© 2017 IJSRSET | Volume 3 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Overlapping Slicing a Method for Privacy Preservation in Medical Applications

Rakshatha V, Supriya Salian
Department of Computer Science and Engineering / SJEC / Mangalore, Karnataka, India

ABSTRACT

Overlapping slicing is the best technique for privacy preserving of the published data set, it also preserves the data usefulness by hiding the sensitive information. Overlapping slicing is also used to hide the individual detail without revealing the identity of any individual at the same time offers better privacy. The different techniques that are used for privacy preservation of data namely generalization, bucketization, Multi-set generalization, slicing and overlapping slicing. Overlapping slicing technique is used to compress the whole dataset in order to preserve the privacy without any separation between quasi attribute and sensitive attribute that hides the individual detail by displaying only the required information which will be helpful for data usefulness and to take further decision to improve medical related technologies and medicines.

Keywords: Privacy, Overlapping Slicing, Privacy Preservation Data Publication.

I. INTRODUCTION

Preserving the privacy for the healthcare data has been the most significant issue. Especially when the data gets published. So while publishing any kind of the data related to medical, certain tools and methods have been implemented to preserve the privacy while publishing the useful data. i.e. make it available for the public to make use of the data that is available. So when the medical data set is published we have to make sure or implement methods that will not reveal the patient's identity especially when we are dealing with very sensitive information about that patient and also if any adversary trying to gain any information about an individual. Published data becomes more useful if and only if the persons privacy is preserved. In this paper work we are making use of the technique called Overlapping Slicing, where it conserves the usefulness of data against privacy threats.

In recent years preserving of micro data privately has been extensively used in many fields such as person, household, or an organization. There are many anonymization techniques that have been proposed. The most popular data anonymization techniques are generalization and bucketization. In both approaches, attributes are partitioned into three attributes

- First types of attributes are identifiers that mainly focus on identifying the individual, like name or social security number.
- The second types of attributes are quasi identifiers which can be known by adversary possibly form public database when considered together can identify a individual like birthdate, zip code, sex.
- Third type of attributes is sensitive attributes, which may not be known to adversary and are considered as sensitive like disease and salary.

Figure 1: Anonymization Process

Here in the above figure data publisher refers to the hospital's dataset that contains original details about the hundreds of patients. Anonymization is a method that converts the text into a Unicode which is non-human readable form. The main goal is to protect the privacy of patient's dataset. In anonymization hiding a particular detail of the patients that will not reveal his/her identity and also use encryption technology so the described

details of any patient remain anonymous Finally data is been published to the medical center who is a data recipient and has also gained lot of importance in very recent decades and this is one of the best techniques to secure the privacy of the published datasets.

II. PROBLEM STATEMENT

In recent times the task of preserving the privacy for data that gets published becomes a very difficult task. So there is a high risk involved in revealing the personal information disclosure, when the data sets related to medical get published to third party vendors or any agencies. There are four main data privacy preserving methods.

- Generalization
- Bucketization.
- Multiset generalization.
- Slicing.

The first anonymization method is generalization which aims at f preserving the patient's identity. In generalization rows are partitioned into buckets and each column value will have unique column value, the tuple with each unique value of the column that will match only single bucket so here again it fails to preserves the privacy of the data that gets published by revealing the patient's identity.

Second anonymization technique is bucketization where it will not help in hiding the individual identity and also bucketization requires the clear separation between quasi attributes and sensitive attributes.

Third anonymization technique is Multiset generalization where each of the columns have the multi-set value represented in the form of ratio.
Fourth anonymization technique is called slicing where it helps to handle the high dimensional data and preserves the privacy better than generalization and bucketization .

In order to avoid drawback of first two methods overlapping slicing is used in this present work.
Steps involved in overlapping slicing are:

- Overlapping slicing technique is the extension of slicing technique where duplication of the attributes take place in one or more columns which will create more co-relation among the existing attributes.
- It will be a very difficult task for any adversary trying to gain any access about any individual.
- The privacy of the data is preserved without revealing the identity of any individual.

