Efficient Anonymization Algorithm for Multiple Sensitive Attributes

S.Srijayanthi, T.Sethukarasi, A.Thilagavathy

Abstract: ***The data of medical applications over the internet contains sensitive data. There exist several methods that provide privacy for these data. Most of the privacy-preserving data mining methods make the assumption of the separation of quasi-identifiers (QID) from multiple sensitive attributes. But in reality, the attributes in a dataset possess both the features of QIDs and sensitive data. In this paper privacy model namely (vi…vj)-diversity is proposed. The proposed anonymization algorithm works for databases containing numerous sensitive QIDs. The real dataset is used for performance evaluation. Our system reduced the information loss for even huge number of attributes and the values of sensitive QID's are protected.***

Keywords : quasi-identifiers, sensitive attribute, anonymization..

I. INTRODUCTION

Privacy is the capability of a person to establish what information is able to be shared, and can make use of right of direct access. When the information exist inside public domain in that case it becomes a danger to individual entities privacy because the information can be apprehended through the data holder. There arises a need that the accountability of the data holder is to make sure the privacy of m customers' data. Other than the information contained in public domains, the customers may sometimes donate to data leakage. The various key privacy threats are disclosure, personal embracement and abuse, surveillance and discrimination.

The data involving the personal information of individuals need to be given a major security. It involves the protection of data from hackers and non authorized users of the data.

Two variants of anonymization algorithm exist. One is the syntactic approach model and the other is the probabilistic approach model. In syntactic method, the data is well preserved with structured format. Any privacy violations on these data can be easily found by inspecting the data manually. In the probabilistic method, data discrepancies are introduced to corrupt the existing data. Hybrid approach involves combining the syntactic and probabilistic methods.

Revised Manuscript Received on November 08, 2019.

***** Correspondence Author

S.Srijayanthi, Assistant Professor, Department of CSE, R.M.K. Engineering College Kavaraipettai, Tamil Nadu, India. Email: ssj.cse@rmkec.ac.in

T.Sethukarasi, Professor, Department of CSE, R.M.K. Engineering College Kavaraipettai, Tamil Nadu, India. Email: tsk.cse@rmkec.ac.in

A.Thilagavathy, Associate Professor, Department of CSE, R.M.K. Engineering College Kavaraipettai, Tamil Nadu, India. Email: atv.cse@rmkec.ac.in

Providing high security to the data with less information loss is a challenging job.

Numerous Privacy preserving methods exist and many methods work on anonymization of data. The different privacy preservation approaches are k-anonymity, l-diversity, t-closeness, Randomization, Cryptographic techniques, Multidimensional Sensitivity Based Anonymization (MDSBA). Many of the existing system of privacy preservation are based on structured data. But majority of the data are unstructured [1]. Bottom up Generalization [2] and Top down Generalization [3] are the conservative approaches of Anonymization for structured data records. The challenges in privacy preservation are as follows:

- To develop tangible result towards protection of privacy in data whether the data is structured or unstructured
- To develop Robust and vigorous procedure for holding huge level of mixed data sets.

Data ought to be permitted to continue in its indigenous form devoid of alteration and data analytics be able to be accepted at the same time ensuring privacy preservation.Data usage must be maximized and at the same time ensuring data privacy.

Nowadays, huge voluminous of data related to personal information of people are being collected and distributed in the Internet. These personal information are being shared to create services and applications and for other management purposes. There is a dire need for the preservation of privacy of data since these data need to be maintained for they are being used in research [4] as well. Data holder is the term used for any institute that holds the original database. Data Analyzer is the term used for those institutes that receive and share this database. The data analyzers perform anonymization on these databases for their use. Many methods prevail in the anonymization of databases involving personal information. Many recent studies have proved that the data holder considers the database in terms of explicit identifiers, quasi-identifiers (QIDs) and sensitive attributes.

Explicit identifiers are those attributes to facilitate explicitly identification of individual person (eg.name). Quasi identifiers are those that are united with additional new attributes to facilitate identification of individual person (eg. zipcode and age). The sensitive attributes are those delicate individual attributes existing in the private environment (eg. salary)[5]. After eliminating each and every one of the unambiguous identifiers present in the database, disclosure can possibly still happen. k-Anonymity [6], l-diversity [7], and t-closeness [8] are the various key privacy models to avoid this problem.

Several works taking place on the privacy models are proposed, such as [9], [10], [11], [12].

In [13] the noise reduction achieved depends on the amount in which the attributes need to be protected. The method is scalable as the data set is increased in its volume.[14] proposed a robust privacy model named β-likeness. The privacy guarantee is expressed as a limit on the relative confidence gain on each single sensitive attribute value. The authors in [15] proposed IRSWAP algorithm. The algorithm creates diverse anonymized databases intended for every data analyser. Wan et al. [16] had proposed a model FF-anonymity. Here, they argued that attribute belonging to both QIDs and sensitive attributes is important. Jin et al. [17] combined the features of sensitive attributes and QIDs. Multiple sensitive attributes was used in Ye et al. [18] which forgoes the relationship among sensitive attributes. Quasi-sensitive attributes was introduced by Shi et al. [19] which is an insensitive attribute but becomes sensitive when used in combination. In [20], the sensitive attributed can be split from the QIDs.

