International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 4, July-Aug-2018, ISSN (Online): 2395-566X

A Survey on Privacy Preserving Mining Features and Techniques

M. Tech. Scholar Deepa Agrawal
Dept.of Computer Science and Engineering
Mittal Institute of Technology
Bhopal, India
deepaagrawalcs@raddifmail.com

Asst. Prof. Jayshree Boaddh
Dept.of Computer Science and Engineering
Mittal Institute of Technology
Bhopal, India

***Abstract* – The daily use of the word privacy concerning secure data sharing and analysis is commonly imprecise and should be dishonest. The branch of study that embody these privacy considerations are referred as Privacy Preserving Data Mining (PPDM). So these papers focus on this problem of increasing the robustness of the data. Here various approaches adopt by researchers are detailed with their field of security. Some of issue related to the papers is also discussed. Various approaches of association rule mining are explained for finding the or hiding the hidden information** as.

***Keywords*- Cryptography, PPDM, Frequent Pattern, Anonymization etc.**

I.INTRODUCTION

Data mining is a familiar technique for intelligently extracting information or knowledge from a large amount of data [1] collected by individuals, governments, hospitals which has great opportunities to take out sensitive knowledge patterns [2].This has lead to better concern about the privacy of the personal information [3]. The complete data about an individual often includes some sensitive information. Distributing such data instantly violates individuals' privacy. The concept of privacy preserving data mining involves in preserving personal information from data mining algorithms. Privacy preserving data mining technique [5] is a new research area in data mining and statistical databases where mining algorithms are analyzed for the side effect they acquire in data privacy. The objective of privacy preserving data mining is to build algorithms for transforming the original information in some way, so that the private data and private knowledge remain confidential even after the mining process [4].

Henceforth the saving privacy of database turns out to be imperative issues [6]. The expression "Database as a Service" (DBaaS) showed up in [7, 12]. DBaaS is the breakaway innovation of the current time. The information proprietor of the association stores their information at the outsider administration providers site and delegates the duty of regulating and dealing with the information to the specialist. This worldview mitigates the need of introducing information administration programming and equipment, enlisting managerial and information administration team (staff) at the organizations site. Because of this, the association can focus on their center business rationale as opposed to on the monotonous activity of information administration prompting the sparing in information administration cost. Cloudant, Amazon DynamoDB, Hosted MongoDB are a few cases of database specialist organizations. Safeguarding the security of the outsourced databases is an incredible test in the current scenario. As the information is put away at the administration providers site, the reality of the situation may prove that specialist is doubtful as far as revealing and abusing the information. For this situation, security of the database can be hampered drastically. On the off chance that appropriate security isn't authorized, at that point there are odds of
information ruptures and hacking the information in an unapproved way. Information breaking implies unveiling the sensitive information deliberately or unexpectedly.

Sequential pattern mining is other important computation. [7]Sequential pattern mining discovers sequence of patterns from the large database. The computation of mining patterns in sequence is specifically carried out over customer purchase behavior analysis in retailing business and medical record analysis [7]. The retailer can analyze the purchase behavior of customers to predict their needs and satisfy their demands. Under privacy limitations, the privacy preserving data mining problem was intensely researched. To solve this problem number of efficient techniques has been proposed for privacy preserving data mining. It can be done without compromising the security of user's data. But most of these methods might result with some drawbacks as information loss and side-effects to some extent. This paper presents a brief survey of different privacy preserving data mining techniques and analyses the specific methods for privacy preserving data mining.

Different procedures are utilized for understanding the security in database outsourcing. These methods incorporate encryption, validated information structures,

management protecting encryption, signature plans, and so forth. In this paper, authors have given the entire investigation of security methods alongside their advantages and disadvantages.

II.TECHNIQUES OF ARM

Privacy preserving data mining techniques can be broadly categorized as three ways [13]-

1.Heuristic approach – Heuristic method is just about used for centralized database, right here two varieties of data is viewed, which is, raw knowledge and aggregated information. Over each forms of knowledge Classification, Association rule mining, Clustering methods are applied, after that hiding procedures are used over the effect of them to preserve it from incorrect utilization.

2.Reconstruction approach – Reconstruction approach is also used for centralized database, but here, only one type of data is used, which is, raw data. The data mining methods are applied over the raw data. Whatever the outcome comes, the statistical distributed based method is used over them.

3.Cryptography approach –Cryptography approach is basically works on distributed database, which is the one, where data is stored in different places. The data which is being stored, may be raw data or aggregated data or both. On applying data mining methods on each type of data some results will come, on them encryption technique will be used. The PPDM techniques can be further categorized, which follows these approaches [14,15]. Those categories are anonymization based approach: The aim of anonymization procedure is to conceal sensitive or private information about an individual.

