International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 8 (2017) pp. 1639-1663

Preserving the Privacy of Sensitive Data using Data Anonymization

M. Jyothi[1] and Dr. M. V. P. Chandra Sekhara Rao[2]

[1] Assistant Professor, Department of IT, GMRIT, Rajam, Srikakulam, Andhra Pradesh, India.
[2] Professor, Department of CSE, RVR & JC College of Engineering, Guntur, AP, India.

Abstract
Objectives: Health care and financial data are very sensitive. There are many methods to provide privacy to the dataset. The objective of this paper is to run the k-anonymity method using arx tool
Methods: This paper is mainly concentrated on anonymization method with is used to provide privacy to the dataset so that the attacker will not gain any sensitive information about the individuals. Anonymization is the best method to provide privacy when compared to the other methods like randomization, perturbation etc. Anonimization can be done in many ways; there are several tools available to perform anonymization.
Findings: Observe the differences between original data and anonymized data. K-anonymity prevents from linking attack using generalization and suppression techniques.

Keywords: Privacy-Preserving, Data Mining, Anonymization, K-Anonymity, l-diversity.

INTRODUCTION

Health care dataset contains some sensitive information like the patient's information with attributes Name, Zip code, Age, Sex and disease. This original dataset will not provide any privacy. This sensitive data may be exposed during the data mining process and it is possible to learn lot of information about individuals from public data. So, to provide privacy the dataset should be modified. Privacy preserving data mining is first introduced by Agarwal and Srikanth[1]. One type of modification is to apply anonymization operations on the dataset. Anonymization is the best method to provide privacy when compared to the other methods like randomization, perturbation etc[2]. Data anonymization is the process of either encryption or removing personal identifiable information from datasets, to provide privacy to the person identification. Anonymization uses two operations generalization and suppression. Generalization means that individual attributes are replaced with a broad category, while suppression means removing the value. Generalization is used for converting categorical attributes, discrete numeric attributes and continues numeric attribute.

To transform the dataset using the anonymization method we have to use some rules in generalization hierarchies. Some anonymization methods use local recoding, some other use global recoding.

Local recoding means different rules can be applied for equal data items whereas global recoding means same rule for all the data items. This global and local recoding can be single – dimensional recoding, data items are values of an individual column, or multi-dimensional r coding, in which data items are combinations of values for different columns.

The attributes in the data set are categorized into personal identification attributes, quasi-identifiers and sensitive attributes. Personal identification attributes identify the persons directly. A set of attributes that can linked with external data to uniquely identify individuals in the population are called quasi-identifiers. Sensitive attributes hold sensitive information.

For example consider a hospital dataset which contains the patient's information with attributes Name, Zip code, Age, Sex and disease as shown in table 1. In Table 1, Name attribute is the personal identification, Disease is the sensitive attribute. Suppose to provide privacy to the data set if we remove the name the table 1 will be modified to table 2.

Table 1: Hospital dataset micro data

Name	Zip code	Age	Sex	Disease
Alice	47677	29	M	Ovarian Cancer
Boby	47678	22	M	Ovarian Cancer
Peter	47602	27	M	Prostate Cancer
Emelee	47909	43	M	Flu
Holdon	47905	32	F	Heart Disease
Cloyce	47906	47	M	Heart Disease

Table 2: Hospital dataset micro data after removing Name attribute

Zip code	Age	Sex	Disease
47677	29	M	Ovarian Cancer
47678	22	M	Ovarian Cancer
47602	27	M	Prostate Cancer
47909	43	M	Flu
47905	32	F	Heart Disease
47906	47	M	Heart Disease

Table 2 does not explicitly indicate the names of patients. However, if an adversary has to access to voter registration list in Table 3, he can easily discover the identities of all patients by joining the two tables on {Age, Sex, Zip code}. These three attributes are, therefore, the quasi-identifier attributes. From table 2 and table 3 using quasi-identifiers an

outsider can easily say that it is peter who is suffering from Prostate cancer.

