Manuscript
Click here to download Manuscript
BinByMeanSlicing_IJSEKE.pdf

BINBYMEAN SLICING: AN EFFICIENT PRIVACY PRESERVING DATA PUBLISHING

Nithya.M

Research Scholar, Faculty of CSE, Sathyabama University,
Assistant Professor CSE, Sri Sairam Eng. College, Chennai, India
Nithya.cse@sairam.edu.in

Dr. T.Sheela

Professor and HOD, Department of IT,
Sri Sairam Engineering College, Chennai, India
Hod.it@sairan.edu.in

Data publishing is a critical step in data mining. Effective management of data during this phase guarantees privacy and accuracy of data. In general Privacy and accuracy are always trade off factors and is hard to retain both. Level of privacy breaching is measured as Privacy Loss and level of inability to interpret data accuracy is measured as Accuracy or utility loss. Both these losses need to be retained low for an efficient data handling system. Generalization, Bucketization and Slicing are well known techniques among the list. Unfortunately they have their own limitation in handling privacy and accuracy. Generalization suffers in handling high dimensional data thus experiencing higher utility loss. Bucketization lacks data privacy where differentiating sensitive and quasi identifier attributes is a challenge. Slicing on the other hand though offers better privacy and accuracy, there is always scope to improve data correlation aiming in reducing utility loss. This paper explains a new technique called BinByMean Slicing which is designed exclusively for Medical data handling. Privacy and accuracy losses are calculated for Generalization, Bucketization, Slicing and BinByMean Slicing using Kullback-Leibler divergence to demonstrate the level of accuracy and privacy losses. Experimental result reveals this method guarantees reduced privacy loss and accuracy loss.

Keywords: Privacy; Accuracy; Slicing; Publishing; Kullback-Leibler.

1. Introduction

Handling medical data is crucial as it deals with sensitive individual information. In data publishing there is a strong assumption that privacy and accuracy are trade off features [1], practically impossible to achieve both. Guaranteeing accuracy may help the society by offering factual medical analysis but on the other hand it is realized at the cost of breaching individual sensitive data. Medical data publishing [2] results in privacy loss when it goes vulnerable for intruders to learn regarding sensitive information. At the same time it results in accuracy gain when the same information is learnt by researchers. Direct- comparison methodology [3] discusses privacy and accuracy loss in detail. Here privacy loss is averaged among individuals and as a result it offers relatively lesser importance to individual privacy loss. Loss of privacy on an individual is always considered critical as it finally contributes to privacy loss level of the system. Thus worst case privacy loss has to be considered for all individuals. Having explained on the privacy and accuracy aspect, it is not advised to compare privacy with accuracy as it

involves judgmental factors. To understand this phenomenon we can relate it to level of risk associated while investing and the corresponding returns we get post investment. In any case the risk and returns cannot be compared to each other. The same way privacy loss and accuracy gain cannot be related. Based on the method selected either privacy or accuracy is guaranteed leaving the other delimited [4].

Anonymization techniques [5], [6] like Generalization [7]; Bucketization [8], [9] and Slicing are well known which handles data anonymization in their own way. In general these techniques manage in manipulating the original data to avoid sensitive information made available for data analysts. During data manipulation, there are always possibilities of data utilization going down. Utilization loss becoming predominant shall directly affect the accuracy of data analysis. In few occasions the analysis results go completely wrong finally unable to solve the very purpose of data mining and publishing. Further there is always confusion in choosing the right privacy requirement from the listed types like k-anonymity or l-diversity or t-closeness. Next comes selecting the right parameter for the selected privacy requirement. For example, it is required to know if to choose l=2 or l=5 for l-diversity type. This paper explains about a new algorithm called BinByMean Slicing which is the successor of Slicing technique. Privacy and accuracy losses are validated using Kullback-Leibler divergence algorithm. Initial part of this paper will detail on merits and demerits of Generalization, Bucketization and Slicing techniques [10]. Further the paper would compare the privacy and accuracy losses for all the methods.

