IJETCS International Journal of Emerging Technology and Computer Science
ISSN: 2455-9954

Privacy Preserving Techniques in Data Stream and challenges

Mrs. Aditi Kalia, , Mrs, Pallavi Ghaste,

Kalia.86aditi@gmail.com, khude.pallavi@gmail.com,

Department of Computer Engineering

DYPCOE,Akurdi

Abstract
Data mining gets valuable data from large amounts of knowledge. In latest, knowledge streams are new sort of knowledge, which are fully totally different from existing static knowledge. The characteristics of information streams are: data has timing preference; data distribution changes perpetually with time; the quantity of is large; knowledge flows in and out quickly; and immediate reply is important. Existing algorithmic program is intended for the static database. If the information changes, it'd be mandatory to rescan the complete dataset, that takes to a lot of computation time and providing late answer the user. The matter of privacy-preserving data processing has wide been studied and lots of techniques are realize. However, existing techniques for privacy-preserving data processing are designed for static knowledge bases and don't seem to be appropriate for dynamic data. Once got to perform computation at that point to providing privacy together so that the privacy preservation drawback of data streams mining is very huge issue. The success of privacy protective knowledge stream mining algorithms is measured in terms of its accuracy, performance, knowledge utility, level of uncertainty or resistance to data processing algorithms etc. but no privacy protective algorithm exists that outperforms all others on all attainable criteria. Rather, an algorithmic program could perform higher than another on one specific criterion. So, the aim of this paper is to present current situation of privacy protective knowledge stream mining framework and techniques.

Keywords—Privacy-preserving big data stream mining, mining big data streams, privacy of big data processing

I. INTRODUCTION

Data Mining is outlined as extracting info from large sets of knowledge. In different words, we will say that data processing is the procedure of mining data from knowledge. There’s a large quantity of knowledge accessible within the data industry. This knowledge is of no use till its regenerate into helpful info. It’s necessary to analyse this large quantity of knowledge and extract helpful information from it. Extraction of data isn't the sole method we want to perform; data processing also involves different processes like knowledge improvement, knowledge Integration, knowledge Transformation, data processing, Pattern analysis and knowledge Presentation [1].

Information is these days most likely the foremost vital and demanded resource, in online worked society that relies on the dissemination and sharing of knowledge within the personal furthermore as within the public and governmental sectors. Governmental, public, and personal establishments are progressively needed to form their knowledge electronically accessible. Therefore there want to protect the privacy of the respondents (individuals, organizations, associations, business institutions, then on) [2]. An information stream could be a sequence of infinite, real time knowledge things with a awfully high rate that may solely browse once by associate degree application. Imagine a satellite-mounted remote device that's perpetually generating knowledge. The information are huge (e.g. terabytes in volume), temporally ordered, quick ever-changing, and potentially infinite. These options cause difficult issues in knowledge streams field. Knowledge Stream mining refers to informational structure extraction as models and patterns from continuous knowledge streams. Data Streams have totally different challenges in several aspects, like machine, storage, querying and mining [1].

