The Study of Slicing A New Approach for Privacy Preserving Data Publishing

Prof. Ghule Shital
Computer Department
SCSCOE, Rahuri.
Ahmednagar, India
shitalghule@scscoe.org

Mr. Jadhav Saurabh
Computer Department
SCSCOE, Rahuri.
Ahmednagar, India
saurabhjadhav@scscoe.org

Miss. Landage Sayali
Computer Department
SCSCOE, Rahuri.
Ahmednagar, India
sayalilandage@scscoe.org

Mr. Anwat Sandip
Computer Department
SCSCOE, Rahuri.
Ahmednagar, India
sandipanwat@scscoe.org

Mr. Arote Parag
Computer Department
SCSCOE, Rahuri.
Ahmednagar, India
paragarote@scscoe.org

***Abstract*—Privacy is an important issue when any one wants to use data that contains individuals sensitive information. To protect individuals sensitive information generalization and bucketization techniques are used. But generalization and bucketization losses considerable amount of data from high dimensional data whereas in bucketization technique does not provide clear separation between quasi identifier attribute and sensitive attribute also it does not prevent membership disclosure protection because bucketization produce quasi identifier as per original dataset. Generalization use k-anonymity and bucketization use diversity check algorithm. Data publishing is not big task but preserving privacy is important issue now days. We introduce a new technique slicing. The basic idea of slicing is to overcome drawbacks of generalization and bucketizationi.e it divide the data both horizontally and vertically. Vertical partitioning is done by grouping attribute into columns and each column contains subset of attributes that are highly related. Horizontal partitioning is done by grouping tuple into buckets. It can efficiently used with high dimensional.**

***Index terms*- l-diversity, k-anonymity, slicing, bucketization, generalization.**

I. INTRODUCTION

Privacy-preserving is an important task for publishing of microdata. Microdata is contains record each information having entity such as a person, a household, or an organization. There are some better anonymization techniques used such as bucketization and generalization that are designed for privacy preserving microdata publishing. In slicing generalization and bucketization are used.

In both approaches, attributes are partitioned into three categories such as identifiers, quasi identifiers and sensitive attribute. In both techniques firstly identifiers are removed then tuples are partitioned into buckets. others step are different in both techniques. In generalization transforms the QI-values in each bucket into less specific but semantically consistent values so that tuples in the same bucket cannot be distinguished by their QI values. In bucketization, one separates the SAs from the QIs by randomly permuting the SA values in each bucket.Our experiments shows that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute also demonstrate that slicing can be used to prevent membership disclosure.

Slicing has several advantages when compared with generalization and bucketization. Slicing provides better data utility than generalization. It preserves more attribute correlations with the SAs than bucketization also handle high-dimensional data and data without a clear separation of QIs and SAs. We an efficient algorithm for computing the sliced table that contain diversity. Our algorithm partitions attributes into columns, applies column generalization, and partitions tuples into buckets. Attributes that are highly-correlated are in the same

column; this preserves the correlations between such attributes.

II. SLICING

Slicing is mostly use for privacy preserving data publishing. It has several advantages compared to previous technologies such as generalization and bucketization. Slicing can also handle high dimensional data and data without clear separation between quasi - identifier and sensitive attribute. We use a technique l-diverse for sliced table. also the column generalization technique are used.

We introduce the module in this system are original data, generalization, bucketization and sliced data with all this module system work with privacy preservation. Following table shows an example of microdata which contain the original table, generalization table, bucketization table.

Figure(1) shows the original table which having attribute age, sex, zip-code this are the QI attribute and the sensitive attribute is disease. Figure(2) shows the generalized table which satisfies 4-anonymity. Figure(3) shows bucketization table which satisfies 2-anonymity and done the horizontal partitioning. Figure(4) shows the one attribute per column slicing table. Slicing partition tuples into bucket also the values within each bucket are randomly permuted to break the linking between different column. figure(5) shows the slicing table.

