IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 18-21
www.iosrjournals.org

Data Preservation Using Anonymization Based Privacy Preserving Techniques – A Review

S.Dhanalakshmi[1], P.S.Ahammed Shahz Khamar[2]
[1]*Assistant Professor, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College*
[2] *III M.Sc Software Systems, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College*

Abstract: *Advancement in database and networking technologies had made information storage and data sharing easier. The data mining techniques are used to access the shared data either in centralized or in the distributed environment for knowledge discovery. The knowledge extracted can be used by the shared organization to have mutual benefits. If the data used in mining contains the person specific information, it is important to protect the customer personal data. At the same time the data is to be utilized to guarantee valid analysis results. Privacy preserving data mining focus on providing protection to the personal information stored in the database and also to provide valid data mining results without violating the privacy. Many techniques have been explored in preservation of privacy data during data mining process. In this paper reviewed the existing anonymization based preservation techniques. The approaches used in anonymization are addressed highlighting the advantages and disadvantages of each method.*

Keywords*: Data Security, Privacy, k-anonymity, l-diversity, t-closeness*

I. Introduction

To extract valuable knowledge via data mining, first data are gathered from various data owners and then mining techniques are applied. But due to the rising concern over individual's privacy has established many challenges in data mining.

Privacy preserving data mining deals with how to modify the data such that modified data ensure valid data analysis result and also to guarantee the privacy right of the data owners. Many techniques have been designed to preserve privacy during mining process. In this paper surveyed the group anonymization based preservation techniques[1].

II. Anonymization

Anonymization based perturbation method makes the individual record to be indistinguishable among a group of records. In dataset the attributes can be grouped as unique identifiers, sensitive attributes and quasi identifiers. Unique identifiers identifies the individuals uniquely (e.g. name, social security number), sensitive attributes contains private information that should be protected and quasi identifiers are a set of attributes when linked to external data set can re-identify individual records.

In many applications the data records are released after removing the unique identifiers. Table 1 is the original medical data set and Table 2 is the published medical data set after removing the unique identifier attribute (name). However the quasi identifiers (such as age, sex, zip) in the data set when linked with other external databases (e.g.: Voters dataset) can re-identify the individual records as shown in Table 2 and Table 3.

2.1 K-Anonymity

One solution to this problem is group anonymization using k-anonymity model proposed by Sweeney [2]. In his work to preserve the privacy of individuals proposed a K-anonymity model which provides k-anonymity protection, i.e. the information of each individual contained in the released data set cannot be distinguished from at least k-1 individuals in the database.

Table 1. Medical data

#	Name	Sex	Age	Nationality	Zip	Disease
1	Arnold	M	25	Indian	641054	Gastritis
2	Saron	M	21	Indian	641064	Viral Infection
3	Bob	M	29	American	641162	Heart Disease
4	John	M	39	American	641064	Cancer
5	Jerry	F	32	Russian	641162	Cancer
6	Alice	M	36	Russian	641162	Cancer
7	Mary	F	52	Indian	671890	Viral Infection
8	Joice	F	50	Indian	671892	Cancer
9	Anie	F	55	American	671892	Heart Disease

Therefore for every group of values of the quasi-identifiers in the k-anonymous table there will be at least k records share the same values. Thus K-anonymous model make sure that individuals cannot be uniquely identified by external linking attacks. K-anonymity uses the techniques such as Generalization and Suppression. The k-anonymous table of Medical Table 2 is shown in Table 4. In Table 4 group of attributes {age, nationality and zip} is considered as quasi-identifiers. Thus the age attribute is generalized into 3 categories and nationality attribute and zip attribute are suppressed.

Table 2. Published Medical Data (Unique Identifier Removed)

#	Sex	Age	Nationality	Zip	Disease
1	M	25	Indian	641054	Gastritis
2	M	21	Indian	641064	Viral Infection
3	M	29	American	641162	Heart Disease
4	M	39	American	641064	Cancer
5	F	32	Russian	641162	Cancer
6	M	36	Russian	641162	Cancer
7	F	52	Indian	671890	Viral Infection
8	F	50	Indian	671892	Cancer
9	F	55	American	671892	Heart Disease

Table 3. External data - voters list

#	Name	Sex	Age	Nationality	Zip
1	July	F	55	American	641032
2	John	M	35	American	641033
3	Bob	M	29	American	641162
4	Bibin	M	35	American	641178
5	Joy	F	45	American	641262

Generalization transforms the original data set to perturbed data set by changing the original data values into generalized values. The record values are generalized to a specific range in order to reduce the granularity of representation. Generalization is depicted with the help of taxonomy tree, the node values can be replaced by a node lying in the path between itself or with the parent node. Generalization for numeric attribute can also be discretized into disjoint interval values.

In the suppression method the value of the attributes is completely removed before the data set is released for analysis. Both the generalization and suppression can be applied globally or locally to the data set. When applied globally the same type of transformation is done to all items of the data set. When applied locally the transformation is done to specific transactions of the dataset. The Transformed information done by the group anonymization method is true, but to some extent results in information loss.