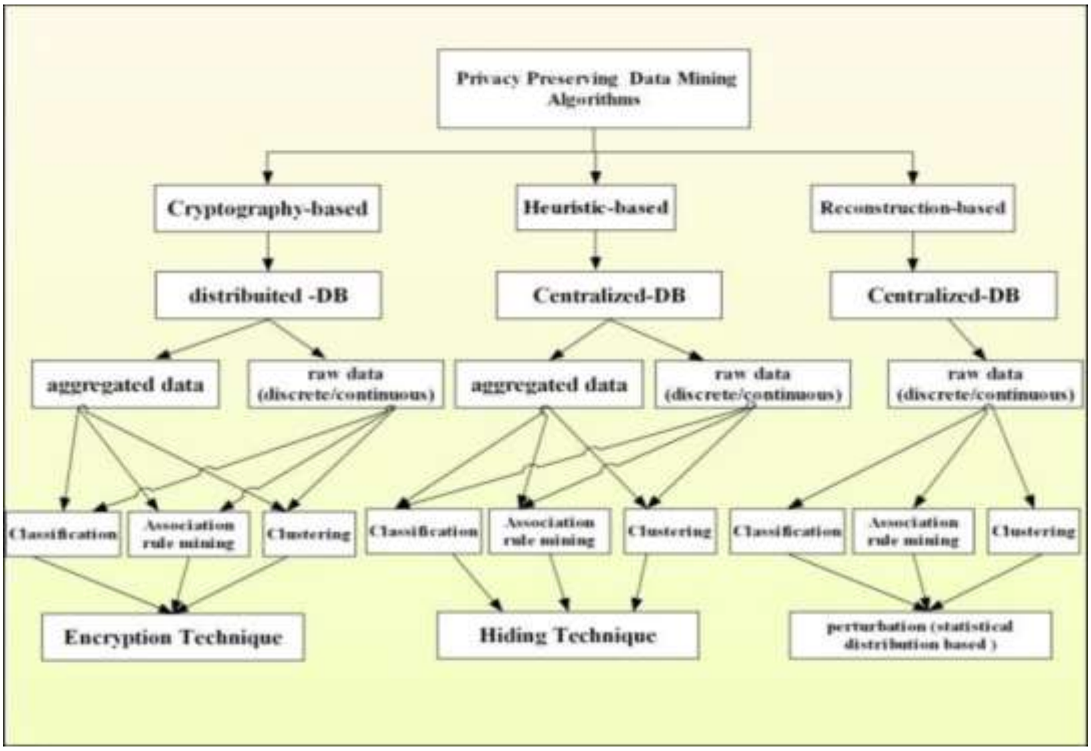


Figure 1 PPDM Algorithms/Approaches

II. LITREATURE REVIEW

Data anonymization could be a promising method among the discipline of privacy protective data processing wont to defend the information in opposition to identity revelation. Information loss and long-established attacks attainable on the anonymized info are vital challenges of anonymization. Not too way back, information anonymization utilizing info mining ways has showed mammoth improvement in info utility. Still the prevailing approaches lack in strong handling of attacks. As a result J. Jesu Vedha Nayahi et al. planned associate anonymization formula established on agglomeration and resilient to similarity attack and probabilistic illation attack is planned [3].

R. Rajeswari et al. Proposes a privacy persevered access control mechanism for knowledge streams. For the privacy security mechanism it makes use of the mixture of each the Kanonymity procedure and fragmentation system. The kanonymity procedure makes use of the suppression and generalization. It prevents the privacy revelation of the sensitive info. The privacy defence mechanism avoids the identity and attributes revelation. The privacy is dead by means that of the high accuracy and consistency of the person experience, i.e., the exactness of the non-public data. [4].

Author et al. Show however the exclusive departments of same cluster mix their information while not harming the privacyof the client for creating strong alternatives in efficient and proper manner. For that reason the approaches vertically data combination, cryptography and decision mining is established. To mine the alternatives from the information a C4.5 resolution tree is employed. The implementation of the projected privacy conserving information mining and decision making methodology is carried out using JAVA technology. In addition the potency of the tactic is computed in phrases of accuracy, error rate, memory consumption and time consumption. Within the finish to justify the effects of the projected data processing system the conventional J4.5 tree utilizing Maori hen instrument is employed with same knowledge for comparative performance learn. The experimental results show the mighty performance and protection at intervals the given privacy conserving procedure [20].

III. PRIVACY PRESRVING DATA MINING TECHNIQUE

Privacy preserving data processing techniques will be generally categorized as 3 ways Heuristic approach – Heuristic methodology is simply concerning used for centralized info, right here 2 sorts of knowledge is viewed, which is, raw information and aggregative information. Over every types of information Classification, Association rule mining, bunch ways are applied, after that concealing procedures are used over the result of them to preserve it from incorrect utilization.

Reconstruction approach – Reconstruction approach is also used for centralized info, but here, just one kind of data is employed, which is, raw data. The information mining ways are applied over the data. Regardless of the outcome comes, the applied math distributed primarily based methodology is employed over them.

Cryptography approach–Cryptography approach is basically works on distributed info that is that the one, where knowledge is keep in numerous places. The information that is being keep, could also be data or aggregative data or each. On applying data processing ways on every kind of knowledge some results can come back, on them encoding technique are used. The PPDM techniques will be additional categorised, which follows these approaches.