All this table shows the implementation of the system. For slicing the original data can be taken as input to preserve privacy. From this approach we preserve better utilization than generalization. Also it support for high diamensional data. For example it use hospital data, sensus record also the big databases of organizations can use this system to preserve privacy.

For example if we take the database of hospital record in this database all three attributes that is quasi identifier ,sensitive attribute and identifiers are used. On this database the all of the methods like generalization, bucketization, slicing operation are performed by ay applying the algorithms which are used in slicing.

Age	Sex	Zip-code	Disease
24	M	422605	Flu
24	F	422605	Flu
35	F	422604	Gastric
54	M	422704	Dyspepsia
60	M	422702	Flu

Table (1)-original table

Age	Sex	Zip-code	Disease
[20-52]	*	42260*	Flu
[20-52]	*	42260*	Flu
[20-52]	*	42260*	Gastric
[54-64]	*	42270*	Dyspepsia
[54-64]	*	42270*	Flu

Table(2):generalize table

Age	Sex	Zip-code	Disease
24	M	422605	Gastric
24	F	422605	Flu
35	F	422604	Flu
54	F	422704	Dyspepsia
60	M	422702	Flu

Table(3):Bucketize table

Age	Sex	Zip-code	Disease
24	M	422702	Gastric
24	F	422704	Flu
35	F	422704	Flu
54	F	422605	Dyspepsia
60	M	422605	Flu

Table(4):One attribute-per column slicing

Age	Sex	Zip-code	Disease
24	M	422702	Gastric
24	F	422704	Flu
35	F	422704	Flu
54	F	422605	Dyspepsia
60	M	422605	Flu

Table(5):slice table

III. SYSTEM ARCHITECTURE

Main focus of proposed architecture is to achieve privacy preservation also simultaneously utilize data in better way. The system architecture is as follows:

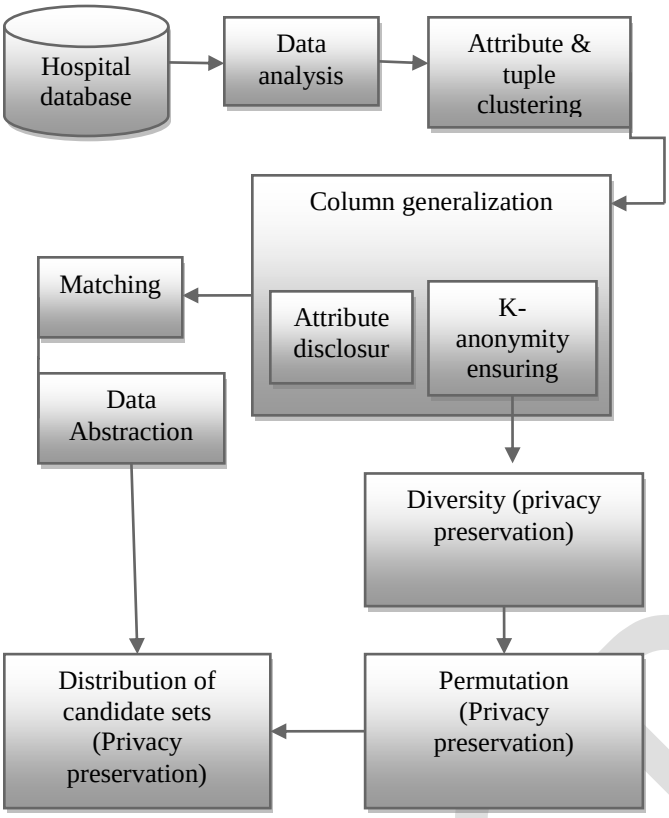

Figure.1: System Architecture.

IV. SYSTEM WORKING

A. Data Analysis Module

Micro data is taken from large database usually for research purpose so one of micro data which is related to hospital is taken here and data analysis is performed for it. Based on data analysis it is broadly classified into three attributes at this point vertical partitioning is done. Once attribute are formed, tuples are clustered based on quasi identifier as identifier remain unique and cannot be used as clustering element. clustering cannot be performed based on sensitive attribute as they are highly sensitive at this point horizontal partitioning is achieved. Columns are generalized in two views 1.admin 2.Client/User. In admin view all the attribute are disclosed representing the original data sets and thus achieving privacy preservation.