Anonymization is a strategy to retain the data in order that original information will be alternate into hid data with the help of several approaches. The k-anonymity method says that data should be undistinguishable within in the k records. This can be done using Generalization and Suppression techniques. Due to the some limitation of the k-anonymity method, L diversity, T-closeness methods are derived.

Table 1Original data Table

Age	Weight	Name
25	50	U1
34	59	U2
49	55	U3

Table2.K-Anonymous data

Age	Weight	Name
[20-30]	[45-55]	U1
[30-40]	[55-65]	U2
[40-50]	[55-65]	U3

4.Randomization response approach: The randomized response approach is a manner to mask the original information by adding some random data or noise in it, so One are not able to say that knowledge from a person contains genuine know-how or now not. The added random data or noise must be as big as possible hence the data about someone cannot be recovered by the un-trusted one. This is statistical approach first proposed by Warner. The randomized response process is done in two phases. In the primary phase, the original information is being randomized and transfer to the receiver side. In the secondary phase, the receiver reconstruct the original data from randomized data by distribution reconstruction algorithm.

5.Perturbation approach: The perturbation approach modified the normal information values with synthetic information values, in order that the data computed from the perturbed data does now not distinguish from the know-how computed from original data. The perturbation approach are of two type.

6.Additive perturbation: In additive type, random noise is added to the original data. Multiplicative perturbation: In multiplicative type, random rotation method is used to perturb data.

7.Cryptography Approach: Cryptographic procedures are ideally meant for such situations the place multiple parties collaborate to compute outcome or share non sensitive mining outcome and thereby averting disclosure of touchy knowledge. Cryptographic procedures to find its utility in such situations given that of two motives: First, it offers a well-defined model for privateness that includes methods for proving and quantifying it. Second, a large set of cryptographic algorithms and constructs to put in force privacy preserving data mining methods are to be had on this area. The information could also be distributed among special collaborators vertically or horizontally.

III. RELATED WORK

In [2] author look on privacy protection mining on vertically distributed databases. In such a circumstance, data proprietors wish to take in the association oversees or persistent element sets from a total instructive list and reveal as weak information about their (delicate) rough data as possible to other data proprietor. To ensure data privacy, authors design a gainful homomorphic encryption plot and a sheltered connection plan. Author by then propose a cloud-supported frequent element set mining game plan, which is used to collect an association rule mining course of action.

Our answers are proposed for outsourced databases that empower different data proprietors to beneficially share their data securely without haggling on data privacy. Our answers discharge less information about the unrefined data than most existing courses of action. Conversely with the principle known plan achieving a relative security level as our proposed courses of action, the execution of our proposed game plans is three to five solicitations of size higher.

In [3] scientist address the regular test is to decide how to team up viably crosswise over restrictive authoritative limits while boosting the utility of gathered data. Since utilizing just neighborhood information gives imperfect utility, strategies for privacy safeguarding community oriented learning revelation must be created. Existing cryptography-based work for security safeguarding information mining is still too ease back to be in any way viable for extensive scale informational collections to confront the present enormous information challenge.

Past work on irregular Decision trees (RDT) demonstrates that it is conceivable to produce comparable and precise models with considerably littler cost. Work misuse the way that RDTs can normally fit into a parallel and completely disseminated design, and create conventions to execute security safeguarding RDTs that empower general and proficient circulated privacy saving learning disclosure. author safely develop RDTs for both on a level plane and vertically apportioned informational indexes. Authors execute the proposed conventions and investigate the calculation and correspondence cost, and security.

In [4] To secure corporate privacy, the data proprietor changes its data and water crafts it to the server, sends mining request to the server, and recovers the certifiable cases from the expelled cases got from the server. In this paper, authors consider the issue of outsourcing the association manage mining undertaking inside a corporate security sparing framework. Authors propose an ambush show in light of establishment data and devise an management for security ensuring outsourced mining. Our management ensures that each changed element is obscure with respect to the aggressor's experience data, from in any occasion k−1 other changed elements.

In [5] If the preparation informational collections are one-sided in what respects biased (sensitive) characteristics like sexual orientation, race, religion, and so forth., oppressive Decisions may follow. Thus, antidiscrimination strategies including segregation disclosure and counteractive action have been presented in information mining. Segregation can be either immediate or backhanded. Coordinate separation happens when Decisions are made in light of sensitive characteristics. Backhanded segregation happens when Decisions are made in light of nonsensitive traits which are unequivocally related with one-sided sensitive ones.