II. PROPOSED METHOD

There exist two steps in the proposed methodology: Step 1 is performed by the data holder where a randomization process is done. The step 2 is performed by the data analyzer where a reconstruction is done. Here, the parameters are $(m_i…m_j)$ and $(e_i…e_j)$ based on the values $(v_i…v_j)$. Based on these parameters, an aggregated expression is formed from each record. The record is then inserted into the anonymized database. The parameters vary for each privacy model. The sensitive QIDs to be analyzed are first determined by the data analyzer.

The proposed algorithm be able to be used intended for several frequency $(v_i…v_j)$ – diversity. The example below discusses the frequency$(v_i…v_j)$.

Consider the table 1(a) with a single record that contains the name, gender, age, city, disease of patient 'A'. Let l1=l2=l3=l4 and that the anonymization also generates $R_{1,1}$ ={M,F}, $R_{1,2}$ = {28,40},$R_{1,2}$={Delhi, Chennai} and $R_{1,4}$={HIV, Flu}. The result of the algorithm is shown in Table 1(b) .The true record of the patient is depicted in record 3.The data analyzer after knowing any two of the sensitive QIDs for patient 'A', the data analyzer cannot indicate the patient 'A' significance to additional sensitive QID by means of a confidence larger than half.

The data holder inserts the record and the quasi identifier as input to the system. The Cartesian product of the extracted values is obtained as the output from the system. For each record, distinct values are extracted randomly from the set of quasi identifiers. For each sensitive quasi identifier, the set of extracted values are created along with maintaining the original value. After computing the Cartesian product, it is inserted into the anonymized database. The above whole procedure is done for each record.

Algorithm: v-Diversity Algorithm

Input: record r_i , $1 \le i \le n$ from Dataholder DH and QID Qid_j
Output: cartesian product of extracted values
Step 1: for each r_i do
Extract V_{j-1} distinct values randomly from $D(Qid_j)\backslash E(r_i,Qid_j)$
Endfor
Step 2: Do
Create a set $R_{i,j}$ from the original set $E(r_i,Qid_j)$
Repeat step2 for every sensitive QID $S_i…S_j$
Step 3: for each r_i do
Calculate Cartesian product of $R_{i,1}…R_{i,q}$
insert every element into the database that is anonymized
Endfor
Step 4: Repeat above steps for every record

For example, assume that the patient A's age, gender is known to the data analyzer which is 'F' and the age is 40. Table 1(a) shows the details of patient 'A'. Here the patient has the disease Flu and stays in the city Chennai.

The data analyzer will not be able to predict if the 7th record or the 8th record in table 1(b) belongs to patient 'A' because both the rows have different values for disease. An collective grouping is generated by the data holder for the anonymized record since the dimension of the Cartesian product $R_{i,1}…R_{i,q}$ might be huge.

Table 1(b) shows the details of eight records. The attributes considered are gender, age, city and disease. From this table , it is inferred that the age lies between 39 and 42. The cities are Chennai, Bangalore and Delhi. The diseases are cancer, flu and HIV.

The records are grouped into much equivalence class. In table 1(b) the 1st and 4th rows make an equivalence similarity in favor of s_3 for the reason that every row has 'M' for s_1, 39 for s_2 and cancer for s_4. Similarly rows 2 and 3 construct an equivalence similarity in favor of s_4 for the reason that every row has F for s_1, 40 for s_2 and Chennai for s_3.Table 1(c) shows the collective grouping of table 1(b)

Table 1(a) : Patient A's original record

Name	Gender	Age	City	Disease
A	F	40	Chennai	Flu

Table 1(b) : Patient A's anonymized record

Gender	Age	City	Disease
M	39	Delhi	Cancer
F	40	Chennai	Cancer
F	40	Chennai	Flu
M	39	Bangalore	Cancer
F	40	Chennai	Cancer
M	42	Bangalore	Cancer
F	41	Chennai	HIV
F	41	Chennai	Cancer

Table 1(c) : Collective grouping of (b)

Gender	Age	City	Disease
{F,M}	{40,39}	{Chennai, Delhi}	{Flu, Cancer}

III. RESULT AND DISCUSSION

The experimentation was conducted on Pentium dual core system with 120 GB hard Disk, 15'' LED Monitor, 1 GB RAM and Keyboard, Mouse as Input Devices. The proposed system was implemented using JAVA/J2EEProgramming language with Windows 7 Operating system, Netbeans 7.2.1 Tool and MYSQL Database. Figure 1 gives the L1 distance against frequency. Figure 2 gives the L2 distance against frequency. From figure 1 and figure 2 it can be noted that the L1 and L2 distance decreased respectively.