Table 4. 3-Anonymous Data of Medical Data

#	Age	Nationality	Zip	Disease
1	<30	*	641***	Gastritis
2	<30	*	641***	Viral Infection
3	<30	*	641***	Heart Disease
4	3*	*	641***	Cancer
5	3*	*	641***	Cancer
6	3*	*	641***	Cancer
7	>40	*	671***	Viral Infection
8	>40	*	671***	Cancer
9	>40	*	671***	Heart Disease

Various algorithms have been proposed for implementing k-anonymity for data mining analysis. Wang et al. [3] proposed a secure data integration of multiple databases for classification analysis which satisfy the k-anonymity requirement. Wong, R. C. et al. [4] proposed a method that extends the k-anonymity model to (α, k) anonymity model. This model protects both identifications and sensitive associations in the disclosed dataset.

Chiu et al. [5] proposed a C-Means Clustering technique for the k-anonymity model. The model considers only the numeric quasi identifier attribute. The weight of each quasi-identifier attribute is adjusted so that the data distortion is controlled and also increases the quality of clustering results.

2.2 l-Diversity

Machanavajjhala et al.[6] pointed out that k-anonymity method is susceptible to homogeneity attack and background knowledge attack. For example if the adversary knows that his neighbour is in the age of 30 and his zip code 641162, from the anonymized published data set (refer Table 4) easily identifies that his neighbor record falls in the record numbers 4,5,6. Note the sensitive attribute disease column it is same for all the three records 4, 5, 6. So the attacker concludes that his neighbor is suffering from cancer.

So to overcome this type of attacks Machanavajjhala et al.[6] proposed the l-diversity technique which focuses on maintaining the diversity of the sensitive attributes. An equivalence class is said to have l-diversity if there are at least L well represented values for the sensitive attribute. The l-diversity technique is applied in various types like distinct l-diversity, entropy l-diversity and recursive l-diversity. In distinct l-diversity there will be L distinct values in each equivalence class. In entropy l-diversity there will be L distinct values and also evenly distributed values. In recursive diversity most frequent values do not appear too frequently. 2-diversity medical table for 3-anonymous Table 4 is shown in Table 5 in which at least two distinct values are there for the sensitive attribute Disease.

The L- diversity method also has some disadvantages. Achieving l-diversity may be hard or it may be unnecessary to achieve is some cases and in some cases l-diversity lacks to prevent attribute disclosure.

Li. N. et al. [7] pointed out that l-diversity technique is susceptible to skewness attack and similarity attack.

Table 5. 2-diversity Medical Data

#	Age	Nationality	Zip	Disease
1	<40	*	6410**	Gastritis
2	<40	*	6410**	Viral Infection
3	<40	*	6410**	Cancer
4	<40	*	6411**	Heart Disease
5	<40	*	6411**	Cancer
6	<40	*	6411**	Cancer
7	>40	*	6718**	Viral Infection
8	>40	*	6718**	Cancer
9	>40	*	6718**	Heart Disease

2.3 t-Closeness

t-closeness [6] is the enhancement of the l-diversity concept. The principle of t-closeness technique is that the equivalence class will have t-closeness, such that the distance between the distribution of a sensitive attribute in that class should be close to their distribution of the attribute in the entire database and should no more than a threshold t.

III. Research Trends In Privacy Preservation

Several privacy preservation research works have been carried out when collaborative mining is done in the field of medical research and business activities. In recent years lots of privacy preserving research works are explored in various fields. Recent works include preserving privacy in social networks against vulnerabilities of misuse, privacy issues in location based services in mobile networks to preserve the privacy of mobile clients and also about the privacy issues in web applications such as e-commerce and stream data mining.

IV. Conclusion

The main objective of privacy preservation techniques is to ensure higher accuracy and also guarantee privacy requirements. In this paper surveyed the approaches used in privacy preserving data mining using group anonymization based techniques. Each method is analyzed with its advantages and disadvantages.

All the anonymization methods try to make the individual record indistinguishable among a group of records. But it results in information loss. Also the k-anonymity method is at the risk of the background knowledge attack and l-diversity method is at risk to similarity attack. Finally, the survey is concluded with the situations where privacy preserving data mining is being used and also pointed out the current research trends of privacy preservation methods.

References

[1] Aggarwal, C. C., Yu, P. S, A general survey of privacy-preserving data mining models and algorithms. Privacy- preserving data mining, 2008, 11-52.

[2] Sweeney, L, k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10 (05), 557-570.

[3] Wang, K., Fung, B., Dong, G,. Integrating private databases for data analysis. Intelligence and Security Informatics, 2005, 23-41.

[4] Wong, R. C. W., Li, J., Fu, A. W. C., Wang, K., (α, k) -anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, 754-759.

[5] Chiu, C. C., Tsai, C. Y, A k-anonymity clustering method for effective data privacy preservation. Advanced Data Mining and Applications, 2007, 89-99.

[6] Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M, l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1 (1), 3.

[7] Li, N., Li, T., Venkatasubramanian, S, t-closeness: Privacy beyond k-anonymity and l-diversity. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference, 2007, 106-115.