# Secure Techniques to Release Private Data using Anonymization

[1]Apoorva Joshi [2] Pratima Gautam
[1]Research Scholar [2]Assosiate Professor
[1] CS Dept [2] Dept of CS&IT
RNTU University Bhopal India

***Abstract:*** Now a day's there is a wide use of internet, the data existing on it must be made available in a mode that an individual's privacy is not exaggerated. In recent times, many organizations are accumulating extremely large amounts of data which are stored in huge databases. Data publisher collect data from data holders, and make public this data to data recipient for data mining and statistical investigation etc. The released data can disclose secret information of an individual. For providing privacy to the data, many anonymization techniques have been planned for privacy preserving and micro data publishing. This paper talks about various anonymization techniques such as generalization, bucketization, slicing and it also provide a technique for magnify security in the annonymization technique. Additional, a comparative study of the proposed method with existing techniques is discussed.[1]

***IndexTerm:*** **Data publishing, data security, data anonymization, privacy preservation, generalization, bucketization, slicing.**

## I. Introduction

Data mining is a logical process primarily designed for exploring data. It is examine the data and summarizing it into positive information. It is extracting of unknown knowledgeable information from a large database. The broad range of applications of Data mining include business, marketing, medical, scientific field etc. One of the important applications of Data mining is in the field of health care. Health mining is pull out previously strange and secret data and from large medical database. It is gradually more popular nowadays. Privacy of an individual in large amount of medical data is very multifaceted and its processing is also very difficult by regular methods.[2]

Data mining offers efficient techniques and methodology for dealing out with medical data. It helps to expand better diagnosis of diseases and treatment. Health mining function includes effective treatment, healthcare management, medicine discovery, digital health records, customer management, fraud and violence [14] with healthcare data. Still Health Data mining has this much of applications, which affects individuals' privacy. Several organizations publish these medical data. The outside Suppliers and the insurance agent can sell this data for various companies as product. Also the healthcare data can be sold by the person who can contact with the cloud where this data is stored [1]. This may affect individuals confidentiality. Therefore privacy preservation is important. To publish the medical data without disturbing the privacy we need to anonymize the medical data. We have to anonymize the data before publishing.

## II. Background

### Method of K-Anonymity

K-anonymity is a key idea that was introduced to tackle the risk of re-identification of anonymised data through connection to other database. It is introduced by L. Sweeney and P. Samarati in a paper published in 1998[13] To guard the anonymity of the entities which is called respondents, the micro data undergoing public or semipublic release refer, data holders often delete or encrypt quasi identifiers such as names, addresses, and phone numbers. De-identifying micro data, yet it provides no guarantee of anonymity. Released micro data contains other data, such as race, birth date, sex, and ZIP code, which

easily linked to publicly available information to re- identify the individual data, thus leaking information that was not intentional for exposé.[3]

## Random Perturbation

Random Perturbation is a unsystematic ordering of a object that is a permutation valued random variable Randomization technique is usually used to conceal the data by adding the random value and adding the cover with the table. The random value added is correctly large so that the individual values of the records can no longer be recovered by any rival. It also define by Pingshui Wang in [15] In general, randomization technique aims at finding a suitable balance between privacy preservation and knowledge discovery. Representative randomization methods contain random-value based perturbation and Randomized reaction [4]

## Blocking Based technique

Blocking Based technique is concerned, positive modifications to the support and confidence structure introduced in Saygin et al.16,17 are offered to account for the dependencies connecting the bounds on the values of the nominator and the denominator during the calculation of the confidence gap. In particular, the proposed blocking algorithm first finds the group of transactions that supports the sensitive rule as well as the group of transactions that partially support this rule. Then, the algorithm sorts these groups based on the weights given similarly to the earlier technique and it blocks a definite number of 0's followed by a definite number of 1's in the list of sorted transactions. [5]

## Cryptographic Technique

Data Confidentiality it provided by one of two type of encryption algorithm called symmetric cryptography and Asymmetric cryptography Sensitive data encrypted by using cryptography technique. In [18], authors launch cryptographic technique which is extremely popular because it gives security and protection to the sensitive attributes [18]. There are many cryptography algorithms are available. But all these techniques have disadvantages like they not succeed to protect the result of computation. The mostly cryptographic algorithm proposed does not give fertile results in case of large database and it is extremely difficult to apply this algorithm to large databases.[6]

