reconstructed distribution may not yield results in conformity of the purpose of data mining.

2.4 Condensation approach based PPDM

Condensation approach constructs constrained clusters in dataset and then generates pseudo data from the statistics of these clusters . It is called as condensation because of its approach of using condensed statistics of the clusters to generate pseudo data. It creates sets of dissimilar size from the data, such that it is sure that each record lies in a set whose size is at least alike to its anonymity level. Advanced, pseudo data are generated from each set so as to create a synthetic data set with the same aggregate distribution as the unique data. This approach can be effectively used for the classification problem. The use of pseudo-data provides an additional layer of protection, as it becomes difficult to perform adversarial attacks on synthetic data. Moreover, the aggregate behaviour of the data is preserved, making it useful for a variety of data mining problems [10]. This method helps in better privacy preservation as compared to other techniques as it uses pseudo data rather than modified data. Moreover, it works even without redesigning data mining algorithms since the pseudo data has the same format as that of the original data. It is very effective in case of data stream problems where the data is highly dynamic. At the same time, data mining results get affected as huge amount of information is released because of the compression of a larger number of records into a single statistical group entity [6].

2.5 Cryptography Based PPDM

Consider a scenario where multiple medical institutions wish to conduct a joint research for some mutual benefits without revealing unnecessary information. In this scenario, research regarding symptoms, diagnosis and medication based on various parameters is to be conducted and at the same time privacy of the individuals is to be protected. Such scenarios are referred to as distributed computing scenarios .The parties involved in mining of such tasks can be mutual untrusted parties, competitors; therefore protecting privacy becomes a major concern. Cryptographic techniques are ideally meant for such scenarios where multiple parties collaborate to compute results or share non sensitive mining results and thereby avoiding disclosure of sensitive information. Cryptographic techniques find its utility in such scenarios because of two reasons: First, it offers a well-defined model for privacy that includes methods for proving and quantifying it. Second a vast set of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms are available in this domain. The data may be distributed among different collaborators vertically or horizontally [2].

All these methods are almost based on a special encryption protocol known as Secure Multiparty Computation (SMC) technology. SMC used in distributed privacy preserving data mining consists of a set of secure sub protocols that are used in horizontally and vertically partitioned data: secure sum, secure set union, secure size of intersection and scalar product. Although cryptographic techniques ensure that the transformed data is exact and secure but this approach fails to deliver when more than a few parties are involved. Moreover, the data mining results may breach the privacy of individual records. There exist a good number of solutions in case of semi-honest models but in case of malicious models very less studies have been made [4][12].

3. EVALUATION OF PRIVACY PRESERVING:

An introductory list of evaluation parameters to be used for evaluating the quality of privacy preserving data mining algorithms is given below:

(i) Performance: the performance of a mining algorithm is measured in terms of the time required to achieve the privacy criteria.
(ii) Data Utility: Data utility is basically a measure of information loss or loss in the functionality of data in providing the results, which could be generated in the absence of PPDM algorithms.
(iii) Uncertainty level: It is a measure of uncertainty with which the sensitive information that has been hidden can still be predicted.

(iv) Resistance: Resistance is a measure of tolerance shown by PPDM algorithm against various data mining algorithms and models.

As such, all the criteria that have been discussed above need to be quantified for better evaluation of privacy preserving algorithms but, two very important criteria are quantification of privacy and information loss. Quantification of privacy or privacy metric is a measure that indicates how closely the original value of an attribute can be estimated. If it can be estimated with higher confidence, the privacy is low and vice versa. Lack of precision in estimating the original dataset is known as information loss which can lead to the failure of the purpose of data mining. So, a balance needs to be achieved between privacy and information loss [6].

