Available online at www.ijiere.com

International Journal of Innovative and Emerging Research in Engineering

e-ISSN: 2394 - 3343 e-ISSN: 2394 - 5494

Survey on Anonymization in Privacy Preserving Data Mining

Freny Presswala [a], Amit Thakkar [b], Nirav Bhatt [c]

[a]Charotar University of Science and Technology Changa, India, freny16presswala@gmail.com

[b]Asso. Prof, Charotar University of Science and Technology Changa, India, amitthakkar.it@ecchanga.ac.in

[c]Assi. Prof., Charotar University of Science and Technology Changa, niravbhatt.it@charusat.ac.in

ABSTRACT:

Providing security to delicate data against unapproved access has been a long term goal for the database security research group and for the administration statistical organizations. Subsequently, the security issue has gotten to be, recently, a significantly more critical territory of examination in data mining. Consequently, as of late, privacy-preserving data mining has been mulled over extensively. Data anonymization, one of the methods of privacy preserving data mining, is a sort of data purification whose expectation is security insurance. It is the methodology of either scrambling or expelling by and by identifiable data from information sets, with the goal that the individuals whom the information portrays stay unacknowledged. In this paper we have reviewed various techniques of data anonymization and have shown the comparative analysis of the same.

Keywords: privacy preserving data mining, data anonymization, k-anonymity, l-diversity, t-closeness

I. INTRODUCTION

Gigantic volume of itemized individual information is consistently gathered and imparting of these information is ended up being gainful for data mining application [1]. Such information incorporate shopping propensities, criminal records, therapeutic history, credit records and so on. Data mining includes the extraction of implied beforehand obscure and possibly valuable learning from vast databases. Data mining is an extremely difficult errand since it includes building and utilizing programming that will oversee, investigate, outline, model, examinations and translate substantial datasets to recognize designs irregularities. On one hand such information is an imperative resource for business association and governments for choice making by investigating it .then again privacy regulations and other privacy concerns may keep information holders from imparting data for information investigation [1]. Privacy preserving in Data mining methods are being utilized progressively as a part of wide verity of utilization [2].

Anonymization method aims at making the individual record be indistinguishable among a group records by utilizing techniques of generalization and suppression [3]. Anonymization refers to a methodology where character or/and delicate data about record holders are to be covered up. It even accepts that delicate data should be retained for analysis [4].
There are four sort of quality of fundamental type of data [4]:

(i) **Explicit Identifiers** is a situated of properties containing information that recognizes a record manager explicitly, for example, name, percentage and so forth.

(ii) **Quasi Identifiers** is a situated of properties that could potentially recognize a record manager when combined with publicly available data.

(iii) **Sensitive Attributes** is a situated of properties that contains touchy individual particular information, for example, illness, salary and so forth.

(iv) **Non-Sensitive Attributes** is a situated of properties that makes no problem if revealed even to conniving gatherings.

The rest of the paper is organized as follows. Section 2 discusses privacy preserving data mining along with classification of Privacy Preserving Data Mining. Section 3 then describes Data anonymization and various method of data anonymization along with its limitations. Section 4 also discusses its related work. Section 5 concludes the whole paper.

II. PRIVACY PRESERVING DATA MINING

A. Privacy Preserving Data Mining

Privacy-preserving data mining finds various applications in surveillance which is naturally expected to be "privacy-violating" applications. The key is to plan routines which continue to be viable, without compromising security. Various methods have been talked about for bio surveillance, facial de-identification, and data fraud [5]. Most systems for privacy calculations utilize some type of change on the data with a specific end goal to perform the privacy protection. Typically, such techniques diminish the granularity of representation to lessen the privacy. This lessening in granularity results in a few loss of viability of data administration or mining algorithms. This is the natural exchange off between information loss and privacy.