1. CLUSTERING MODULE

In this clustering module the data sets are horizontally partitioned by grouping the tuples into buckets. This grouping is based on k-means cluster. The quasi-identifiers are chosen for clustering. The clustering is basically done to group related data sets together. The k-means clustering is chosen for following reasons.

2. MATCHING AND DIVERSITY MODULE

This module serves admin purpose when he/she may want to match the client viewwith admin view. One may use this module to ensure the data displayed on client sideis privacy protected. For instance, this matches and checks for one attribute relatedwith multiple attribute. Distributing data with duplicate values is diversity whichis achieved by adding fake tuples to the original data, there by achieving privacypreservation.

3. RANDOM PERMUTATION MODULE

Random permutation module, within each bucket values in each column are randomly permutated(or sorted) to break the linking between different columns. This ensures the maximum level of privacy preservation.

4. DATA DISTRIBUTION MODULE

In this module the data sets is distributed depending on the login made. in adminview the original table along with the indexed values of the client table is displayed.in user view the sliced table will be presented as a result which is fully privacyprotected.

V. ALGORITHM

To develop this system we used algorithm for implementation. we use the l-diversity algorithm and column generalization algorithm. This algorithm is as follows:

Tuple –partition algorithm

1 Q=S; SBB=ϕ

2 while Q is not empty

3 remove the first bucket from Q;

Q=Q-bucket

4 split bucket into 2 buckets bucket1 and bucket2.

5 if diversity-check(S, Q ∪ bucket 1, bucket2∪SBB, l)

6 Q=Q∪bucket1, bucket2.

7 else SBB=SBB ∪ bucket.

8. return bucket

Diversity check algorithm

1. For each tuple l ε T, L[t]=ϕ
2. For each bucket in T

3. Record f(v) for each column value v in bucket.
4. For each tuple tε T.
5. Calculate p(t,B) and find D(t,B)
6. L[t]=L[t] ʊ hp(t,B),D(t,B)i.
7. For each tuple t ε T.
8. Calculate (t,S) for each s based on L[t].
9. If p(t,s)>1/l,return false.
10. Return true.

VI. EXPERIMENTAL RESULTS

In this system, we conduct the experiments. We found 2 experiments. In first experiments we evaluate that slicing preserve its effectiveness in data utility. Also protect against attribute disclosure as compared to bucketization and generalization. We use for that purpose l-diversity algorithm & use 3 anonymization techniques such as generalization, bucketization and slicing.

In second experiment, it contain the effectiveness of slicing in membership disclosure protection. For that we count fake tuples also compare how many bucket are matched to original data.

A. Experimental Data: We use the Adult dataset from the UC Irvine machine learning repository, which is comprised of data collected from the US census. Tuples with missing values are eliminated and there are 45222 valid tuples in total. The adult dataset contains 15 attributes in total. In our experiments, we obtain two datasets from the Adult dataset. The first dataset is the "OCC-7" dataset, which includes 7 attributes: QI = Age, Workclass, Education, Marital-Status, Race, Sex and S = Occupation. The second dataset is the "OCC-15" dataset, which includes all 15 attributes and the sensitive attribute is S = Occupation. In the "OCC-7" dataset, the attribute that has the closest correlation with the sensitive attribute Occupation is Gender, with the next closest attribute being Education. In the "OCC-15" dataset,
the closest attribute is also Gender but the next closest attribute is Salary.