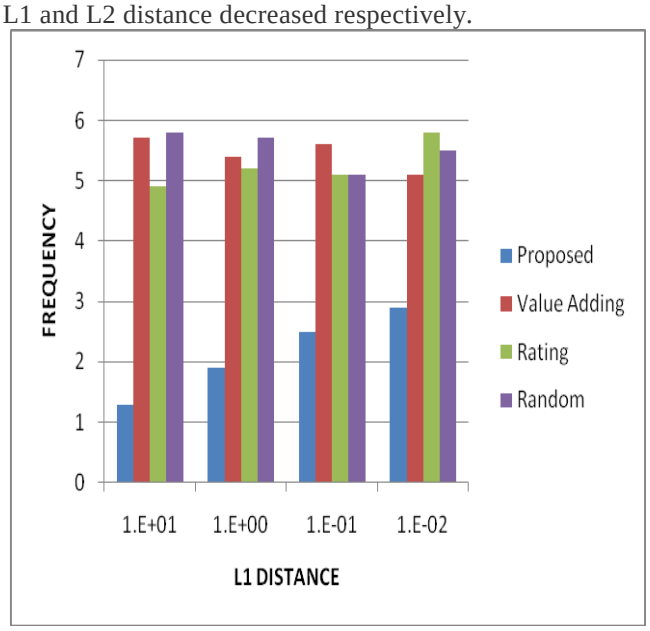

Figure 1: L1 Distance

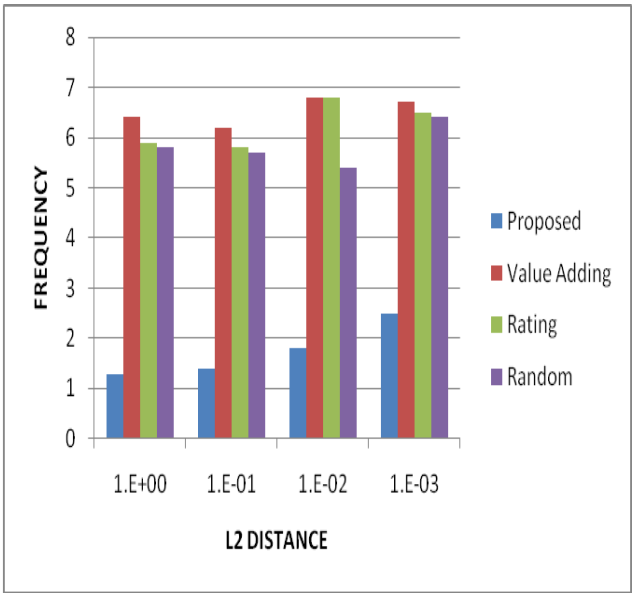

Figure 2: L2 Distance

The proposed method reduced the L1 distance and L2 distance compared to the existing methods. Our proposed algorithm increased the L1, L2 distances with the increase of very large frequencies.

IV. CONCLUSION

In this paper privacy model namely ($v_i \ldots v_j$)-diversity is proposed. The proposed algorithm works for databases containing numerous sensitive QIDs. The real dataset is used for performance evaluation. Our system reduced the information loss for even huge number of attributes and the values of sensitive QID's are protected

REFERENCES

1. G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
3. H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
4. B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
5. E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," *IEEE Trans. Antennas Propagat.*, to be published.
6. J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.
7. C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
8. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," *IEEE Transl. J. Magn.Jpn.*, vol. 2, Aug. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].
9. M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.
10. (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). *Title* (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)
11. J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com
12. (Journal Online Sources style) K. Author. (year, month). Title. *Journal* [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)
13. D. Sanchez, J. Domingo-Ferrer, S. Martinez, and J. Soria Comas, "Utility-preserving differentially private data releases via individual ranking microaggregation," Information Fusion, vol. 30, pp. 1–14,2016.
14. J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee," Proc. VLDB, vol. 5, no. 11, pp. 1388–1399, 2012.
15. J. Soria-Comas and J. Domingo-Ferrer, "Probabilistic k-anonymity through microaggregation and data swapping, "IEEE International Conference on Fuzzy Systems,pp. 1–8, 2012.
16. K. Wang, Y. Xu, A. W. C. Fu, and R. C. W. Wong, "FF-Anonymity: When Quasi-identifiers Are Missing," IEEE ICDE, pp. 1136–1139,2009.
17. X. Jin, M. Zhang, N. Zhang, and G. Das, "Versatile publishing for privacy preservation," ACM KDD, pp. 353–362,2010.
18. Y. Ye, Y. Liu, C. Wang, D. Lv, and J. Feng, "Decomposition: Privacy Preservation for Multiple Sensitive Attributes," DASFAA, pp. 486–490,2009.
19. P. Shi, L. Xiong, and B. Fung, "Anonymizing data with quasisensitive attribute values," ACM CIKM, pp. 1389–1125 ,2010.
20. J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, "t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 11, pp. 3098–3110, 2015.