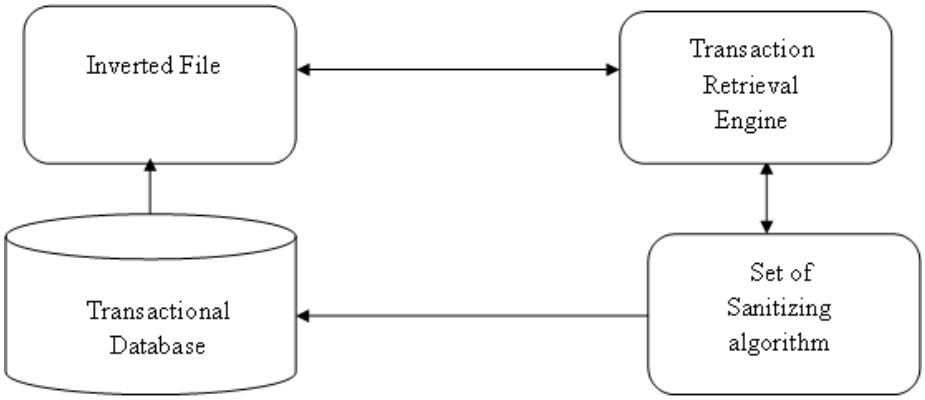

Figure 1: Framework of Privacy Preserving

B. Classification of Privacy Preserving Data Mining[6]

There are many approaches which have been adopted for privacy preserving data mining.

We can classify them based on the following dimensions:

Data distribution: This measurement refers to the dispersion of data. A percentage of the methodologies have been developed for centralized data, while others refers to a disseminated data situation. Circulated data situations can also be classified as horizontal data appropriation and vertical data circulation. Horizontal conveyance alludes to these situations where distinctive database records dwell in better places, while vertical data dissemination, refers to the situations where all the values for diverse characteristics live in better place.

Data modification: data adjustment is utilized as a part of request to change the original values of a database that needs to be released to the public and along these lines to guarantee high privacy assurance. It is vital that a data change strategy should be working together with the privacy policy received by an association.
Techniques for change include:

- Perturbation, which is accomplished by the alteration of a quality value by a new value (i.e., changing a 1-value to a 0-value, or adding commotion).
- Blocking, which is the replacement of an existing quality value with a "?".
- Aggregation or merging which is the combination of several values into a coarser classification.
- Swapping that alludes to interchanging values of individual records.
- Sampling, which alludes to releasing data for only a sample of a population?

Data mining algorithm: This measurement refers to the data mining algorithm, for which the data change is taking place. This is actually something that is not known in advance, however it facilitates the analysis and configuration of the data hiding algorithm. We have included the problem of hiding data for a combination of data mining algorithms, into our future exploration plan. For the present, different data mining algorithms have been considered in isolation of one another. Among them, the most important thoughts have been developed for classification data mining algorithms, like choice tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.

Data or rule hiding: The measurement alludes to whether crude data or aggregated data should be covered up. The complexity for hiding collected data in the form of rules is obviously higher, and therefore, mostly heuristics have been developed. The lessening of the measure of public information causes the data miner to deliver weaker inference rules that will not allow the inference of confidential values. This procedure is also known as "rule confusion".

Privacy preservation: The measurement which is the most imperative refers to the privacy protection system utilized for the selective adjustment of the data. Selective adjustment is needed so as to accomplish higher utility for the adjusted data given that the privacy is not risked.

The techniques that have been applied for this reason are:

- Heuristic-based methods like versatile adjustment that adjusts only selected values that minimize the utility loss as opposed to all available values.
- Cryptography-based systems like secure multiparty processing where a calculation is secure if toward the end of the computation, no gathering knows anything with the exception of its own input and the results.
- Reconstruction-based procedures where the original appropriation of the data is recreated from the randomized data

It is paramount to realize that data adjustment results in degradation of the database execution. To evaluate the degradation of the data, we mainly utilize two measurements. The first, measures the confidential data insurance, while the second measures the loss of functionality.

III. DATA ANONYMIZATION

Data anonymization enables the transfer of information across a boundary, for example, between two departments within an agency or between two agencies, while decreasing the risk of unintended disclosure, and in certain environments in a manner that enables evaluation and analytics post-anonymization. In the context of medical data, anonymized data alludes to data from which the patient cannot be identified by the recipient of the information. The name, address, and full post code must be uprooted together with any other information which, in conjunction with other data held by or disclosed to the recipient, could identify the patient [24].

A. K-Anonymity

Definition: (K-anonymity). A data set T satisfies K-anonymity if it is divided into a partition and each group G_i ($1 \leq i \leq p$) in the partition contains at least K records, and T is either generalized or anatomized [7].