III. IMPLEMENTATION

1) Generalization

Generalization has a slight difference compared to the original data i.e. the last three digits of the DOB column is hidden, also the sex will be hidden, last three digits of zip code will also be hidden and finally grouping the age into different buckets based on the age limits that we specify. But rest of all the fields that present in the database remains the same and it reveals the identity of the patient.

2) Bucketization

In bucketization we are sorting the age group based on the limit specified by different buckets in generalization. Where it will sort according to the priority specified in the buckets with particular fashion and displays it and we are not making a clear separation between quasi attribute and sensitive attribute and it prints the quasi attribute in the real form as that of the original database. This will be the major cause of revealing the individual identity.

3) Multi-set generalization

In multi-set generalization it will display the single row containing multiple values i.e. the first column of multi-set generalization is displayed in the form of the ratio obtained from a single bucket by taking the count on number of females and males age, second column only based on the males and females sex, the third column zip code is also displayed in the same form of ratio depending on the number of zip code present in one particular bucket.

4) Slicing

In slicing technique we are grouping the highly co-related attribute into a single column and display it. The age and sex are considered to be the highly co-related attributes which can be combined and display as one column, and the other column which consists of sensitive attribute disease. The final table looks like the

whole sub-division of the original table wherein all the details, including the individual's identity is hidden.

5) Overlapped slicing

The final method of data anonymization is overlapping slicing technique which is extension of the slicing technique, which is used to duplicate the attribute in one or more column which will preserve the data in better manner; here in overlapping slicing the quasi attribute are grouped into one column and include a sensitive attribute disease in first column and it is duplicated and inserted in the second column to create a better co-relation among the attributes . So if any adversary trying to gain any information will not be able to differentiate between the original value and the duplicate attribute present in the table.

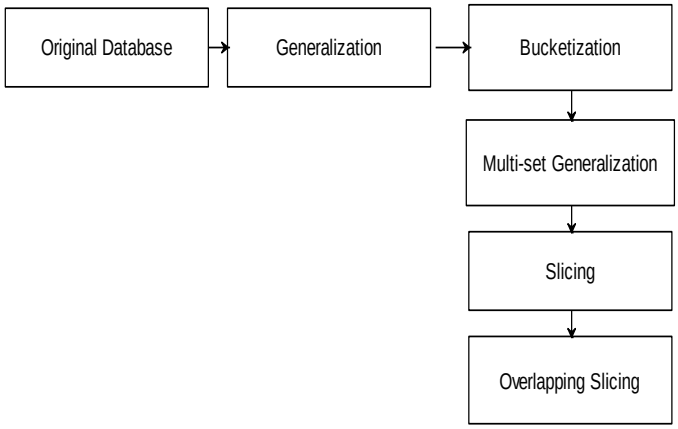


Figure 2: Overlapping slicing architecture

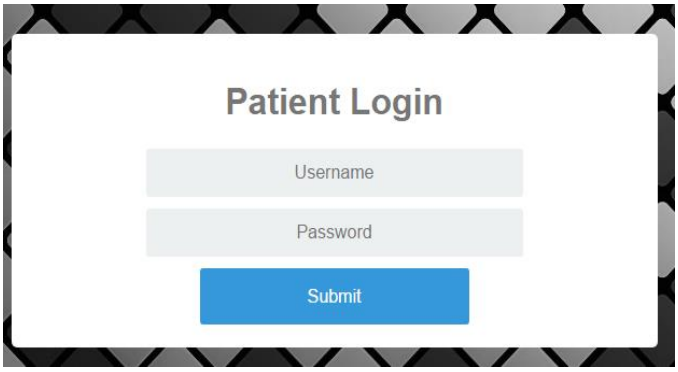
IV. RESULT

A) Homepage

The concept of hiding individual detail using overlapping slicing has been implemented and results are shown below. The proposed paper is implemented in Java technology. The home page describe the how the individuals privacy get disclosed and patient login, admin login is there.

B) Patient Profile and login:

Here this page provides the details of patient and options to change password and logout.



This is patient login page with option to enter his username and passsword.

C) Original data

This web page contains all the detail information about each individual like age, sex, zip code, DOB etc. printing such kind of detailed information to the outside world will bring threat to the individual's privacy.

D) Generalization

In generalization, the age is replaced with boundary value so that it difficult to get perfect age of a person and privacy is preserved.