Table 3: Voters dataset micro data

Name	Zip code	Age	Sex
Alice	47677	29	M
Peter	47602	27	M
Holdon	47905	32	F
Dan	47912	29	M
Boby	47678	22	M
carrol	47900	50	M
Ellen	47930	24	F

So removing the personal identification information will not provide complete privacy to the data[3]. To provide privacy to the dataset first we have to remove the personal Identification information and then we have to anonymize the quasi-identifiers. The sensitive attributes should always be released directly because researcher's want this information. Different privacy preserving methods have been proposed[4-8].
To anonymize the quasi-identifiers we can use any one of the following approaches: K-anonymity[9, 10] or l-diversity.

K-ANONYMITY

This method is used to anonymize the quasi-identifiers to provide privacy to the data[11, 12]. The approach is as follows: The information for each person contained in the released dataset cannot be distinguished from at least k-1 individuals whose information also appears in the data. For example: if an attacker with the only information of birthdates and gender is trying to identify a person in the released dataset. There are k persons in the table with the same birth date and gender. In k-anonymity any quasi-identifier present in the released table must appear in at least k records.
The goal of K-anonymity is to make each record indistinguishable from at least k-1 other records. These K-records form an equivalence class. K-anonymity uses generalization and suppression. Using generalization, k-anonymity replaces specific quasi-identifiers with less specific values until it gets K identical values. And it uses suppression when generalization causes too much information loss, which is referred as outliers.
Form the table 1 we have 3 quasi-identifiers which can be generalized as shown in the figure 1

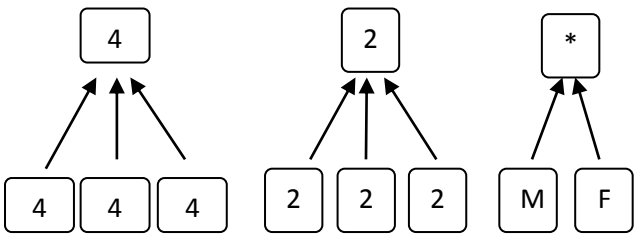

Figure 1: Generalization on Quasi-identifiers like zip code, age and sex

By applying k=2 anonymity and quasi-identifier {zip code, Age, sex} on table 2 we will get table 4. Now if we compare table 2 and table 4 it is difficult for an outsider to find the sensitive information because there are three people with generalized zip code and age. In table 4 first three records form one equivalence class and last two records another equivalence class.

Table 4: k-anonymity on table 2

Zip code	Age	Sex	Disease
476**	2*	M	Ovarian Cancer
476**	2*	M	Ovarian Cancer
476**	2*	M	Prostate Cancer
479**	3*	F	Heart Disease
479**	4*	M	Flu
479**	4*	M	Heart Disease

Any records which has not come into any equivalence class should be suppressed. In table 4 record 4 will not fall into any of the equivalence class so it should be suppressed. All applying the generalization and suppression on table 3 it results to a Table 5

Table 5: Generalization and suppression on table 3

Zip code	Age	Sex	Disease
476**	2*	M	Ovarian Cancer
476**	2*	M	Ovarian Cancer
476**	2*	M	Prostate Cancer
*	*	*	*
479**	4*	M	Flu
479**	4*	M	Heart Disease

The problem with the k-anonymity is it will not provide privacy if sensitive values in an equivalence class lack diversity and also if the attacker has background knowledge. Consider Table 6 the first 3 records which form an equivalence class have the same sensitive attribute values where there is no privacy and attacker can direct gain the information. And last three records if attacker has some background knowledge about the person (ex. The person father is a heart patient) then this information may be useful for the attacker to gain the sensitive information[13].

Table 6: Sensitive attributes lack diversity

Zip code	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease

4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

L-DIVERSITY

L-diversity solves the k-anonymity problems with equivalence class sensitive values[14]. L-diversity says that sensitive attributes must be "diverse" within each quasi-identifier equivalence class. L-diversity is defined as each equivalence class has at least l well-represented sensitive values. Table 7 shows 3-diversity which has two equivalence class with diverse sensitive attributes.