While releasing private tables for exploration reason identifiers are expelled from the table to de-distinguish the individual yet at the same time by matching semi identifiers from private table with public table one can easily recognize the individual. In this way k-Anonymization is utilized to make in any event k tuples similar by using generalization or suppression. Generalization is procedure of substituting property values with semantically predictable yet less exact values. For example, the month of conception can be replaced by the year of conception which happens in more records, so that the ID of a particular individual is more difficult. Suppression alludes to removing a certain trait value and replacing events of the value with a special value "*", indicating that any value can be placed instead [8].

Attack on k-anonymity[9]

a) Homogeneity Attack: Alice and Bob are hostile neighbours. One day Bob falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Alice sets out to find what ailment Bob is suffering from. Alice finds the 4-unacknowledged table of current inpatient records published by the hospital, thus she knows that one of the records in this table contains Bob's data. Since Alice is Bob's neighbour, she knows that Bob is a 31-year-old American male who lives in the postal division 13053. In this way, Alice knows that Bob's record number one among 9, 10, 11, or 12. Presently, all of those patients have the same medical condition (disease), along these lines Alice concludes that Bob has cancer.

b) Background Knowledge Attack: Alice has a pen companion named Umeko who is admitted to the same hospital as Bob, and who has some patient records. Alice knows that Umeko is a 21 year old Japanese female who currently lives in postal district 13068. In light of this information, Alice learns that Umeko's information is present in record number 1, 2, 3, or 4. Without additional information, Alice is not certain whether Umeko contracted an infection or has heart disease. On the other hand, it is also well-known that the Japanese have a too low incidence of heart disease. Along these lines Alice concludes with close certainty that Umeko has a viral infection.

B. L-Diversity

Definition: (L-diversity for a single sensitive attribute) an equivalence class is said to have l-diversity if there are at least l- "well-represented" values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity[10].

Machanavajjhala et al. gave a number of interpretations of the term "well-represented" in this principle[10]:

a) Distinct l-diversity: The simplest understanding of "well represented" would be to guarantee there are at any rate l- distinct values for the delicate trait in every equivalence class. Distinct l-diversity does not forestall probabilistic inference attacks. An equivalence class may have one value seem significantly more frequently than different values, enabling a foe to conclude that an element in the equivalence class is likely to have that value. This motivated the development of the following two stronger thoughts of l-diversity.

b) Entropy l-diversity: Entropy diversity degree can be defined as:

$$D(E)= - \sum_{s \epsilon S} p(E,s) \log p(E,s)$$

Where $p(E, s)$ is the fraction of tuples in the equivalence class E with sensitive attribute level s.

A table is said to have entropy l-diversity if for each equivalence class E, Entropy(e) ≥ log l. Entropy l - diversity is solid than distinct l-diversity. As pointed out earlier, to have entropy l -diversity for every equivalence class, the entropy of the whole table must be in any event log (l). Now and then this may excessively prohibitive, as the entropy of the whole table may be low if a couple of values are extremely normal. This leads to the following less traditionalist notion of l –diversity.

c) Recursive (c, l)**-diversity:** Diversity degree of the table T can be defined as:

$$D(T)=\min(D(E_i))$$

Where E_i denotes the i-th equivalence class which obtained by anonymizing T.

Recursive (c, l)-diversity makes beyond any doubt that the most continuous value does not show up too frequently, and the less regular values don't show up too rarely. Let m be the quantity of values in an equivalence class, and ri, $1 \leq i \leq m$ be the quantity of times that the ith most continuous delicate value shows up in an equivalence class E. At that point E is said to have recursive (c, l)-diversity if $r1 < c(rl+rl+1 +...+rm)$.

A table is said to have recursive (c, l)-diversity if all of its equivalence classes have recursive (c, l)-divers.

Limitation of l-diversity[10]

a) L-diversity may be difficult and unnecessary to achieve.
b) L-diversity is insufficient to prevent attribute disclosure.

Attacks on l-diversity[10]

a) Skweness Attack: At the point when the overall dispersion is skewed, satisfying l-diversity does not counteract characteristic disclosure.
b) Similarity Attack: At the point when the touchy attributes values in an equivalence class are distinct however semantically similar, an adversary can learn essential information.

C. T-Closeness

Definition: (The t**-closeness Principle :)** An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness [9].

