A Survey on Data Mining Techniques for Privacy Preserving and Hiding Sensitive Itemset

Tasneem Jahan, Prof. Amit Saxena

Department of Computer Science & Engineering

Truba Institute of Engineering and Information Technology

Abstract: Protecting the personal information of users is a crucial concern for society nowadays. There have been continuous security breaches in the databases which are a potential threat to the users and companies. Cambridge Analytica and Aadhar (UID) are the recent examples. To protect the privacy of individuals and their data, many strategies are applied based on the methods of mining. The branch of study that embody these privacy considerations is referred as Privacy Preserving Data Mining (PPDM). This paper focuses on the problem of increasing the robustness of the data. Here various approaches adopted by researchers so far are discussed in detail. Some of shortcoming and result issues in the papers are also discussed. Various approaches of association rule mining are explained for hiding the information as well.

Index Terms— Perturbation, Cryptography, SMC, Randomization, Condensation, Anonymization .

I. INTRODUCTION

"Privacy Preservation" in information mining implies that confidential or imperative information must be safeguarded or secured by the unapproved individual or attacker. The issue of privacy preserving information mining has turned out to be more important so as to expand the capacity of storing individual information about clients, and corporate information of private foundations and to outsource and a wide range of different purposes [1, 13]. As of late, the privacy of outsourced databases is a prevalent research theme. The third party systems are used to enable the clients of an organization to create, store and access their databases at supplier end. Utilizing outsourced database can enable associations to decrease equipment gear cost, framework building, yet in addition diminish cost of the work force office. But the major threat to an outsourced database is that if the supplier isn't trusted, sensitive information may get leaked, causing a huge damage to business. Henceforth, the privacy of database turns out to be a crucial issue [6]. The expression "Database as a Service" (DBaaS) is discussed in [7, 12]. DBaaS is the breakaway innovation of the current time. The information proprietor of the association stores their information at the outside site and delegates the duty of regulating and dealing with the information through the specialist at outside third party. 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 challenge in the current scenario. As the information is being stored at the administration provider's site, a situation might arise that the specialist gets doubtful in revealing and abusing the information. For this situation, security of the database can be hampered drastically. There are chances of unauthorized data access so as to hack the information or outbreak it in an unapproved way. Information outbreak implies unveiling the sensitive information deliberately or unexpectedly. As per the review taken by Trust wave Global Security [1], out of 450 information break tests, 63% of examinations were identified with the organization of outsider specialist. As indicated by the information rupture examination done by Trust wave in 2012, 76% of security lacks were caused by the outsider specialist [2]. In this manner, it is exceptionally basic for the organizations to know about the level of security in their outsourced databases to keep the information classified and consequently conforming to the administration tenets and controls. Secrecy, honesty in setting of culmination and accuracy, credibility, responsibility, and so on are considered during the establishment of security

parameters. Along with their foundation, executing them in an efficient way is essential from the security perspective. Different methods 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 are trying to present the entire investigation of security methods alongside their advantages and disadvantages.

II. TECHNIQUES OF ALGORITHM

Apriori Algorithm

It is an essential system for removing successive patterns by creating applicants. As the name suggests, it requires the earlier learning of regular element set properties. It is an incremental approach where frequent k-element set is utilized to produce frequent (k+1) element set. At first, the database is checked for discovering frequent items. At this step, the cutoff value of frequent 1-element sets are separated. A cross join on the resultant is connected to get all conceivable 2-element sets blends. Again database is examined for the numbers of those element sets and the procedure rehashes until there is no new successive element set. To lessen the quantity of competitors, calculation utilizes apriori property, additionally called downward closure property, says "If an element set isn't frequent, its supersets will never be frequent". Consequently, the calculation works in two stages: joining (cross join is performed on k-element sets to create k+1 element sets) and pruning (casting out rare element sets in view of apriori property). The inconvenience of utilizing this calculation is that database is required to be filtered numerous time which expands the execution time. Apriori algorithm also suffers from bottleneck in candidate generation, thus resulting in its slow execution.

FP Growth

It is a strategy that concentrates frequent element sets in divide and conquer technique. FP-Growth works in two stages: Constructing and Mining FP tree. Root is constantly NULL. In the event that some succession of a transaction is already existing, then the rest of the elements are joined underneath it and the number of subset elements is expanded by one. Tree is mined by developing its contingent pattern which incorporates the ways to achieve the node through root. A sub tree is developed and designs are produced by linking the element with its way. Hunt space is decreased because of the age of contingent patterns. It gives great outcomes for even long patterns. Since there is no need of constructing candidates, space complexity was decreased. When there exists large number pf patterns, FP growth proves out to be more efficient and scalabale.