Table 7: 3-diverse patient's information

Zip code	Age	Disease
476**	2*	Heart Disease
476**	2*	Cancer
476**	2*	Flu
476**	3*	Flu
476**	3*	Cancer
476**	3*	Heart Disease
4790*	≥40	Flu
4790*	3*	Heart Disease
4790*	2*	Cancer
4790*	2*	Flu
4790*	≥40	Heart Disease
4790*	3*	Cancer

Equivalence class

Equivalence class

L-diversity can be probabilistic l-diversity, entropy l-diversity or recursive (c,l)- diversity. According to probabilistic l-diversity the frequency value in an equivalence class is bounded by 1/l. Entropy l-diversity says that the entropy of the distribution of sensitive values in each equivalence class is at least log (l) and recursive (c, l) diversity says that the most frequent value does not appear too frequently.

But l-diversity does not consider the semantic meaning of sensitive attributes example gastric ulcer and stomach cancer are considered as two values but semantically same, so sometimes attacker may gain some sensitive information. Other anonymization techniques are also available[15].

DATA ANONYMIZATION TOOL

To perform anonymization on dataset many tools are available[16]. This paper explains about ARX anonymization tool[17]. For analyzing we have selected adult dataset. Adult dataset consists of 15 attributes and 30477 tuples. For applying anonymization on this dataset, using ARX tool, out of 15 attributes 9 attributes are considered. The 9 attributes are {age, workclass, education, marital status, occupation, sex, race, native country, salary}. Out of these 9 attributes {age, sex} are quasi-identifiers and {occupation} is a sensitive attribute.

ARX provides user interface to anonymize the data. As a first step data set has to be imported. Before importing the user has to create a project. ARX supports .csv, .xls, .xlsv. After importing the dataset we have to apply the k-anonymity method to anonymize the dataset. To use the k-anonymity we have to generalize the dataset. After this we have to set the property of each attribute as sensitive, quasi-identifier or insensitive. To generalize the hierarchy is created for each attribute which defines the privacy level. In this paper, hierarchy is defined for age attribute. After creating the hierarchy K-anonymity can be applied. . In the similar fashion hierarchies can be defined to the other quasi-identifiers to provide privacy. As an example k=2 anonymity is applied and the comparison of are attribute for original dataset and anonymized dataset is show in fig 2 and 3 respectively.

Now this anonymized dataset can be exported and is saved as csv format. On this anonymized data mining applications can be applied using the weka tool. The weka tool support only arff format. First the anonymized data in csv format should be converted to arff format and then the data set can be used in weka.

Age	Work class	Education	Marital status	Occupation	Race	Sex	Native Country	Salary
39	State-gov	Bachelors	Nevermarried	Adm-clerical	White	M	United stastes	<=50k
50	Self-emp-not-inc	Bacchelors	Married-civ-spouse	Exec-managerial	White	M	United states	<=50k
38	Private	Hs-Grad	Divorced	Handlers Cleaners	White	M	United States	<=50k
53	Private	11th	Married-civ-spouse	Prof-Specialty	Black	M	United-states	<=50k
28	Private	Bachelors	Married-civ-spouse	Prof-sleciality	Black	F	Cuba	<=50k
37	Private	Masters	Married-civ-spouse	Exec-managerial	White	F	United States	<=50k
49	Private	9th	Married-spouse-absent	Other-service	Black	F	Jamica	<=50k

Figure 2: Micro Data for Original Dataset

Sex	Age	Race	Marital status	Education	Native Country	Work class	Occupation	Salary-Class
Female	[51-60]	White	Spouse not present	Higher education	North America	Government	Nontechnical	<=50k
Female	[51-60]	White	Spouse not present	Higher education	North america	Government	Nontechnical	<=50k
Female	[51-60]	White	Spouse not present	Higher education	North america	Government	Nontechnical	<=50k
Female	[51-60]	White	Spouse not present	Higher education	North america	Government	Nontechnical	<=50k
Female	[51-60]	White	Spouse not present	Higher education	North america	Government	Nontechnical	<=50k
Female	[51-60]	White	Spouse not present	Higher education	North america	Government	Nontechnical	<=50k
Female	[51-60]	White	Spouse not present	Higher education	North america	Government	Nontechnical	<=50k

Figure 3: Micro Data of Anonymized Dataset using ARX Tool.