## Condensation Approach

Condensation is one more privacy preserving approach. Charu C. Aggarwal and Philip [19] introduced this method, which constructed constrained clusters in the database and then produces fake data. condensation of data into multiple sets of predefined amount is the main concept of this technique. Statistics are predefined for each set group. This method is used in vibrant data update such as stream problems. Each set of group has a specific size of data at least k which is set to as the specific of that privacy-preserving approach. The higher the level, the higher is the amount of privacy. Statistics from each group are used to generate corresponding pseudo-data. Even though this is a simple privacy preservation approach, it is not efficient one as it leads to loss of information.[7]

## III. Comparison chart of Different technique

Table 1. show the comparison chart of the techniques which used to hide the sensitive data

| S.No. | Technique used for Privacy | Merits | Demerits |
|---|---|---|---|
| 1 | **K-Anonymity** | k-anonymity is basically used to protect quasi-identifier while exposing truthful information., it does not give sufficient protection against attribute disclosure.[3] | Firstly, it might be very tough for the proprietor of a database to take decision which of the attributes are allow or are not available in external tables. The another problem is that the *k*-anonymity model assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods. |
| 2 | **Random Perturbation** | Autonomous action of the unusual attributes by the perturbation approach. It is a simple approach.[4] | Under such conditions ,this approach is easier to break and gives us little privacy. |
| 3 | **Blocking Based** | discover sensitive attribute and then they replace know sensitive values with random value .therefore random values help in privacy preserving[5] | Rebuild of old dataset is pretty difficult |
| 4 | **Cryptographic Technique** | It offers a well defined sculpt for privacy, which includes techniques for proving and Quantifying it. [6] | This method is particularly difficult to scale when more parties are involved. Also, it does not tackle the question of whether the disclosure of the final data mining result |
| 5 | **Condensation Approach** | This Method works with pseudo-data rather than | It uses pseudo data no longer necessitates the recunstruction of data mining algorithms, |

| | | | |
|---|---|---|---|
| | | with modifications of main dataset, this provide high preservation of privacy than other techniques which simply does some changes in original data.[6] | since they uses the same format as the original database have. |

**Table.1**

## IV. K-Anonymity Technique

When releasing data for analysis and statistical purposes, it is must to disclosure the risk on a suitable state while maximizing data utility. To limit instructive this hazard, Samarati Sweeney [13] introduced the *k*-anonymity privacy preservation requirement, which needs every record should be anonymized Table data to be identical with at least k other records within the dataset, with respect to a compilation of quasidentifier attributes. To achieve the *k*-anonymity requirement, they used each generalization and suppression for data anonymization. Unlike usual privacy protection techniques e.g. data swapping and adding noise, information in a k-anonymous table is using generalization and suppression leftovers truthful. Particularly, a table is k- anonymous if the values of each and every tuple are identical.

While k-anonymity provides security against identity leak, it does not give sufficient protection beside attribute disclosure. There are two attacks called homogeneity attack and background knowledge attack. The limitations of the k-anonymity model relate from the two assumptions. Firstly, it may be very tough for the holder of a database to decide which of the attributes are or are not available in external tables. The second limitation is that the *k*-anonymity model assumes a [8] certain method of assault, whereas in genuine situations there's no cause why the enemy should not strive with other methods.

Steps of K-Anonymity

**Step 1**: Select database and Table T

**Step 2**: Select Key attribute, Quazi-identifier attribute and Sensitive Attribute from give n attribute list

**Step 3**: Select the set of most sensitive values A from list of all sensitive values that is to be preserved

**Step 4**: For each tuple whose sensitive value belongs to set A they move all these tuples to Table T1 and rest to table T2.

**Step 5**: Find the statistics of quazi attributes of table T1 i.e. distinct values for that attribute and total no of rows having that value.

**Step 6:** Apply generalization on quazi identifiers of table T1 to make it anonymized

**Step 7 :** Append rows of table T1 and table T2. T*=T1+T2 which is table ready to release. [12]

Figure 1. shows two windows input and output window .

Input window represents the original data set.

Output window represents the modified data set. Using anonymization technique.