2. Data Analysis

Any source data shall have identifiers which can uniquely identify individuals (Name, SSO), Quasi Identifiers (Age, Sex) which are available for the analyst and sensitive data (Disease, Salary) whose privacy need to be secured. Medical records from Hospital, salary records from a Company are considered as sensitive data. These data when leaked out could be threat to individual privacy. Similarly military records from Government when leaked out could be threat to whole nation. These data could be of any data type, any volume and size. The data source can be manipulated with certain level of privacy maintained and released to certain group of people. In parallel another group of people might receive manipulated data with different degree of privacy. If both the data variants are somehow accessible by an intruder then there is always a possibility to compare both the data variants and exploit the privacy factor. Further data analysis results should always respect analysis requirement. In few occasions maintaining data privacy is expected than accuracy of data. Advanced privacy preserving algorithms like differential privacy [11] can be adopted to ensure privacy. In other cases accuracy of data is mandate when compared to privacy. For example through analysis it is found that in a particular city, majority of the female population are affected by a severe disease do to consumption of particular milk brand. When this data is published accurately, we have the possibility to find the root cause and eradicate it. On the other hand privacy of individual is leaked out. Here individual privacy loss is relatively having lesser weightage compared to collective accuracy of data, which is helping the society to find solution for a common problem. Also it is evident that privacy loss is a measure of individual data and accuracy

loss is a measure of aggregated data. Thus data publishing techniques should have the flexible of generating reports according to need [12].

3. Inspiration for BinByMean Slicing

To realize a better approach, it is necessary to understand the drawbacks of existing methods. A sensitive data table is evaluated against known techniques like Generalization, Bucketization and Slicing. Finally BinByMeanSlicing technique is evaluated as well to measure the privacy and accuracy loss. Since all these techniques follow anonymization methodology, the entire data table is grouped to smaller groups called bucket with each row in the bucket referred as tuple. These tuples are grouped to form equivalence class, where each equivalence class should have atleast k-unique sensitive data for realizing K-anonymity privacy metrics. Similarly L-diversity metrics looks for L-distinct sensitive data per equivalence class and T-closeness metrics looks for similar distribution of sensitive data within the equivalence class and source data. It is recommended to have the size of equivalence class with more tuples, as it would promote data closeness resembling similar distribution as that of the source data. While in this process there is always a tradeoff as increasing the size of equivalence class would result in duplication of the sensitive attributes. This would directly affect the K-anonymity metrics, which expects k unique sensitive attributes per equivalence class. As a result the maximum number of tuples in an equivalence class is limited to the number of distinct sensitive attributes present in the source data. To achieve balance state on K-anonymity and T-closeness metrics, each equivalence class can have the tuples added in the same proposition of sensitive attributes present in the source data and in parallel maintaining distinct tuple addition for an equivalence class. In general it is difficult to satisfy all the metrics for a methodology. Inspiration of BinByMeanSlicing is based on creating a method which can satisfy K-anonymity, L-diversity and T-closeness metrics thereby providing low privacy and accuracy loss.

Table 1. Description of Medical records

Attributes	No. of values
Location	5
Gender	2
Age	32
Disease	5

Table 1. shows attributes and number of occurrence of medical records which are sensitive in nature and considered for publishing. Unique identifier data “Name” is removed from source table for further processing. “Location”, “Gender” and “Age” are considered as quasi identifiers and “Disease” as sensitive attribute. Combination of the quasi identifiers in the table could easily reveal the sensitive information.

3.1. *Generalization*

Various generalization and suppression methods are discussed in [13]. Generalization based on k-anonymity has higher data utility loss for high dimensional data [14], [15], [16]. This is due to the fact that data generalization in a bucket requires data closeness. If the tuples distances are far apart then generalization could be a challenge. In general data closeness cannot be expected for high dimensional data. Further uniform distribution assumption is required for tuples falling in a bucket which further impacts data utility. Adding to the above utility issues, generalization is done separately for each attribute thus impacting data correlation between attribute columns. K-Anonymity aims in securing privacy of data. K-Anonymity expects each record is indistinguishable from at least k-1 other records in an equivalence class formed by K records.

Table 2. Generalized Data

Location	Gender	Age	Disease
6000**	*	[28-36]	Malaria
6000**	*	[28-36]	Malaria
6000**	*	[28-36]	Cancer
6000**	*	[28-36]	Diabetics
6000**	*	[28-36]	Typhoid

Table 2. shows one of the 8 buckets with location, gender and age attributes generalized. Quasi identifiers are replaced by less specific, but semantically consistent values. The location attribute is generalized with last 2 digits, gender attribute completely generalized and the age attribute grouped to achieve k-anonymity factor = 4 on an equivalence class of 5 tuples for the disease attribute. Due to lack of sensitive attribute diversity, k-Anonymity is prone to homogeneity attacks. Though k-anonymity prevents from identity disclosure, it fails to protect against attribute disclosure. Also this method fails when the intruder has back ground knowledge leading to background knowledge attack. Accuracy or data utility is directly related to number of generalization steps, average size of equivalence class and the discernibility metric (DM) which sums up the squares of equivalence class sizes.In the example considered, there are 3 generalization steps, with 5 as average size of equivalence class.