Anonymization based approach: The aim of anonymization procedure is to hide sensitive or personal information concerning a private. Anonymization may be a strategy to retain the info} so as that original information will be alternate into hid knowledge with the assistance of many approaches. The k-anonymity methodology says that knowledge ought to be indistinguishable inside within the k records. This will be done victimisation Generalization and Suppression techniques. Due to the some limitation of the k-anonymity methodology, Ldiversity, T-closeness ways are derived.

Randomization response approach: The randomised response approach may be a manner to mask the first information by adding some random knowledge or noise in it, so One don't seem to be ready to say that information from an individual contains real ability or no longer. The adscititious random data or noise should be as huge as doable therefore big data concern cannot be recovered by the un-trusted one. This is statistical approach 1st projected by Warner. The randomized response method is finished in

2 phases. In the primary section, the first data is being randomised and transfer to the receiver facet. Within the secondary section, the receiver reconstruct the first knowledge from randomised knowledge by distribution reconstruction algorithmic program.

Perturbation approach: The perturbation approach modified the conventional data values with artificial information values, so as that the information computed from the perturbed knowledge will no longer distinguish from the ability computed from original knowledge. The perturbation approach are of 2 kind.

Additive perturbation: In additive kind, random noise is added to the first knowledge. Multiplicative perturbation: In increasing kind, random rotation methodology is used to perturb knowledge.

Condensation approach: Condensation methodology constructs restricted clusters in dataset once that generates pseudo information from the data of those clusters. It is known as condensation owing to the sooth that of its strategy of applying condensed facts of the clusters to generate pseudo knowledge. It creates units of multiple size from the data, specified it's definite that every and each record lies in a very suite whose size is a minimum of alike to its obscurity level. Evolved, pseudo information are generated from each and every set so you'll create an artificial data set with the equal mixture distribution because the selected information.

Cryptography approach: cryptanalytic procedures are ideally meant for such things the place multiple parties collaborate to cipher outcome or share non sensitive mining outcome and thereby averting disclosure of touchy knowledge. Cryptanalytic procedures to search out its utility in such things provided that of 2 motives: 1st, it offers a well-defined model for privacy that features ways for proving and quantifying it. Second, an oversized set of cryptographic algorithms and constructs to place in effect privacy preserving data mining strategies are to be had on this area. The data might even be distributed among special collaborators vertically or horizontally.

IV. CHALLENGES IN PRIVACY-PRESERVING BIG DATA STREAM MINING

Privacy-preserving big data stream mining could be a consistent area of analysis that opens the door to many analysis challenges and directions to be thought of by near-future research efforts.

Emerging Domains: it's clear enough that, thanks to the precise research focus, i.e. privacy-preserving huge information stream mining, practical applications and systems drive and confirm the effective necessities for corresponding privacy-preserving big information stream mining algorithms. Hence, rising domains, such as social networks, intelligent TV provisioning, intelligent transportation systems, can play a superior role within the future.

Accuracy vs Privacy: Accuracy and privacy are conflict properties for large information stream mining algorithms. Indeed, determining the right trade-off between these 2 properties is a basic analysis issue. The way to increase privacy whereas preserving accuracy? The latter could be a relevant question for future research activities.

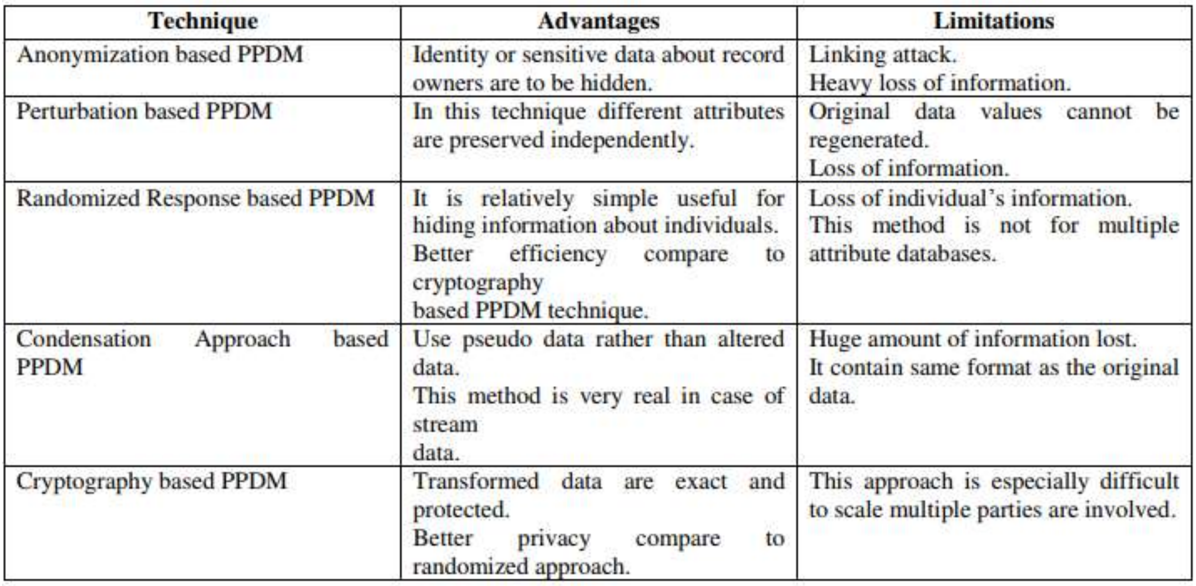
Concept-Drift problems: huge information streams are laid low with concept-drift issues. This makes more durable the privacy preserving requirement, because of, in general, protective the privacy of knowledge is performed in dependence on a planned set of attributes/concepts of the target data model.