The t-closeness model is a further enhancement on the concept of l-diversity. One characteristic of the l-diversity model is that it treats all values of a given attribute in a similar way irrespective of its distribution in the data. This is rarely the case for real data sets, since the attribute values may be exceptionally skewed. This may make it more difficult to create feasible l-assorted representations. Often, an adversary may utilize background knowledge of the global distribution with a specific end goal to make deductions about sensitive values in the data. Furthermore, not all values of an attribute are equally sensitive. For example, an attribute corresponding to a disease may be more sensitive when the value is positive, rather than when it is negative. A t-closeness model was proposed which utilizes the property that the distance between the distributions of the sensitive attribute within an anonymized group should not be different from the global distribution by more than a threshold t. The Earth Mover distance metric is utilized as a part of request to quantify the distance between the two distributions. Furthermore, the t-closeness

approach tends to be more effective than many other privacy-preserving data mining methods for the case of numeric attributes [11].

Limitations of t-closeness[9]

a) Using EMD in t-closeness doesn't includes the measures that combine distance-estimation properties of EMD with the probability scaling nature of KL Distance.

IV. RELATED WORK

In 2013, researchers Abou-el-ela Abdou Hussien, Nermin Hamza, Hesham A. Hefny proposed a techniques which tried to solve the problem of common attacks on data mining by applying common attacks techniques for anonymization based Privacy Preserving Data Mining & Privacy Preserving Data Publishing. It uses k-anonymity and l-diversity method of anonymization [12].

In 2006, Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer proposed a k-anonymity model helps prevent from linking attack but it is unable to handle background knowledge attack and homogeneity attack which is caused by absent of diversity in database. So, new approach is needed which helps to prevent from both attack. So, a new approach called l-diversity is introduced [10].

In 2007, Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian proposed a k-anonymization model helps prevent from linking attack but it is unable to handle background knowledge attack and homogeneity attack which is caused by absent of diversity in database. So, new approach was introduced called *l-diversity which helps prevent from* both problems. But *l-diversity has number of limitations such as it is* neither necessary nor sufficient to prevent attribute disclosure. So new approach is needed which helps to solve problem of both k-anonymity and *l-diversity. This paper proposed a new privacy notion which* distributes sensitive attribute values in a table such that it satisfies *l-diversity* and helps to prevent the attribute disclosure problem [9].

In 2005, Kristen LeFevre, David J. DeWitt, Raghu Ramakrishnan proposed a method that tries to introduce a set of algorithms for producing minimal full-domain generalizations, and show that these algorithms perform up to an order of magnitude faster than previous algorithms on two real-life databases [13].

In 2006, Kristen LeFevre, David J. DeWitt, Raghu Ramakrishnan proposed a new multidimensional model which tries to propose a new multidimensional model, which provides an additional degree of flexibility not seen in single-dimensional approaches. K-anonymity has been proposed to reduce the risk of this type of attack [14].

In 2012, Dr.R. Sugumar, Dr.A. Rengarajan, M.Vijayanand tries to solve the problem of hidden linkage that can reveal basic information even after k-anonymity is applied which helps the attackers to make decision and in justifying decisions. In order to preserve the privacy of the client in data mining process, a variety of techniques based on k-anonymity of data records have been proposed recently. The main motivation of the proposed association rule hiding (ARH) algorithm is that it has reduced information loss by means of hiding those transactions that supports the specific sensitive rule [15].

In 2012, Tamas S. Gal, Zhiyuan Chen, Aryya Gangopadhyay attempt to solve the healthcare privacy issues using k-anonymity *l*-diversity model for multiple sensitive attributes. Previous research shows that k-anonymity model can be extended to multiple sensitive attributes but l-diversity not because *l*-diversity for each individual, sensitive attribute doesn't guarantee *l*-diversity over all sensitive attributes. So other approach is needed which uses both technique and should be applicable to multiple sensitive attributes also. The motivation behind this proposed work is that it produces less distortion then existing approaches and satisfies *l*-diversity for multiple sensitive attributes also [16].

In 2013, R. Mahesh, T. Meyyappan attempt to solve the problem of privacy of individual's sensitive data from attacks such as record linkage attack and attribute linkage attack [17].

In 2010, Siava Kisi levich, Lior Rokach, Yuval Elovici and Bracha Shapira attempt to solve the problem of prior knowledge regarding the domain hierarchy taxonomy of the original database that is done in the form of generalization [18].

V. CONCLUSION

This paper tries to summarize the basics about Privacy Preserving Data Mining with its various classifications. Then it introduces to Data Anonymization. Further various anonymization techniques/method such as k-anonymity, l-diversity and t-closeness has also been described along with attacks on them and their limitations. Also, some works related to Data anonymization have been shown. Overall this paper tries to show conceptive form of Data Anonymization.