3.2. *Bucketization*

Bucketization technique has better data utility when compared to Generalization but has serious privacy concerns. Membership disclosure cannot be prevented [17] in this technique as all the quasi identifiers are published in their natural state. This can create opportunity for intruder to decode the identifier information based on the quasi identifier relation. In general when certain quasi identifier combinations are undistorted, then it can lead to information loss [18]. Another drawback is this technique requires perfect demarcation between quasi identifiers and sensitive attributes. In most cases there is always confusion in identifying quasi identifiers vs. sensitive attributes. Thus this

concern adds to Bucketization drawback. Finally attribute correlation gets affected as this technique needs separation of quasi identifier and sensitive attribute. Bucketization is correlated to L-Diversity methodology. Each equivalence class is expected to have at least L – well represented sensitive attributes as per L-diversity model. This model infact is a solution for K-Anonymity issue as it demands in an equivalence class, each sensitive value can occur at a frequency of at most 1/L. Though L diversity prevents from homogeneity attack, it fails when dealing with background knowledge attack.

Table 3. Bucketized Data

Location	Gender	Age	Disease
600053	M	37	Typhoid
600052	F	39	Diabetics
600054	M	40	Diabetics
600052	M	41	Flu
600053	M	43	Flu

In Table 3. quasi identifiers are bucketized with 5 tuples each resulting in 8 buckets. Location, gender and age attributes are kept as it is per bucket and the disease attribute is random shuffled to realize anonymity. While doing this anonymization few buckets result in only 3 diversity, where the expected L-diversity metrics is 5.Table.3 shows one of the bucket with worst case diversity as few sensitive values (Diabetics, Flu) repeats leading to attribute disclosure.

3.3. *Slicing*

Slicing has an upper hand with respect to Generalization and Bucketization which can manage high dimensional data. In this technique the source data table is handled both vertically [19] and horizontally [20]. In this technique the source data table is divided column wise. This division brings certain quasi identifiers together on one side (vertical X) and the other with a combination of quasi identifier and sensitive attribute (vertical Y). Further the data table is bucketed and random shuffling is executed on the second part (vertical Y) of bucketed tuples with respect to sensitive attribute. Though (vertical Y) is random shuffled as a combination of age and disease attributes, it reflects considerable privacy loss as there is good correlation existing between the location and gender attribute in vertical X component with respect to age and disease attribute in vertical Y component. The correlation between X and Y components are preserved leading to less accuracy loss. This creates an opportunity for realizing an efficient technique which can fine tune privacy and accuracy loss to a minimum level. Care should be taken when there are repeating sensitive attributes with ineffective shuffling resulting in greater probability of membership disclosure.

Table 4. Sliced Data

Location	Gender	Age	Disease
600053	M	34	Diabetics
600053	F	32	Malaria
600043	M	28	Malaria
600056	F	34	Cancer
600053	M	36	Typhoid

Table 4. shows one of 8 buckets having correlated attributes (location, gender) as X vertical and (age, disease) as Y vertical. Horizontal partition is realized by bucketizing the tuples. Further Y vertical is randomly shuffled such that the associations between uncorrelated groups (X and Y) are broken. This secures better privacy when compared to Generalization and Bucketization. On the other hand accuracy is also sustained since the associations between correlated attributes are intact. Compared to accuracy loss, privacy loss is relatively high in this technique. There is still opportunity to reduce the privacy loss without compromising data accuracy. To further reduce privacy loss BinByMean Slicing algorithm is proposed.

4. BinByMean Slicing Technique

Considering the drawbacks of Generalization, Bucketization and Slicing there is need to create an improvement in publishing technique. BinByMean Slicing is designed to overcome the above drawbacks. This version is a hybrid slicing which aims in relatively reducing the privacy loss without compromising the data accuracy.

4.1. *Algorithm*

BinByMean Slicing

A. Horizontal partitioning (Bucketization) with L-diversity:

1: N: Dataset, n (N) =Number of tuples in N
2: N has 5 attributes: N= $\{N_V,N_W,N_X,N_Y,N_Z\}$ where N1= $\{N_W,N_X,N_Y,N_Z\}$
3. Sorting the dataset based on N_X
4: Partitioning data set N1 into buckets based on $(B_i\text{-}B)^2$ distinct attribute values in N_Z where $N_Z = N_{Z1}\ldots..N_{ZN.}$
5: N1 $=B_1UB_2UB_3\ldots\ldots B_M$ where M=No.of buckets with L- diversity of N_{ZN} in N_Z for every B Equivalence class
6: Increase size of B until N_Z has all the distinct values in it.

B. Finding attribute correlation:

7: Calculate correlation coefficient between the column attribute using Pearson constant. If r=> 0.7 then group the column attribute.