Generalized Table

ID	NAME	BLOOD GROUP	DISEASE	EMAIL	MOBILE	CITY	DATE OF BIRTH	AGE	GENDER	ZIP CODE
1	gana	o+	Fever	xyz@gmail.com	9234567890	mlr	199***	1-30	*	575***
2	sara	a-	cancer	aa@gmail.com	7734567850	mlr	199***	1-30	*	575***
3	fahara	o+	StomachPain	qwerty@gmail.com	9134567850	mlre	199***	31-60	*	581***
4	dhanush	ab-	ankle swollen	dhanu@yahoo.com	9741214515	bangl	199***	31-60	*	587***
5	harish	b+	ulcers in mouth	har@gmail.com	9548761325	mlre	198***	1-30	*	575***
6	ashok	b-	knee swelling	ashu@gmail.com	9548761325	mlre	199***	31-60	*	575***
7	sahana	a+	lumps in breast	sana@gmail.com	9234567890	bangl	198***	1-30	*	575***
8	banni	o-	cold	banni@yhaoo.com	9741214515	mlre	199***	1-30	*	581***

E) Bucketization

Shows the Bucketized table with sorted age group and based on the priority given to female or male the output is displayed, quasi attributes are grouped in one column and disease is casually permuted in other column.in bucketization buckets are created in manner where 1-40 age group falls in the first bucket and 40-100 falls in the second bucket.

Bucketized Table

ID	NAME	BLOOD GROUP	EMAIL	MOBILE	CITY	GENDER	DATE OF BIRTH,AGE	DISEASE, ZIP CODE
8	banni	o-	banni@yhaoo.com	9741214515	mlre	female	(1995-07-28,24)	(cold,581343)
7	sahana	a+	sana@gmail.com	9234567890	bangl	female	(1988-02-22,25)	(lumps in breast,575002)
21	aaa	o+ve	aaa@gmail.com	1234567890	mangalore	female	(2017-03-11,25)	(Fever,575003)
2	sara	a-	aa@gmail.com	7734567850	mlr	female	(1990-04-05,26)	(cancer,575003)
9	vidya	a+	vidya@gmail.com	9548761325	bangl	female	(1992-06-15,27)	(blood-cancer,575015)
1	gana	o+	xyz@gmail.com	9234567890	mlr	female	(1995-11-10,28)	(Fever,575002)
10	hanitha	ab-	hanni@yhaoo.com	9234567890	mlre	female	(1988-02-22,29)	(StomachPain,581343)
5	harish	b+	har@gmail.com	9548761325	mlre	male	(1988-02-22,30)	(ulcers in mouth,575003)
4	dhanush	ab-	dhanu@yahoo.com	9741214515	bangl	male	(1990-04-05,32)	(ankle swollen,587620)
3	fahara	o+	qwerty@gmail.com	9134567850	mlre	female	(1995-07-28,33)	(StomachPain,581343)
11	harini	o-	harini@gmail.com	9234567890	mlr	female	(1995-07-28,36)	(ankle swollen,581343)

F) Multi-set Generalization

Shows the Multiset Based Generalization table which is displays a multi-set value in the form of ratio. Based on the no of patient's belonging to respective buckets, gender is also displayed in the ratio form and finally the zip code.

Multiset Based Generalization Table

AGE	GENDER	ZIP CODE
24:1,25:2,26:1,27:1,28:1,29:1,30:1,32:1,33:1	male:3,female:9	581343:4,575002:2,575003:4,575015:1,587620:1
24:1,25:2,26:1,27:1,28:1,29:1,30:1,32:1,33:1	male:3,female:9	581343:4,575002:2,575003:4,575015:1,587620:1
24:1,25:2,26:1,27:1,28:1,29:1,30:1,32:1,33:1	male:3,female:9	581343:4,575002:2,575003:4,575015:1,587620:1
24:1,25:2,26:1,27:1,28:1,29:1,30:1,32:1,33:1	male:3,female:9	581343:4,575002:2,575003:4,575015:1,587620:1
24:1,25:2,26:1,27:1,28:1,29:1,30:1,32:1,33:1	male:3,female:9	581343:4,575002:2,575003:4,575015:1,587620:1
24:1,25:2,26:1,27:1,28:1,29:1,30:1,32:1,33:1	male:3,female:9	581343:4,575002:2,575003:4,575015:1,587620:1
24:1,25:2,26:1,27:1,28:1,29:1,30:1,32:1,33:1	male:3,female:9	581343:4,575002:2,575003:4,575015:1,587620:1
24:1,25:2,26:1,27:1,28:1,29:1,30:1,32:1,33:1	male:3,female:9	581343:4,575002:2,575003:4,575015:1,587620:1
24:1,25:2,26:1,27:1,28:1,29:1,30:1,32:1,33:1	male:3,female:9	581343:4,575002:2,575003:4,575015:1,587620:1

G)Slicing

Shows the Slicing table which is being displayed after the partioning the data horizontally and vertically where only required details of patients are being displayed while hiding the individual identity and attributes which are highly co-related are grouped together.

Slicing Table

AGE,GENDER	ZIP CODE,DISEASE
(24,female)	(581343,cold)
(25,female)	(575002,lumps in breast)
(25,female)	(575003,Fever)
(26,female)	(575003,cancer)
(27,female)	(575015,blood-cancer)
(28,female)	(575002,Fever)
(29,female)	(581343,StomachPain)
(30,male)	(575003,ulcers in mouth)
(32,male)	(587620,ankle swollen)
(33,female)	(581343,StomachPain)
(36,female)	(581343,ankle swollen)
(39,male)	(575003,knee swelling)
(40,female)	(575015,skin allergy)

H) Overlapping Slicing

In overlapping slicing the attributes are duplicated in both columns so the privacy of the patients is preserved in better manner and if adversary trying to gain any personal information about any individual will lead to dilemma by knowing which is the original value and duplicate value.

Overlapping Slicing Table

AGE,GENDER	AGE,ZIP CODE,DISEASE
(24,female)	(24,581343,cold)
(25,female)	(25,575002,lumps in breast)
(25,female)	(25,575003,Fever)
(26,female)	(26,575003,cancer)
(27,female)	(27,575015,blood-cancer)
(28,female)	(28,575002,Fever)
(29,female)	(29,581343,StomachPain)
(30,male)	(30,575003,ulcers in mouth)
(32,male)	(32,587620,ankle swollen)
(33,female)	(33,581343,StomachPain)
(36,female)	(36,581343,ankle swollen)
(39,male)	(39,575003,knee swelling)
(40,female)	(40,575015,skin allergy)

V. CONCLUSION

The disadvantages of generalization and bucketization are overcome by using slicing and overlapping slicing techniques. Overlapping slicing has a capacity to hold the huge quantity of data. Using overlapping slicing, the size of the data has been reduced, as the partioning takes place in column, also it maintains the data usefulness while not revealing the identity of an individual Overlapping slicing is where the duplication of attributes takes place in more than in one column which will help to keep the medical data sets preserved and to create better co-relation among the attributes. Main advantage of overlapping slicing is that, it can be used without any separation between sensitive attribute and quasi attribute.

VI. REFERENCES

[1]. Tiancheng Li, Ninghui Li and Jian Zhang,Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3, 2012.

[2]. Amar Paul Singh and Ms. Dhanshri Parihar "A Review of Privacy Preserving Data Publishing Technique".International Journal of Emerging Research in Management &Technology Volume-2, Issue-6, ISSN: 2278-9359 .

[3]. Zankhana Prajapatia and Rushirajsinh Zalab "Privacy Preserving For High-Dimensional Data using Overlapping Slicing" International Journal of Innovative and Emerging Research in Engineering Volume 2, Issue 4, 2015.

[4]. Jeevanandhini M and Ruby gnanaselvam C "Slicing techniques: A new approach to privacy preserving data publishing" International Journal of Multidisciplinary Research and Development,Volume 3,Issue 6, 2016 ISSN: 2349-4182.

[5]. S.Giri and Mr.Nilav Mukhopadhyay "Overlapping Slicing With A New Privacy Model" International Journal of Scientific and Research Publications, Volume 4, Issue 6,2014, ISSN 2250-3153.