In this paper, authors handle separation aversion in information mining and propose new systems relevant for immediate or roundabout segregation counteractive action independently or both in the meantime. Authors talk about how to clean preparing informational collections and outsourced informational indexes such that direct and additionally aberrant unfair Decision tenets are changed over to authentic (nondiscriminatory) arrangement rules. Authors likewise propose new measurements to assess the utility of the proposed methodologies and authors look at these methodologies.

In [6] author intend to comprehend this test and propose a component that can check whether the utility of the distributed information is equivalent to the utility guaranteed by the distributer without trading off the information security, to be specific unveiling the crude information, notwithstanding when the distributer is exploitative. Since the differential security display is getting to be accepted standard for privacy preserving as it can give thorough security insurance, our work in this paper centers around differentially private information distributing components.

In [7] This paper exhibits and investigates the experience of applying certain information mining strategies and methods on 932 Systems Engineering understudies' information, from El Bosque University in Bogotá, Colombia; exertion which has been sought after keeping in mind the end goal to develop a prescient model for understudies' scholastic execution. Past works were checked on, related with prescient model development inside scholarly conditions utilizing Decision trees, counterfeit neural systems and other characterization strategies.

As an iterative disclosure and learning process, the experience is investigated by the outcomes acquired in every one of the procedure's cycles. Each got outcome is assessed in regards to the outcomes that are normal, the information's info and yield portrayal, what hypothesis manages and the relevance of the model acquired as far as forecast precision. Said congruity is assessed considering specific insights about the populace examined, and the particular needs showed by the foundation, for example, the backup of understudies along their learning procedure, and the taking of opportune Decisions keeping in mind the end goal to forestall scholastic hazard and departure.

IV.CHALLENGES WITH MINING

Authors concentrated on various elements should have been taken care while performing association rule mining on information streams. Because of the distinctive idea of information stream, regular calculations like Apriori and FP-Growth can't be utilized as these require in excess of one output of database which is greatly unfortunate case in stream information mining condition. Two sort of issues, general and application subordinate were talked about. General issues are pertinent for all applications that management with stream information.

Information Treatment Model: Data stream emerges in never-ending and limitless way too in huge volumes. The issue is to draw out transactions from an extensive information stream that would support in association rule mining. Three structures were presented for information treatment. In Landmark show, a point known as historic point is chosen.

Every one of the transactions starting there to the current are dug for finding continuous patterns. In Damped display, every transaction is allocated some esteem and this esteem decreases with their timestamp. Late transaction is having more an incentive when contrasted with more established. In Sliding window demonstrate, a sliding window is kept up in which a bit of stream is stacked in and handled.

Memory Management: Sufficient space for obliging element sets and their frequencies when an extensive volume of information touches base on the double is the greatest issue. Additionally, with the landing of crisp stream, the frequencies of element sets differ the greater part of the circumstances. Along these lines, it is basic to get together slightest measure of data. Yet, this data ought to be adequate to yield association rules.

Decision of Algorithm: The calculation ought to be picked by the necessity of results. A few calculations give correct outcomes and some give surmised comes about with false positives or false negatives.

Idea Drift Problem: The element set which is regular can end up rare with the coming transactions and the other way around. Because of this fluctuating nature of information, expectations of association tenets can wind up erroneous. This issue is known as Concept Drift and to deal with it, incremental calculations are required. Asset mindful calculations: Resource mindful calculations are required which can change their preparing rate as per the accessibility of assets [6]. This idea will extremely accommodating in the earth where assets are shared by numerous procedures. Each application has its own needs and issues.

Clients ought to have the capacity to change the mining parameters as indicated by their necessities notwithstanding when the calculation is running. Mining multidimensional information stream is another issue which expands the many-sided quality. The applications need to create reactions as per client's inquiries. In the event that information is touching base from in excess of one source, it prompts the expanded correspondence cost. Incorporating the recurrence checks is likewise an issue.

V.PRIVACY THREATS AND FRAMEWORK

The principle objective of security is to uncover the personality and individual data, which is delicate for the particular one. There are some sort of security dangers which may reveal ones sensitive data:

- personality exposure [8]: In personality declaration risk, interloper can get the individual personality from distributed information. This risk is related to direct identifier property.
- Attribute revelation [9]: In property exposure risk, interloper can uncover person's delicate data. This risk is related to delicate property.
- Membership declaration [10]: Any data concerning individual is revealed from informational collection, known as participation exposure. This may happen when information isn't shielded from personality revelation.
- A lot of protection safeguarding strategies are existing to take care of the mystery breaking issues. The general diagram for these systems can be arranged in five stages in which information is experiences [11].
- Distribution: The circulation of information can be either brought together or conveyed. In brought together conveyance, every one of the information kept in archive on focal server, while all information are put away on various databases.
- Modification: This depicts how information is altered for covering the first information. To satisfy this prerequisite, different methods for change connected on information like bother, total, swapping, examining, concealment, clamor expansion.
- Data Mining Algorithm: The information mining approaches includes the methods for producing basic leadership comes about because of the information. This phase\stage manages different calculations like Decision tree, bunching, harsh sets, affiliation administer, relapse, grouping.
- Data concealing: The information concealing involves crude learning or total information which wants to be covered up.
- Privacy Preservation Technique: The protection safeguarding approach incorporates diverse ways to deal with accomplish security, which are, speculation, information mutilation, information sanitation, blocking, cryptographic and anonymization.

VI.CONCLUSION

This paper presents a brief survey on various standard techniques for privacy preserving data mining was presented namely: randomization, anonymization, secure multiparty computation. Because of the increasing capability to trace and gather large amount of

sensitive information, privacy preserving in data mining applications has become an important concer. Here detailed discussion of different techniques and combination of those are done. In future, work can develop one single model which overcomes some challenges and threats discuss in the paper.

REFERENCES

[1]. Das, A., Ng, W.-K., And Woon, Y.-K. 2001. Rapid Association Rule Mining. In Proceedings Of The Tenth International Conference On Information And Knowledge Management. ACM Press, 474-481.

[2]. Kim-Kwang Raymond Choo, Senior Member, IEEE, Anwitaman Datta, And Jun Shao. "Privacy-Preserving-Outsourced Association Rule Mining On Vertically Partitioned Databases". Ieee Transactions On Information Forensics And Security, Vol. 11, NO. 8, AUGUST 2016 1847

[3]. Lichun Li, Rongxing Lu, Senior Member, IEEE, Jaideep Vaidya, Senior Member, Basit Shafiq, Wei Fan, Member, Danish Mehmood, And David. "A Random Decision Tree Framework For Privacy-Preserving Data Mining". Lorenzi. Ieee Transactions On Dependable And Secure Computing, VOL. 11, NO. 5, SEPTEMBER/OCTOBER 2014

[4]. Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, And Hui (Wendy) Wang. "Privacy-Preserving Mining Of Association Rules From Outsourced Transaction Databases". Ieee Systems Journal, Vol. 7, NO. 3, SEPTEMBER 2013 385.

[5]. Sara Hajian And Josep Domingo-Ferrer. "A Methodology For Direct And Indirect Discrimination Prevention In Data Mining". Ieee Transactions On Knowledge And Data Engineering, Vol. 25, NO. 7, JULY 2013

[6]. Jingyu Hua, An Tang, Yixin Fang, Zhenyu Shen, And Sheng Zhong "Privacy-Preserving Utility Verification Of The Data Published By Non-Interactive Differentially Private Mechanisms ". Ieee Transactions On Information Forensics And Security, Vol. 11, NO. 10, OCTOBER 2016

[7]. S. M. Merchán, Member, IEEE And J. A. Duarte. "Analysis Of Data Mining Techniques For Constructing A Predictive Model For Academic Performance". Ieee Latin America Transactions, Vol. 14, NO. 6, JUNE 2016.

[8]. Hajian, S. & Domingo-Ferrer, J. (2012). A Methodology For Direct And Indirect Discrimination Prevention In Data Mining. Manuscript.

[9]. C. Clifton. Privacy Preserving Data Mining: How Do Authors Mine Data When Authors Aren't Allowed To See It? In Proc. Of The ACM SIGKDD Int. Conf. On Knowledge Discovery And Data Mining (KDD 2003), Tutorial, Washington, DC (USA), 2003.

[10]. D. Pedreschi, S. Ruggieri And F. Turini, "Discrimination-Aware Data Mining," Proc. 14th Conf. KDD 2008, Pp. 560-568. ACM, 2008.

[11]. M. Mahendran, 2Dr.R.Sugumar "An Efficient Algorithm for Privacy Preserving Data Mining Using Heuristic Approach" International Journal of Advanced Research in Computer and Communication Engineering. Vol. 1, Issue 9,pp,. 737-744 November 2012.

[12]. Pedreschi, D., Ruggieri, S. & Turini, F. "Measuring Discrimination In Socially- Sensitive Decision Records". Proc. of the 9th SIAM Data Mining Conference, pp. 581-592, SDM 2009.

[13]. Hajian, S., Domingo-Ferrer, J. & Martinez-Ballesté, A. "Discrimination Prevention In Data Mining For Intrusion And Crime Detection". Proc. of the IEEE Symposium on Computational Intelligence in Cyber Security (CICS 2011), pp. 47-54. IEEE 2011.