ECLAT

ECLAT generates frequent itemset in a DFS manner. It is a variation of apriori calculation. It utilizes vertical information management (element: transaction id set). The element sets having number not as much as least support cutoff will be disposed. Cross join is performed to create three element sets. The 2-element set subsets of 3-element set are assessed from past table. From the descending conclusion property, 2-element sets which are not frequent, their 3-element set will likewise be occasional. Calculation rehashes till no new continuous element set is produced. Because of this procedure, numerous sweeps of database are not required since transaction id set contains all the required data for numbering underpins. DFS reduces requirement of memory space, and ECLAT is considerably faster than Apriori. It does not need to scan the database for finding support of (k+1) item sets, for k>=1.

Rapid Association Rule Mining (RARM) [1] is another association rule mining philosophy that uses the tree structure to mine the underlying data. By using the SOTrieIT structure RARM will produce huge 1-itemsets and 2-itemsets rapidly. Every node of the SOTrieIT contains Item Label and its corresponding Item Support. SOTrieIT is in a sorted order of nodes with respect to their support count.

III. RELATED WORK

In [2] author look on privacy protection mining on vertically distributed databases. In such databases, data proprietors could harm the associations, by revealing sensitive and important information other data proprietor. To ensure data privacy, authors designed a homomorphic encryption plot and a sheltered connection plan. Author then proposed a cloud-supported frequent element set mining plan which is used to collect an association rule mining. Our answers are proposed for outsourced databases that empower different data proprietors to beneficially share their data securely without haggling on data privacy. This paper has proposed work on discharging less information about the unrefined data than most existing courses of action.

In [3], authors have discussed the regular tests to decide how to boost the utility of gathered data. Since utilizing just neighborhood information gives imperfect utility, strategies for privacy safeguarding must be created. Existing cryptography-based work for security safeguarding information mining is still easier to break and to gain the insight of viable information. Past work on Random Decision trees (RDT) demonstrates that it is conceivable to produce comparable and precise models with considerably lesser cost. 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. This work has shown excellent results on vertical databases in terms of calculations and corresponding 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 s only the certifiable cases from the server. In this paper, authors consider the issue of outsourcing the association's data to the corporate security sparing framework. Authors have devised the measure for the management of security ensured outsourced mining.

In [5] If the informational collections are one-sided in respect to biased (sensitive) characteristics like sexual orientation, race, religion, and so forth, oppressive decisions may follow. Thus, antidiscrimination strategies including segregation disclosure and counteractive actions might arise. Segregation can be either immediate or backhanded. Coordinate separation happens when decisions are made in light of sensitive characteristics. In this paper, authors handle separation aversion in information mining and propose new systems relevant for immediate or roundabout segregation counteractive actions independently. Authors also discussed about how to clean information collections and outsourced informational and likewise proposed new measurements to assess the utility of the proposed methodologies.

In [6] author intend to comprehend the tests 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. 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 of 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 received outcome is assessed and compared with estimated results.

IV. CHALLENGES WITH MINING

Authors concentrated on various elements that 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 which manage stream information.

Information Treatment Model: Data stream emerges in never-ending and emerges 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 element sets and their frequencies is needed. When an extensive volume of information emerges, space is the greatest issue. Additionally, the amount of data needed to carry experiments needs 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 outcomes with false positives or false negatives.

Idea Drift Problem: The element set which is regular can end up rare with the forthcoming transactions. 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 be 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 not withstanding when the calculation is running. Mining multidimensional information stream is another issue. 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 personal or individual data, which is delicate for the particular one. There are a few security dangers which may reveal someone’s 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 five dimensions needed to be considered under this are:

• Distribution: The circulation of information can be either brought together or conveyed. In brought together conveyance, entire information is kept in archive on server, whereas the other alternative is that entire information is stored on various databases.

• Modification: This depicts how information is altered for covering the original information. To satisfy this prerequisite, different methods are swapping, examining, concealment, clamor expansion.

• Data Mining Algorithm: The information mining approaches includes the methods for producing the information that is needed. This stage manages different calculations like Decision tree, bunching, harsh sets, affiliation administer, relapse, grouping.

• Data concealing: The information concealing involves crude learning on entire information which needs to be covered up.

• Privacy Preservation Technique: The privacy protection approach incorporates diverse ways to accomplish security, such as, speculation, information mutilation, information sanitation, blocking, cryptographic and anonymization.

VI. CONCLUSION

This paper tends to outline the issues for extracting only the useful information from database without damaging the security of entire data contained with information proprietors. To achieve the security of information, data mining rules are implemented to gain only the required information, and rest of the sensitive data is kept hidden. Here detailed discussion of different techniques and combination of those are done. In few works both numeric and text information was protected, so the time and space required for those calculation has been evaluated as high. The proposed work will be the perturbation of the sensitive information.

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.