Table 1: Advantages and Limitations of PPDM Techniques

Technique	Advantages	Limitations
Anonymization based PPDM	Identity or sensitive data about record owners are to be hidden.	Linking attack. Heavy loss of information.
Perturbation based PPDM	In this technique different attributes are preserved independently.	Original data values cannot be regenerated. Loss of information.
Randomized Response based PPDM	It is relatively simple useful for hiding information about individuals. Better efficiency compare to cryptography based PPDM technique.	Loss of individual's information. This method is not for multiple attribute databases.
Condensation Approach based PPDM	Use pseudo data rather than altered data. This method is very real in case of stream data.	Huge amount of information lost. It contain same format as the original data.
Cryptography based PPDM	Transformed data are exact and protected. Better privacy compare to randomized approach.	This approach is especially difficult to scale multiple parties are involved.

4. CONCLUSION:

The main purpose of privacy preserving data mining is developing algorithm to hide or provide privacy to certain sensitive or private information so that they cannot be disclosed to unauthorized parties or intruder. Although a Privacy and accuracy in case of data mining is a pair of ambiguity. Succeeding one can lead to adverse effect on another. In this, we made an effort to survey a good number of existing PPDM methods. Finally, we conclude there does not exists a single privacy preserving data mining algorithm that outperforms all other algorithms on all possible criteria like accuracy, performance, utility, cost, complexity, tolerance against data mining algorithms etc. Different algorithm may perform better than another on one particular criterion. So here we conclude this survey and analysing the existing work and develop the new method in the future.

5. ACKNOWLEDGMENT:

I am very obliged to Dr. N. D. Shah, Principal of Parul Institute of Engineering and Technology for providing facilities to achieve the desire milestone. I also extend my thanks to Head of Department Dr. Gokulnath K for his inspiration and continuous support. I wish to warmly thank my guide, Ms. Tejal Patel for all her diligence, guidance, encouragement, inspiration and motivation throughout. Without her treasurable advice and assistance it would not have been possible for me to attain this landmark. I take this opportunity to thank all my friends for their support and help in each and every aspect.

RFERENCES:

1. Kiran Patel, Hitesh Patel, Parin Patel, "Privacy Preserving in Data stream classification using different proposed Perturbation Methods ", *IJEDR,* 2014, Volume 2, Issue 2 | ISSN: 2321-9939.
2. Radhika Kotecha, Sanjay Garg, "Data Streams and Privacy: Two Emerging Issues in Data Classification", *5th Nirma University International Conference on Engineering (NUiCONE), IEEE* 2015.
3. Rupinder Kaur and Meenakshi Bansalt, "Transformation Approach for Boolean Attributes in Privacy Preserving Data Mining " *1st International Conference on Next Generation Computing Technologies, IEEE* 2015,644-648.
4. Neha Pathak, Shweta Pandey, "An Efficient Method for Privacy Preserving Data Mining in Secure Multiparty Computation ", *Nirma University International Conference on Engineering, IEEE* 2013,1-3.

5. Dhanalakshmi.M, Siva Sankari.E "Privacy Preserving Data Mining Techniques-Survey", *ICICES, IEEE 2014*, ISBN No.978-1-4799-3834-6/14.
6. Hina Vaghashia, Amit Ganatra "A Survey: Privacy Preserving Techniques in Data Mining ", *International Journal of Computer Applications* (0975 – 8887) Volume 119 – No.4, June 2015
7. C. Clifton, M. Kantarcioglu, and J. Vaidya, "Defining Privacy for Data Mining", *Next Generation Data Mining, AAAI/MIT Press*, 2004.
8. Sweeney L, "Achieving k-Anonymity privacy protection uses generalization and suppression" *International journal of Uncertainty, Fuzziness and Knowledge based systems*, 10(5), 571-588, 2002.
9. Gayatri Nayak, Swagatika Devi, "A survey on Privacy Preserving Data Mining: Approaches and Techniques", *ternational Journal of Engineering Science and Technology,* Vol. 3 No. 3, 2127-2133, 2011
10. Charu C. Aggarwal, Philip S. Yu "Privacy-Preserving Data Mining Models and algorithm" *advances in database systems 2008 Springer Science*, Business Media, LLC
11. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concepts and Techniques: 3rd Edn; Morgan Kaufmann Publishers is an imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA.

WEB REFERENCE:

http://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf