IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Privacy Preserving Data Mining Framework and Techniques

G.L.Anand Babu [1], G.Sekhar Reddy[2], A.Sriram[3]

[1,2]*Associate Professor, Department of IT, Anurag Group of Institutions, Hyderabad, Telangana.*

[3] *Professor & Head, Department of IT, Anurag Group of Institutions, Hyderabad, Telangana.*

***Abstract- Preserving security in data mining has emerged as total precondition for switch over secret data as far as data Analysis, validation, and distributing. The data found by different data mining methods may contain private data about individuals or business. Preservation of privacy is imperative part of data mining and achieves data mining objectives without losing the security of the people. A genuine concern is that non-delicate information even may convey sensitive data, including individual information, realities or patterns. Alternately no privacy preservation algorithm exists that outflanks all others on every conceivable condition. Generally, an algorithm may perform superior to another on one particular condition. In this way, the point of this paper is to present current situation of privacy preserving data mining system and techniques.***

***Keywords: data mining, privacy preserving, sensitive attributes, Privacy preserving data mining, privacy preserving techniques.***

## I. INTRODUCTION

The advancement of data mining has many preferences in data analytics and knowledge discovery. It is utilized as a part of an extensive variety of fields from monetary data analytics to intrusion detection frameworks. Consider a circumstance where at least two associations owing secret databases like to run data mining algorithms on the union of their databases without uncovering superfluous data. In current years, the issue of privacy preserving data mining has turned out to be more vital on account of expanding capacity to store individual information about clients and enhanced advancement of data mining algorithms to impact this data. Various methods like randomization and k-anonymity are recommended to perform privacy preserving data mining. Privacy preserving data mining methods are developed with the end goal that secret information is extricated is not revealed to the clients running the technique.

The real worry of Privacy Preserving Data Mining is delicate raw data like names, locations are modified from unique database, henceforth the clients of the information won't have the capacity to arrangement someone else's privacy furthermore, and sensitive knowledge got from mining which can bargain data privacy must be rejected.

## II. PRIVACY PRESERVING TECHNIQUES

Preserving privacy in data mining has appeared as total precondition for switch over secret data as far as data analytics, validation, and distributing. To maintain a strategic distance from information abuse, the information is anonymized. Numerous data mining techniques are changed to guarantee protection [5].

The techniques can be categorized into:

1. **Heuristic-based procedures**: It is a versatile change that alters just chose values that minimize the adequacy misfortune as opposed to every single accessible esteem.

2. **Cryptography-based methods**: This strategy incorporates secure multiparty computation where a computation is secure if toward the fruition of the computation, nobody can know anything with the exception of its own information and the outcomes. Cryptography-based algorithms are considered for defensive protection in a conveyed circumstance by utilizing encryption systems.

3. **Reconstruction-based strategies**: where the first circulation of the information is reassembled from the randomized information.

In view of these measurements, distinctive PPDM strategies might be ordered into taking after five classifications.

1. Anonymization based PPDM

2. Perturbation based PPDM

3. Randomized Response based PPDM

4. Cryptography based PPDM

We examine these in detail in the accompanying subsections.

### A. Anonymization based PPDM

Anonymization alludes to an approach where character or/and sensitive data about record proprietors are to be hidden. It even expects that delicate information ought to be held for analysis. Clearly express identifiers ought to be evacuated yet at the same time there is a risk of security interruption when semi identifiers are connected to freely accessible information. Such attacks are called as linking attacks.. For instance traits, for example, DOB, Sex, Race, and Zip are accessible openly records, for example, voter list.

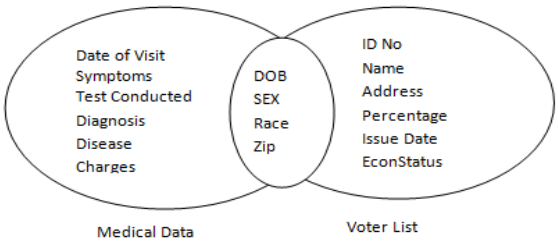


*Fig.1 Linking Attack*

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 fig.1.

Sensitive data in restorative record is illness or even medicine recommended. The semi identifiers like DOB, Sex, Race, Zip and so on are accessible in restorative records furthermore in voter list that is openly accessible. The express identifiers like Name, SS number and so forth have been expelled from the medicinal records.

Still, personality of individual can be anticipated with higher likelihood. Sweeney proposed k-anonymity display utilizing speculation and concealment to accomplish k-anonymity i.e. any individual is recognizable from at any rate k-1 different ones as for semi identifier trait in the anonymized dataset. In spite of the fact that the anonymization strategy guarantees that the changed information is valid however endures substantial data misfortune. Additionally it is not invulnerable to homogeneity assault and foundation learning assault for all intents and purposes [3].

### B. Perturbation Based PPDM

Perturbation being utilized as a part of measurable revelation control as it has a characteristic property of straightforwardness, productivity and capacity to save factual data. In bother the first values are changed with some engineered information values so that the factual data processed from the annoyed information does not contrast from the measurable data registered from the first information to a bigger degree. The annoyed information records don't consent to genuine record holders, so the aggressor can't play out the attentive linkages or recoup sensitive knowledge from the available information.