CONCLUSION

Privacy is very important to protect the sensitive data from the attacker. To provide privacy to the data anonymization methods can be used. In this paper anonymization is done by using K-anonymity method in arx tool. The experimental results show the difference between the original and the anonymized data. On this anonymized datamining applications can be applied using the weka tool.

REFERENCES

[1] Agrawal, R., Srikant, R. "Privacy preserving data mining". In: Proceedings of the ACM SIGMOD Conference of Management of Data, pp. 439–450. ACM 2000

[2] Jun-Lin Lin , Yung-Wei Cheng, "Privacy preserving item set mining through noisy items", Elsevier Expert Systems with Applications 36 (2009) 5711–5717

[3] L. Sweeney, Computational disclosure control – a primer on data privacy protection, Ph.D. thesis, Massachusetts Institute of Technology, 2001.

[4] C. Clifton, T. Tassa, on syntactic anonymity and differential privacy, Tans. Data Priv. 6 (2) (2013) 161–183.

[5] K. El Emam, E. Jonker, L. Arbuckle, B. Malin, A systematic review of re identification attacks on health data, PloS One 6 (12) (2011).

[6] A. Gkoulalas-Divanis, G. Loukides, J. Sun, Publishing data from electronic health records while preserving privacy: a survey of algorithms, J. Biomed. Inform. 50 (2014) 4–19.

[7] C. Dwork, Encyclopedia of Cryptography and Security, Springer, 2011. Chapter: Differential Privacy.

[8] Sridhar Mandapati, Dr. Raveendra Babu Bhogapathi and Dr. M.V.P.Chandra Sekhara Rao, Swarm optimization Algorithm for Privacy Preserving in Data Mining", International Journal of Computer Science Issues, Vol. 10, No. 2 (March 2013).

[9] Sweeney, L."Achieving k-anonymity privacy protection using generalization and suppression", International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 10(5), 571– 588 (2002)

[10] Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira "Efficient Multi-Dimensional Suppression for K-Anonymity", in proceedings of IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 3. (March 2010), pp. 334-347, IEEE 2010.

[11] L. Sweeney. "K-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[12] P. Samarati and L. Sweeney, "Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression," Technical Report SRI-CSL-98-04, 1998.

[13] TianchengLi ,Ninghui Li "Towards optimal k-anonymization", Elsevier, Data & Knowledge Engineering 65 (2008) 22–39

[14] Ashwin Machanavajjhala, Daniel kifer, Johannes gehrke, and Muthurama Krishnan Venkitasubramaniam"l -Diversity: Privacy

[15] Beyond k-Anonymity", ACM Transactions on Knowledge Discovery from Data, Vol. 1, No. 1, Article 3, Publication date: March 2007

[16] Ninghui Li, Yiancheng Li, and Suresh"t-Closeness: Privacy Beyond k-Anonymity and l-Diversity". IEEE 23rd International Conference on Data Engineering pp 106 - 115 2007

[17] Clifton, c., kantarcioglu, m., vaidya, j., lin, x., and zhu,m. Y. "Tools for privacy preserving data mining", Sigkdd explorations 4, 2, 28–34. 2002.

[18] Fabian Prasser*, Florian Kohlmayer*, Putting Statistical Disclosure Control Into Practice: The ARX Data Anonymization Tool. In: Gkoulalas-Divanis, Aris, Loukides, Grigorios (Eds.): Medical Data Privacy Handbook, Springer, November 2015. ISBN: 978-3-319-23632-2