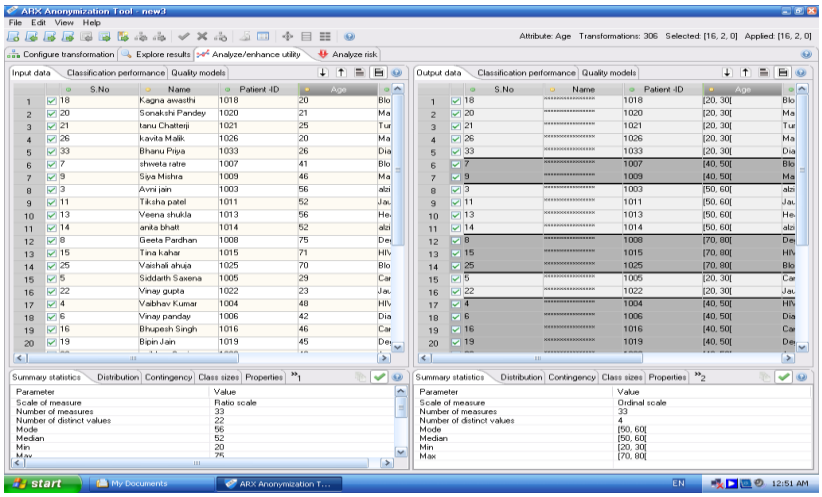

figure.1

Figure. 2 Represents the Risk analyzation in given dataset in which minimum records are in risk level which do not affect the data confidentiality

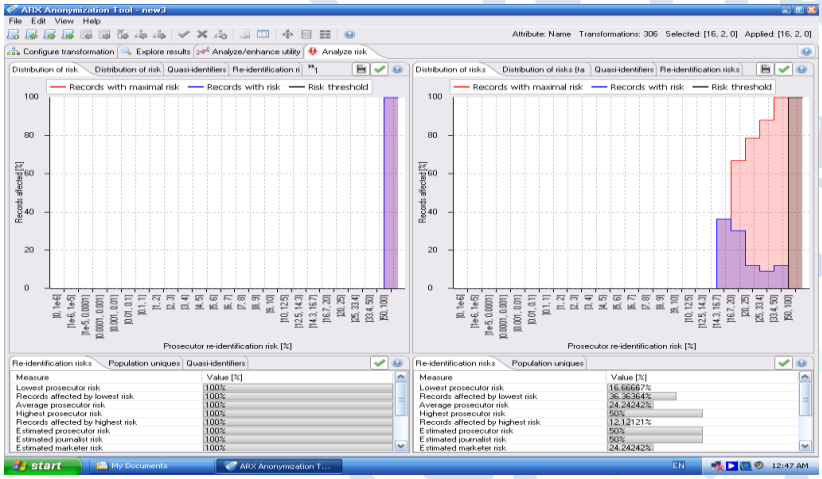

figure.2

## V. Problem with current Technique

This paper talks about anonymization algorithm that provides differential privacy preservation and provides effective categorization analysis. The used method outcome linked with the classical generalization technique connects with output perturbation to effectively anonymize data. But data mining in healthcare is different from the other fields, because the micro data are heterogeneous like ethical, legal, and social other constraints relate to private medical information. It is important to concern of these databases because medicine itself link with the status of life. Data from medical sources are huge, but they come from many dissimilar sources, not all appropriate structure or value. The doctor's interpretations are an crucial component of these database. The additional mathematical models are badly characterized compared to the physical sciences, medical data mining relate to privacy and security considerations, and it necessary to maintain stability the probable benefits of research beside any problem or possible harm to the patient. The main issues or threat in anonymization technique which considers all of sensitive attribute values at same level and applies generalization on all, this leads to some issues like, [9]

- Information Loss
- Data Utility
- Privacy
- Only work for centralized data

## VI. Proposed Work

By comparison and studying lots of technique of annomnization all have some negative and some positive aspect no one technique is work well with data set and provide complete privacy. Especially when we work with medical dataset here we require one or more methodologies [11]. In future we apply two or more technique in one dataset and may be take association rule hiding technique [10] to overcome the current issues and problem in privacy preservation data mining technique