In the perturbation approach any distribution based data mining algorithm works under an understood supposition to treat every measurement freely. Pertinent data for data mining algorithms, for example, classification remains hidden in inter-attribute correlations. This is on the grounds that the perturbation approach treats distinctive characteristics autonomously. Thus the distribution based data mining algorithms have a natural burden of loss of concealed data accessible in multidimensional records. Another branch of privacy preserving data mining that deals with the weaknesses of annoyance approach is cryptographic methods.

### C. Randomized Response Based PPDM

In Randomized reaction, the information is turned in a manner that the focal place can't state with chances superior to a pre-characterized limit, whether the information from a client contains adjust data or wrong data. The data got by every single client is wound and if the quantity of clients is vast, the total data of these clients can be evaluated with great amount of precision. This is exceptionally important for choice tree characterization. It depends on consolidated estimations of a dataset, to some degree singular information things. The information gathering process in randomization strategy is done utilizing two stages [3]. Amid initial step, the information suppliers randomize their information and exchange the randomized information to the information recipient. In second step, the information collector modifies the first dissemination of the information by utilizing a

distribution reconstruction algorithm. The randomization reaction model is appeared in fig.2.

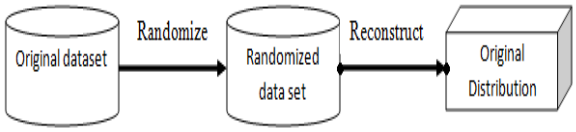


*Fig.2 Randomization Response Model*

Randomization technique is moderately extremely straightforward and does not require knowledge of the appropriation of different records in the information. Consequently, the randomization strategy can be executed at information gathering time. It doesn't require a trusted server to contain the whole unique records with a specific end goal to play out the anonymization procedure [1].

### D. Cryptography Based PPDM

Consider a situation where various therapeutic organizations wish to direct a joint research for some shared advantages without uncovering superfluous data. In this situation, explore with respect to manifestations, analysis and medicine in light of different parameters is to be led and in the meantime security of the people is to be ensured. Such situations are alluded to as circulated figuring situations [4]. Cryptographic systems are in a perfect world implied for such situations where numerous gatherings team up to process results or share non delicate mining comes about and in this manner keeping away from revelation of touchy data. Cryptographic systems locate its utility in such situations as a result of two reasons: First, it offers an all around characterized demonstrate for security that incorporate techniques for representing and measuring it. Second an inconceivable arrangement of cryptographic algorithms and builds to execute protection safeguarding data mining algorithms are accessible in this space.

## III. ASSESSMENT CRITERIA OF PRIVACY PRESERVING ALGORITHM

Privacy preserving data mining an imperative trademark in the advancement and assessment of algorithms is the distinguishing proof of appropriate assessment criteria and the improvement of related standards. For some situation, there is no security saving algorithm exists that beats the other whole algorithm on every single conceivable measure. Generally, an algorithm may perform better that another on particular measures, similar to execution or potentially information utility [2].

An initial rundown of assessment parameters to be utilized for assessing the nature of privacy preserving data mining algorithms is given underneath:

(i) **Performance:** the execution of a mining algorithm is measured regarding the time required to accomplish the security criteria.

(ii) **Data Utility:** Data utility is essentially a measure of data misfortune or misfortune in the usefulness of information in giving the outcomes, which could be created without PPDM algorithms.

(iii) **Uncertainty level:** It is a measure of instability with which the sensitive data that has been hidden can even now be anticipated.

(iv) **Resistance:** Resistance is a measure of resilience appeared by PPDM algorithm against different data mining algorithms and models.

Thus, every criterion that has been talked above should be measured for better assessment of privacy preserving algorithms at the same time; two vital criteria are evaluation of protection and data misfortune. Evaluation of security or protection metric is a measure that shows how intently the first estimation of a property can be assessed. In the event that it can be evaluated with higher certainty, the security is low and the other way around. Absence of exactness in evaluating the first dataset is known as data misfortune which can prompt to the disappointment of the motivation behind data mining. Along these lines, an adjust should be accomplished amongst security and data misfortune.

## IV. CONCLUSION

The fundamental goal of privacy preserving data mining is creating algorithm to cover up or give protection to certain sensitive data with the goal that they can't be unveiled to unapproved gatherings or gatecrasher. In spite of the fact that a Privacy and precision if there should be an occurrence of data mining is a couple of equivocalness. Succeeding one can prompt to antagonistic impact on other. In this, we attempted to audit a decent number of existing PPDM strategies. At last, we finish up there does not exists a solitary protection saving data mining algorithm that beats every single other algorithm on all conceivable criteria like execution, utility, cost, multifaceted nature, resistance against data mining algorithms and so on. Distinctive algorithm may perform superior to another on one specific paradigm.

## REFERENCES

[1] Charu C. Aggarwal, Philip S. Yu “Privacy-Preserving Data Mining Models and algorithm”advances in database systems 2008 Springer Science, Business Media, LLC.

[2] S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004, “State of the Art in Privacy Preserving Data Mining” Published in SIGMOD Record, 33, 2004, pp: 50-57.

[3] 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.

[4] Benny Pinkas, "Cryptographic Techniques for Privacy preserving data mining", SIGKDD Explorations, Vol. 4, Issue 2, 12-19, 2002.

[5] G.L.Anand Babu, Dr. K.Sudheer Reddy, G.Sekhar Reddy, “Security Challenges in BigData”, International Journal of Research in Electronics and Computer Engineering (IJRECE), issn: 2393-9028, Vol-4, Issue-3, July-September,2016.