C. Vertical Partitioning:

8: Based on the r value vertically partition the data set
Vert $X=B_WB_{Y,}$ Vert $Y=B_XB_Z$: N1=$B_WB_Y \cup B_XB_Z$, $B_WB_Y \cap B_XB_Z=\emptyset$.
9: $B_WB_Y= \cup B_{W\,i}B_{Yi}$, $B_XB_Z=\cup B_{Xi}B_{Zi}$ where i =1 to M
10: $B_WB_Y= B_{W\,1}B_{Y1} \cup B_{W\,2}B_{Y2}\cup B_{W\,3}B_{Y3}\ldots\ldots \cup B_{W\,M}B_{YM}$
11: $B_XB_Z=B_{X1}B_{Z1} \cup B_{X2}B_{Z2} \cup B_{X3}B_{Z3}\ldots\ldots\ldots\ldots \cup B_{XM}B_{ZM}$
W, X, Y =quasi identifier (Location, Age, Gender) Z=Sensitive attribute (Disease)

D. Random shuffling :Fisher–Yates:

12: For each $B_X B_Z$ perform random shuffling:
13: Vertical $Y=B_X B_Z$. which has N_{ZN} elements (indices $Y_0 \ldots Y_{n-1}$):
14: for i from n − 1 down to 1 do
15: j ← random integer such that $0 \le j \le i$
16: exchange Y[j] and Y[i]

E. Apply BinByMean value on N_X:

17: For each B,
18: $Nx = \sum_i^m Xi / m$, where m = No. of tuples in the bucket.

4.2. *Working procedure*

Source data table which has to be published is first classified into identifiers, quasi identifiers and sensitive attributes. The source data is sorted based on N_X. Horizontal partition [21] is done by grouping all distinct sensitive attributes and associated tuples into buckets. Each bucket forms into equivalence class with distinct sensitive attributes. When the size of the bucket is increased it will result in better data closeness resembling similar distribution as that of the source data. On the other hand building the bucket with more tuples will result in duplication of sensitive attributes hampering the L-diversity metrics [22]. Semantic extraction techniques [23] on L-diversity can be used to increase the size of the bucket by adding more diversity. Ideally the number of tuples in a bucket is restricted to number of distinct sensitive attributes. Next each bucket is sliced into columns. This division brings certain quasi identifiers together on one side (vertical X) and the other with a combination of quasi identifier and sensitive attribute (vertical Y). This column wise segregation is done for associating highly correlated attributes together to realize accuracy and at the same time disintegrating uncorrelated attributes to retain privacy. Further to breakdown the correlation between vertical X and Y group, random shuffling of vertical Y group is done with respect to sensitive attribute. This step will improve the privacy of data relative to accuracy of data.

Table 5. BinByMean Sliced Data

Location	Gender	Age	Disease
600053	F	11	Malaria
600043	M	11	Cancer
600053	F	11	Diabetics
600053	F	11	Typhoid
600052	M	11	Flu
600056	M	24	Cancer
600052	M	24	Typhoid
600054	F	24	Malaria
600053	F	24	Diabetics
600054	F	24	Flu

In every bucket, N_X age attribute is normalized by averaging the age value and replacing all age values with average value. This brings data closeness within the bucket resulting in improved data accuracy. Data closeness will help in relating the bucket data to the source data improving T-closeness metrics. Finally the bucketized vertical X and resorted vertical Y components are combined. All the buckets are stacked back to construct the sliced data. Privacy and accuracy loss are calculated for source and sliced data set. BinByMean Slicing offers better privacy while retaining data accuracy to acceptable limit. This can be realized with experimental results. Table.5 shows first two bucketized data managed by BinByMean Sliced algorithm. To realize horizontal partitioning the tuples need to be grouped into buckets. L-diversity matrices is used to identify number of tuples per bucket and thereby number of buckets B_M to cover all the records. The algorithm checks every attribute (N_W, N_X, N_Y, N_Z) in the tuple to identify the sensitive attribute N_Z. The algorithm runs through the complete sensitive attribute column to count the number of distinct attributes ($N_Z = N_{Z1}.....N_{ZN}$). First bucket B1 is formed by grouping tuples which has distinct sensitive attributes. B2 to B8 is formed in similar way to fill all the tuples. The sample record has 5 distinct sensitive attributes and thus the size of each bucket is 5 resulting in 8 buckets. Attribute correlation is used to group vertical slicing to identify column attributes with $r > 0.7$. Attribute analysis [24] is done in order to identify correlations.