## V. Conclusion

Anonymization of the data is one of the important method to secure the publish data. Popular techniques of data anonymization like suppression, generalization, bucketization perturbation have been used for preserving privacy of publish data. There are various constraints with these techniques like suppression reduces the quality of data drastically, generalization is inadequate in handling multi dimensional data There are various algorithms which is been studied to find out the frequent itemset and produce the association rules. Herewe use annonymization technique. It uses the medical database.By using this approach healthcare dataset can easily share with with each other without the fear of sensitive information receiving exposed and also the database remains secure.[11]

## REFERENCES

[1] Disha Dubli and D.K Yadav" Secure Techniques of Data Anonymization for Privacy Preservation", International Journal of Advanced Research in Computer Science", Volume 8, No. 5, May-June 2017

[2] 1Somy.M.S, 2Gayatri.K.S, 3Ashwini.B "Privacy Preserving Health Data Mining" International Journal of Computer Science And Technology, IJCST Vol. 6, Iss ue 4, Oct - Dec 2015

[3] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati"k-Anonymity", Springer US, Advances in Information Security (2007)

[4] Avinash Kumar Singh Narayan P. Keer Anand Motwani "A Review of Privacy Preservation Technique" International Journal of Computer Applications (0975 – 8887) Volume 90 – No.3, March 2014

[5] Vassilios S. Verykios∗ ""Association rule hiding methods", WIREs Data Mining Knowl Discov 2013, 3: 28–36 doi: 10.1002/widm.1082

[6] K. Naga Prasanthi" A Review on Privacy Preserving Data Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016

[7] Ms.R.Kavitha1, Prof.D.Vanathi2" A Study Of Privacy Preserving Data Mining Techniques", International Journal of Science and Applied Information Technology, Volume 3, No.4, July - August 2014

[8] Krzysztof J. Cios, G. William Moore" Uniqueness of medical data mining"elesvier Artificial Intelligence in Medicine 26 (2002) 1–24

[9] Noman Mohammed, Rui Chen" Differentially Private Data Release for Data Mining"ACM KDD'11, August 21–24, 2011, San Diego, California, USA.

[10] Ruchi.P.Kanekar1, Prof. Rachel Dhanaraj2" Adding Dummy Items To Hide Sensitive Association Rules" National Conference On Advances In Computational Biology, Communication, And Data Analytics 6 | Page (ACBCDA 2017)

[11] Apoorva Joshi,, Pratima Gautam "A survey on Sanitizing Methods in Association Rule Hiding Technique"International Journal of Scientific Research in Computer science,Engineering and information Technology",volumwe 2 Issue 6 2017

[12] ARX Data Anonymization Tool https:\\arx.deidetifier.org

[13]Samarati P, Sweeney L (1998). Generalizing data to provide anonymity when disclosing information (Abstract). In Proc. of the 17th ACM-SIGMOD- SIGACT-SIGART Symposium on the Principles of Database Systems, p. 188, Seattle, WA, USA.

[14]H.C.Koh, G.Tan,"Data mining applications in health care", Journal of Healthcare Information Management Vol. 19, No. 2, pp. 64-73.

[15]Pingshui Wang, Jiandong Wang, Xinfeng Zhu ,Jian Jiang "Research on Privacy Preserving Data Mining" 2012 International Conference on Biological and Biomedical Sciences Advances in Biomedical Engineering, Vol.9.

[16]Saygin Y, Verykios VS, Elmagarmid AK. Privacy preserving association rule mining. In: Proceedings of the 12th International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/EBusiness Systems (RIDE'02); 2002, 151.

[17]Saygin Y, Verykios VS, Clifton C. Using unknowns to prevent discovery of association rules. ACM SIGMOD Record 2001, 30:46–54.

[18]K. Alotaibi, V. J. Rayward-Smith, W. Wang and Beatriz de laIglesia, "Non-linear Dimensionality Reduction for Privacy-Preserving Data Classification" in proceedings of 2012 ASE/IEEE International Conference on Social Computing and2012 ASE/IEEE International Conference on Privacy,Security,Risk and Trust, IEEE 2012. [7] H.

[19]Kargupta and S. Datta, Q. Wang and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", in proceedings of the Third IEEE InternationalConference on Data Mining, IEEE 2003.