*B. Expected Output***:** We compare slicing with generalization and bucketization on data utility of the anonymized data for classifier learning. For all three techniques, we employ the Mondrian algorithm [5] to compute the `-diverse tables. The ` value can take values 5,8,10 (note that the Occupation attribute has 14 distinct values). In this experiment, we choose ` = 2. Therefore, the sensitive column is always Gender, Occupation. Classifier learning. We evaluate the quality of the anonymized data for classifier learning, which has been used in (11, 18, 4). We use the Weka software package to evaluate the classification accuracy for Decision Tree C4.5 (J48) and Naive Bayes. Default settings are used in both tasks. For all classification experiments, we use 10-fold cross-validation. In our experiments, we choose one attribute as the target attribute (the attribute on which the classifier is built) and all other attributes serve as the predictor attributes. We consider the performances of the anonymization algorithms in both learning the sensitive attribute Occupation and learning a QI attribute Education. Learning the sensitive attribute. In this experiment, we build a classifier on the sensitive attribute, which is "Occupation"[12]. We _x c = 2 here and evaluate the effects[14] .

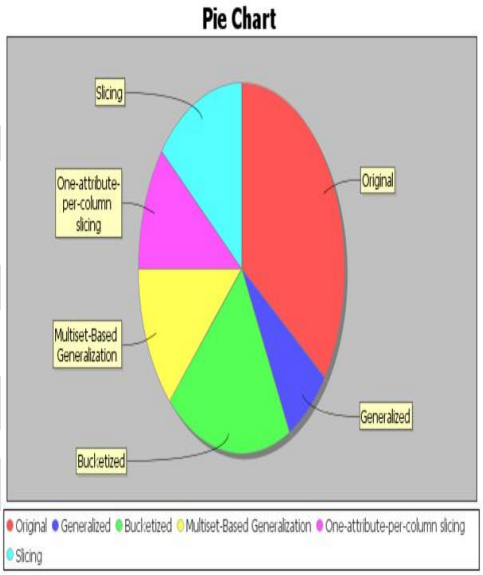

Figure.2: Expected Output

Analysis - In all experiments, slicing outperforms both generalization and bucketization that confirms that slicing preserves attribute correlations between the sensitive attribute 63 and some QIs (recall that the sensitive column is Gender, Occupation). Another observation is that bucketization performs even slightly worse than generalization.

VI. CONCLUSION

We formalize the slicing technique and compare it with generalization and bucketization. We define `l-diverse slicing for attribute disclosure protection and develop an efficient algorithm to achieve `l-diverse slicing. We explain how slicing prevents membership disclosure.

REFERENCES

[1] Tiancheng Li, NinghuiLi,JiaZhang,and Ian Molloy, \Slicing: A New Approachfor Privacy Preserving Data Publishing", Proc. ieee transactions on knowledgeand data engineering, Vol.24,No.3,March 2012

[2] Fuad Ali Mohammed, Al-Yarimi, SonajhariaMinz\Data Privacy in Data Engineering, the Privacy Preserving Models and Techniques in Data Mining andData Publishing: Contemporary A_rmation of the Recent Literature.",Volume 6- No.3, December 2012.

[3] Yogendra Kumar Jain, Vinod Kumar Yadav, Geetika S. Panday ,\An EffcientAssociation Rule Hiding Algorithm for Privacy Preserving Data Mining", IJCSE,Vol.4,No.18 ,2011

[4] AlinaCampan, Traian Marius Truta, Nicholas Cooper, \P-Sensitive K-Anonymity with Generalization Constraints",transactions on data privacy,Vol.65,No.3,2010

[5] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre and AshwinMachanavajjhala,\Privacy-Preserving Data Publishing",Foundationsand Trends in Databases Vol.2,No.12,2009.

[6] A. Inan, M. Kantarcioglu, and E. Bertino, In ICDE, \Using anonymizeddatafor classify"Vol.28,No.30,2009.

[7] Yeye He, Je_rey F. Naughton,\Anonymization of SetValued Data via TopDown,Local Generaliztionl.5,No.7,2009.

[8] T. Li and N. Li; In KDD, \On the tradeo_ between privacy and utility in datapublishing",Vol.78,No.86,2009.

68

[9] T. Li and N. Li. Injector: In ICDE, \Mining backgroundknowledge for data anonymization."Vol.67,No.12,2008.