Security problems: protective the privacy of big data streams implies accessing huge information streams, of course. This involves in a problematic side-effect: the way to make sure the security of the same huge information streams whereas accessing them? Combining security and privacy (as well as privacy and security) is an annoying drawback for big data stream mining analysis. Cryptography. Historically, the privacy-preserving huge information stream mining drawback has been self-addressed by means that of model-based or algorithmic-based approaches. Along with these initiatives, the usage of crypto logical methodologies is emerging as a promising approach to be explored by future research efforts.

Quality and Utility of data. Guaranteeing the privacy of information streams whereas mining huge information streams could deteriorate the same quality and utility of such information. As a consequence, the latter are important problems for the long run.

Stream Analytics. Models, techniques and algorithms projected by active literature should converge in appropriate unifying frameworks for finally supporting privacy-preserving huge information stream analytics, an important analysis challenge at currently. Here, several problems arise: from field of study necessities to parameter calibration, from framework trade-offs to performance, and so forth.

Performance. Last however not least, performance problems invariably arise once process huge information streams (to preserve their privacy, during this case). As a consequence, production models and optimizations that enable U.S. to confirm performance of privacy-preservation mining strategies over huge information streams could be a relevance challenge for the long run [5].



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.

V. CONCLUSION

The main purpose of privacy preserving data processing is developing algorithm to cover or offer privacy to bound sensitive or non-public info so they can't be disclosed to unauthorized parties or trespasser. Though a Privacy and accuracy just in case of knowledge mining may be a try of ambiguity. Succeeding one will result in adverse result on another. In this, we made an effort to survey an honest variety of existing PPDM strategies. Finally, we conclude there doesn't exists one privacy protective data mining algorithmic rule that outperforms all different algorithms on all potential criteria like accuracy, performance, utility, cost, complexity, tolerance against data processing algorithms etc. completely different algorithm could perform higher than another on one explicit criterion. Thus here we conclude this survey and analysing the prevailing work and develop the new methodology within the future.

REFERENCES

[1]. 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.

[2]. C. Clifton, M. Kantarcioglu, and J. Vaidya, “Defining Privacy for Data Mining”, Next Generation Data Mining, AAAI/MIT Press, 2004.

[3]J. Jesu Vedha Nayahi and V. Kavitha,” Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop”,Future Generation Computer Systems, 0167-739X/© 2016 Elsevier.

[4] R. Rajeswari and Mrs R. Kavitha ,”Privacy Preserving Mechanism for anonymizing data streams in data mining”, International conference on current research in Engineering Science and Technology(ICCREST-2016).

[5]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.

[6] A. Cuzzocrea, L. Bellatreche, and I.-Y. Song, “Data Warehousing and OLAP\ over Big Data: Current Challenges and Future Research Directions”, *Proceedings of ACM DOLAP 2013*, pp. 67-70, 2013.

[7] A. Cuzzocrea, “Analytics over Big Data: Exploring the Convergence of DataWarehousing, OLAP and Data-Intensive Cloud Infrastructures”, *Proceedings of IEEE COMPSAC 2013*, pp. 481-483, 2013.