Table 6. Vertical correlation

	Location	**Gender**	**Age**	**Disease**
Location	1			
Gender	**0.086742**	1		
Age	-0.04169	-0.02101	1	
Disease	0.035943	-0.05152	**0.092742**	1

Table 6. shows correlation between all quasi identifiers. When r is close to 1, then the respective columns are highly correlated and if r = 0, then the respective columns lack correlation. r = 0.086742 for attributes location and gender and similarly r = 0.092742 for attributes age and disease. These two groups qualify for vertical partition. On the other hand r = -0.04169 for attributes (age, location), r = 0.03594 for attributes (location, disease), r = -0.02101 for attributes (gender, age) and r = -0.05152 for attributes (gender, disease) resulting in adverse groups. Each bucket B_M is vertically grouped into vertical X (location, gender) and vertical Y (age, disease). Vertical Y is random shuffled with respect to sensitive attribute disease, using Fisher–Yates shuffle algorithm. To shuffle a combinational array: Vertical $Y=B_XB_Z$ which has N_{ZN} sensitive elements with indices $Y_0 \ldots Y_{n-1}$, start a routine from:

for i from n − 1 down to 1 do
    j ← random integer such that $0 \leq j \leq i$.
    exchange Y[j] and Y[i].

In spite of the performing Fisher–Yates shuffle algorithm, privacy loss will be still significant, since there is a possibility of few random shuffled vertical Y tuples occupying the same position as it was before shuffling. This will directly impact the

privacy information, when the vertical Y and vertical X groups are merged and analysed.To avoid this issue, further anonymization can be applied to the random shuffled vertical Y group. BinByMean Slicing algorithm is applied to the N_X attribute of every bucket. N_X attribute is anonymized by averaging the age attribute in a bucket. The average value will replace the individual age value in each tuple in the bucket. Proposed technique which falls under micro aggregation method [25] will still hold the accuracy loss very low, because 100% anonymization is not applied but the age attribute is grouped to its mean value bringing closed association between them. Thus privacy loss decreases further compared to conventional slicing method, still keeping accuracy intact.

5. Comparative Analysis

This section compares the proposed BinByMeanSlicing with the conventional methods. Privacy and accuracy parameters are used to rate the quality of methods. Direct comparison methodology of privacy and accuracy may not be a right solution as there is no common measure for both the parameters. Privacy and accuracy cannot be directly proportional as the intruder and publisher may not have same level of knowledge on the data and may not use the same learning techniques. Privacy is measured based on the level of sensitive attribute learning acquired by the intruder from the rest of quasi identifiers. Accuracy is measured based on the level of correlation between the sensitive and other quasi identifiers. Different methods like variance distance and earth movers distance [26], [27] are used to measure the distribution of data in the source and anonymized data sets.

5.1. *Privacy Loss*

Privacy loss is measured as worst case data loss on an individual in a given set of data. False or true information regarding an individual may also cause privacy loss. Kullback-Leibler divergence method is used to measure privacy loss. This method measures the difference between two probability distributions P(t) and Q, where Q denotes distribution of sensitive attributes in the sample source data and P (t) denotes distribution of sensitive attributes in the equivalence class. In general Kullback Leibler distance (KL-distance) is a natural distance function from a "true" probability distribution to a "target" probability distribution. It can be interpreted as the expected extra message-length per datum due to using a code based on the wrong (target) distribution compared to using a code based on the true distribution.

For discrete probability distributions, P = {P1... Pn} and Q = {Q1... Qn}, the KL-distance is defined to be

$$Ploss(t) = KL(P, Q) = \Sigma i\ Pi.\ \log2\ (Pi / Qi)$$

For continuous probability densities, the sum is replaced by an integral.

$$KL(P, P) = 0;\ KL(P, Q) \geq 0$$

Worst case privacy loss is measured as the maximum privacy loss for all tuples in the sample data table:

$$Ploss = \max Ploss(t)$$

When the distance between the two probability distributions P (t) and Q is smaller, privacy loss will be smaller. So an ideal data without privacy loss will have Ploss = 0.

Privacy loss occurs when the intruder learns privacy information beyond the distribution Q. In order to calculate privacy loss for BinByMean Slicing, the sensitive attribute distribution on the source data is calculated as Q = [8/40 8/40 8/40 8/40 8/40], where each disease repeats 8 times in a sample data table. Next each bucket is processed with KL-distance algorithm in the BinByMean Sliced table. Since L-diversity is respected in the buckets, each disease attribute repeats only once resulting in P B1 (t) = [1/5 1/5 1/5 1/5 1/5], where the total number of tuples in the bucket is t = 5.