VI. REFERENCE

[1] Sinha, B. K., and J. Kumar. Privacy Preserving Clustering In Data Mining. Diss. 2010.
[2] Shweta Shrma Hitesh Gupta Priyank Jain "A Study Survey of Privacy Preserving Data Mining". International Journal of Research & Innovation in Computer Engineering , Vol 2, Issue 2 , April 2012.
[3] Pingshui WANG. "Survey on Privacy Preserving Data Mining". International Journal of Digital Content Technology and its Applications Volume 4, Number 9, December 2010.
[4] Seema Kedar1, Sneha Dhawale2, Wankhade Vaibhav3,Pavan Kadam4 , Siddharth Wani5 , Pavan Ingale, "Privacy Preserving Data Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013
[5] Mynavathi, R., N. Sowmiya, and D. Vanitha. "Survey of Various Techniques to Provide Multilevel Trust in Privacy Preserving Data Mining."
[6] Charu C. Aggarwal, Philip S. Yu, "Privacy-Preserving Data Mining: Models And Algorithms", Kluwer Academic Publishers, Boston/Dordrecht/London
[7] Jian-min, Han, Cen Ting-ting, and Yu Hui-qun. "An improved V-MDAV algorithm for l-diversity." Information Processing (ISIP), 2008 International Symposiums on. IEEE, 2008.
[8] Snehal M. Nargundi, Rashmi Phalnikar, k-Anonymization using Multidimensional Suppression for Data De-identification, International Journal of Computer Applications (0975 – 8887) Volume 60– No.11, December 2012
[9] Ninghui Li Tiancheng Li,Suresh Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l-diversity, ICDE 2007, pp. 106–115
[10] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam. "ℓ-Diversity: Privacy Beyond k-Anonymity". Department of Computer Science, Cornell University.
[11] Aggarwal, Charu C., and S. Yu Philip. A general survey of privacy-preserving data mining models and algorithms. Springer US, 2008.
[12] Hamza, Nermin, and Hesham A. Hefny. "Attacks on anonymization-based privacy-preserving: a survey for data mining and data publishing." Journal of Information Security 4: 101, 2013.
[13] LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "Incognito: Efficient full-domain k-anonymity." Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005.
[14] Kristen LeFevre, David J. DeWitt, Raghu Ramakrishnan. "Multidimensional K-Anonymity". University of Wisconsin, Madison Department of Computer Sciences Technical Report 1521 Revised June 22, 2005.
[15] Dr.R. Sugumar, Dr.A. Rengarajan, M.Vijayanand. "Extending K-Anonymity to Privacy Preserving Data Mining Using Association Rule Hiding Algorithm". International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.
[16] Gal, Tamas S., Zhiyuan Chen, and Aryya Gangopadhyay. "A privacy protection model for patient data with multiple sensitive attributes." International Journal of Information Security and Privacy (IJISP) 2.3 (2008): 28-44.
[17] Mahesh, R., and T. Meyyappan. "Anonymization technique through record elimination to preserve privacy of published data." Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on. IEEE, 2013.
[18] Kisilevich, Slava, et al. "Efficient multidimensional suppression for k-anonymity." Knowledge and Data Engineering, IEEE Transactions on 22.3: 334-347, 2010.
[19] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," int'l J. Uncertaillty, Fuzizness, and knowledge-Based Systems, vol. 10, no. 5, pp. 571-588, 2002.
[20] Seema Kedar1, Sneha Dhawale2, Wankhade Vaibhav3,Pavan Kadam4 , Siddharth Wani5 , Pavan Ingale, "Privacy Preserving Data Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April 2013
[21] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "K-ANONYMOUS DATA MINING: A SURVEY", Springer US, Advances in Database Systems (2008)
[22] Shweta Shrma Hitesh Gupta Priyank Jain "A Study Survey of Privacy Preserving Data Mining". International Journal of Research & Innovation in Computer Engineering , Vol 2, Issue 2 , April 2012.

Book Reference:

[23] Charu C. Aggarwal, Philip S. Yu. A book on "Privacy-Preserving Data Mining: Models and Algorithms". Springer.

Web Reference:

[24] http://en.wikipedia.org/wiki/Data_anonymization