Table 7. Privacy loss for BinByMean Slicing

BinByMeanSlicing	**Malaria**	**Cancer**	**Diabetics**	**Typhoid**	**Flu**	**Ploss**
Source Data	0.2	0.2	0.2	0.2	0.2	
BinByMeanSlicing B1…B8	0.2	0.2	0.2	0.2	0.2	**0**

Table 7. shows KL-distance algorithm applied to (P B1 (t), Q) to calculate the PB1loss. KL-distance is calculated for rest of the buckets. Privacy loss on the whole is calculated as the worst case value occurring in any of the buckets. Since L-diversity is respected privacy loss = 0 for all the B1….B8 buckets. Thus maximum privacy loss is achieved in this method.

Table 8. Privacy loss for conventional methods

Slicing	**Malaria**	**Cancer**	**Diabetics**	**Typhoid**	**Flu**	**Ploss**
Source Data	0.2	0.2	0.2	0.2	0.2	
Slicing B1, B2, B5…..B8	0.2	0.2	0.2	0.2	0.2	0
Slicing B3	0.4	0	0.2	0.2	0.2	0.0437
Slicing B4	0.2	0.2	0.2	0.2	0.4	**0.0437**

Bucketization	**Malaria**	**Cancer**	**Diabetics**	**Typhoid**	**Flu**	**Ploss**
Source Data	0.2	0.2	0.2	0.2	0.2	
Bucketization B1, B2, B6….B8	0.2	0.2	0.2	0.2	0.2	0
Bucketization B3	0.4	0.2	0	0.2	0.2	0.0437
Bucketization B4	0.2	0	0.2	0	0.6	**0.1167**
Bucketization B5	0.2	0.4	0	0.2	0.2	0.0437

Generalization	**Malaria**	**Cancer**	**Diabetics**	**Typhoid**	**Flu**	**Ploss**
Source Data	0.2	0.2	0.2	0.2	0.2	
Generalization B1, B2, B6...B8	0.2	0.2	0.2	0.2	0.2	0
Generalization B3	0.4	0.2	0	0.2	0.2	0.0437
Generalization B5	0	0.2	0.4	0	0.4	**0.0592**

Table.8 shows KL-distance algorithm applied to conventional methods like Generalization, Bucketization and Slicing. Bucketization has worst case impact on privacy loss as one of the bucket B4 has disease flu repeated 3 times resulting in Ploss = 0.1167. Buckets B3 and B5 are impacted as well with Ploss = 0.0437. Worst case privacy loss = 0.1167 is reflected for Bucketization. Generalization is also impacted equally but with relatively lesser intensity. Bucket B5 has diseases diabetics and flu repeated 2 times resulting in Ploss = 0.0592.

Slicing has better privacy compared to Bucketization and Generalization. Bucket B4 has disease Flu repeated 2 times resulting in Ploss = 0.0437. Influence of K-anonymity, L diversity and T closeness plays a major role in deflating the privacy loss. BinByMean Slicing shows superior performance than other methods in terms of reduced privacy loss.

5.2. *Accuracy Loss*

Accuracy loss is calculated based on utility loss acquired during the anonymization process. Accuracy loss is calculated comparing anonymized data with the sample source data. Accuracy loss is calculated for a larger group of tuples. It is advisable to have buckets with more tuples as it would help in resembling the source data. Accuracy can also be calculated with respect to anonymized data, without comparing with the source data. In this method, if the source data has low utility, then the anonymized data will also result in low utility. Hence this method is not advisable to calculate accuracy. In order to find accuracy loss, Apriori algorithm is used. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. Using this algorithm, maximum population groups [28] are identified in the anonymized and source data. This large population is identified using support values contributed by only quasi identifiers. P1 = "Age > 28 &< 44 and Sex = M & F" is one of the support function to identify number of records in the anonymized and source data. Similarly Apriori algorithm is used to identify other large populations P2 = "Age > 45 &<53 and Sex = M & F" and P3 = "Age > 37 &< 60 and Sex = M & F". Post finding the large populations we calculate sensitive attribute distribution P1…N for all the large populations of anonymized data. Next sensitive attribute distribution P'1…N are calculated for all the large populations of source data. The distributions are fed to KL-distance algorithm to find the utility loss. Uloss N = KL (P N, P'N). Since accuracy loss is an aggregated loss, the utility losses corresponding to all the populations are averaged to find the overall accuracy loss. Accuracy loss is calculated as

$U_{loss} = (1 / N) * \sum U_{loss\ 1\ldots N}$

Where N = number of large populations identified.

Table 9. Accuracy loss for BinByMean Slicing

BinByMean	Malaria	Cancer	Diabetics	Typhoid	Flu	Aloss	Aloss Avg
P'1	0.27	0.09	0.18	0.18	0.27		
P1	0.3	0.1	0.2	0.2	0.2	0.0153	**0.02673**
P'2	0.2	0.3	0.1	0.2	0.2		
P2	0.2	0.2	0.2	0.2	0.2	0.0523	
P'3	0.2	0.2	0.15	0.2	0.25		
P3	0.2	0.2	0.2	0.2	0.2	0.0126	

Table 9. shows KL-distance algorithm applied to large population of anonymized and source data. P'1 = [3/11 1/11 2/11 2/11 3/11] and corresponding P1 = [3/10 1/10 2/10 2/10 2/10]. P'1 population has 11 records and P1 with 10 records. Accuracy loss = 0.0153 for the first large population. Similarly accuracy loss is calculated for other large populations. Finally the average of all large populations is calculated as 0.0267. This value becomes overall accuracy loss for BinByMean Slicing method. Since this value tends closer to zero, accuracy loss is tolerable and is intact.

Table 10. Accuracy loss for conventional methods

Slicing	Malaria	Cancer	Diabetics	Typhoid	Flu	Aloss	Aloss Avg
P'1	0.27	0.09	0.18	0.18	0.27		
P1	0.3	0.1	0.2	0.2	0.2	0.0153	**0.0051**
P'2	0.2	0.3	0.1	0.2	0.2		
P2	0.2	0.3	0.1	0.2	0.2	0	
P'3	0.2	0.2	0.15	0.2	0.25		
P3	0.2	0.2	0.15	0.2	0.25	0	

Bucketization	Malaria	Cancer	Diabetics	Typhoid	Flu	Aloss	Aloss Avg
P'1	0.27	0.09	0.18	0.18	0.27		
P1	0.3	0.1	0.2	0.2	0.2	0.0153	**0.06163**
P'2	0.2	0.3	0.1	0.2	0.2		
P2	0.1	0.3	0.1	0.1	0.4	0.1386	
P'3	0.2	0.2	0.15	0.2	0.25		
P3	0.15	0.2	0.15	0.15	0.35	0.031	

Generalization	Malaria	Cancer	Diabetics	Typhoid	Flu	Aloss	Aloss Avg
P'1	0.27	0.09	0.18	0.18	0.27		
P1	0.2	0.1	0.3	0.2	0.2	0.0521	**0.08033**
P'2	0.2	0.3	0.1	0.2	0.2		
P2	0.3	0.3	0.2	0.1	0.1	0.1269	
P'3	0.2	0.2	0.15	0.2	0.25		
P3	0.1	0.25	0.2	0.25	0.2	0.062	

Table.10 shows KL-distance algorithm applied to conventional methods like Generalization, Bucketization and Slicing. Generalization has poor accuracy loss compared to other methods, since generalization of data results in utility loss. Bucketization has considerable utility loss compared to Slicing, since data correlation is not well managed in Bucketization. Slicing has better data utility compared to all other methods including BinByMean Slicing.

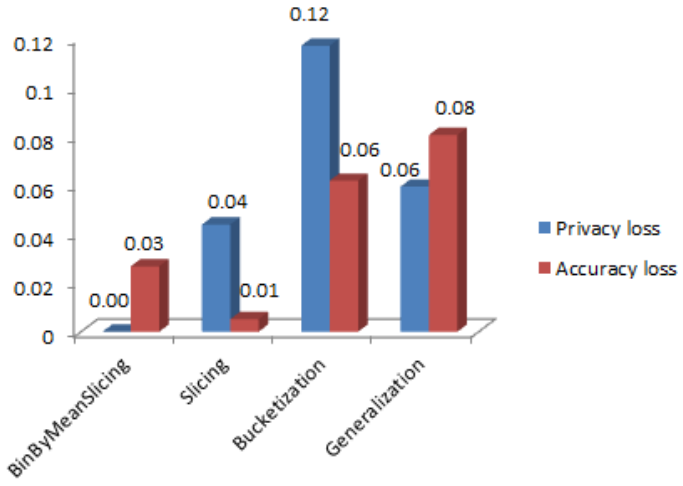

Fig. 1. Result Analysis

Fig.1. shows privacy and accuracy losses plotted for different methods. It is evident that BinByMean Slicing exhibits better privacy and optimum accuracy when compared to conventional methods. Though slicing has better accuracy than BinByMean Slicing, considering both privacy and accuracy parameters BinByMean Slicing has relatively higher performance than conventional slicing. Applying BinByMean value for the age attribute brings anonymization into effect, resulting in reducing the privacy loss. Anonymizing the age attribute by applying average value retains accuracy. In the generalization method, anonymizing the location and gender by replacing with "*" results in worst case accuracy loss.

Table 11. Privacy and Accuracy loss metrics

Method	**K-Anonymity**	**L-Diversity**	**T-Closeness**
BinByMeanSlicing	0	5	0
Slicing	4	4	0.0437
Bucketization	3	3	0.1167
Generalization	4	4	0.0592

Table 11. shows comparition of K-anonymity, L-diversity and T-closeness metrics for all the 4 methods. BinByMean slicing lacks K-anonymity as this method is unable to relaize k tuples in an equivalence class. This is due to fact that the vertical X group is left without any generalization leading to distinct data for vertical X group. This gap is filled by having better L-diversity for this method resulting in preventing attribute disclosure. Other methods like extended K-anonymity [29] can also be used to support the gap. Even though L-diversity might help in preventing attribute disclosure, it is possible for an intruder to gain information about a sensitive attribute as long as she has information about the global distribution of this attribute. To overcome this problem, it is recommended to have T-closeness [30] factor close to 0. T-closeness requires the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. This prevents leakage of individual information.

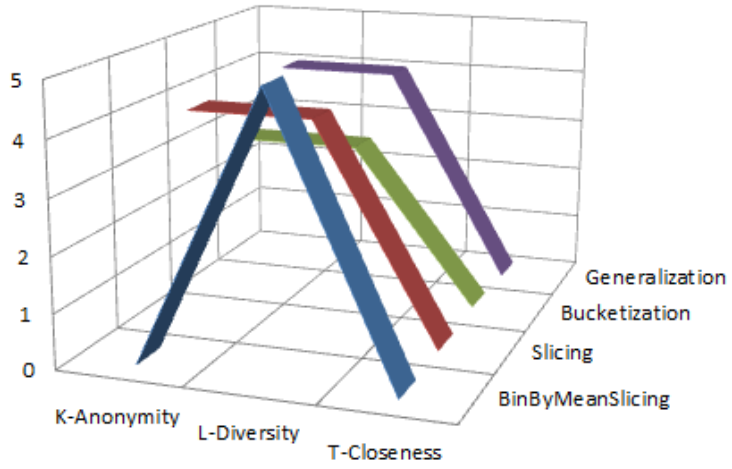

Fig. 2. Metrics Triangle

Fig.2. shows the metrics traingle chart comparing all the 4 methods against K-L-T metrics. From the results BinByMean slicing reflects perfect triangle waveform with K and T metrics falling at 0 indices and L metrics with 5 index. This index positions reflects an ideal method which offers better privacy and accuracy leaving behind other methods.

6. Conclusion

BinByMean Slicing looks promising as it offers better privacy and accuracy for sensitive medical data sets with (Ploss, Aloss) = (0,0267). This method has enhanced L-diversity and T-closeness values resulting in superior performance against conventional methods. Compared to privacy, there is a limitation in accuracy which is evident from KL-distance algorithm results. This paper leaves room for future work on improving the accuracy while still retaining the privacy loss level.

References

[1] Brickell J and Shmatikov V 2008 The cost of privacy: destruction of data-mining utility in anonymized data publishing. Int'l Conf. Knowledge Discovery and Data Mining KDD, pages 70–78.

[2] Mathew G and Obradovic Z 2011 A privacy-preserving framework for distributed clinical decision support, Proceedings of IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), pp.129-134.

[3] Li T and Li N 2009 On the Tradeoff between Privacy and Utility in Data Publishing. Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 517-526.

[4] Yaping li, Minghua Chen, Qiwei Li and Wei Zhang 2012 Enabling multilevel trust in privacy preserving data mining. IEEE Transaction on Knowledge and Data Engineering, YoI.24, No.9.

[5] Durgesh Kumar Mishra, Priyanka Jangde and Gajendra Singh Chandel 2011 Hybrid Technique for secure sum protocol. World of Computer Science Information Technology Journal (WCSIT), Vol.I, No. 5, pp 198-201.

[6] Bhanumathi S and Sakthivel P 2012 Privacy Preserving Multiparty Collaborative Mining using Integer Programming Model. Conference on Recent Trends in Computer and Networking Technologies, pp 21-23

[7] Samarati P 2001 Protecting Respondent's Privacy in Microdata Release. IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027.

[8] Martin D.J, Kifer D, Machanavajjhala A, Gehrke J and Halpern J.Y 2007 Worst-Case Background Knowledge for Privacy- Preserving Data Publishing. Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135.

[9] Koudas N, Srivastava D, Yu T and Zhang Q 2007 Aggregate Query Answering on Anonymized Tables. Proc. IEEE 23rd Int'l Conf.Data Eng. (ICDE), pp. 116-125.

[10] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan and Yong Ren 2014 Information Security in Big Data: Privacy and Data Mining. IEEE Access, Vol 2.

[11] Liyue Fan and Hongxia Jin 2015 A Practical Framework for Privacy-Preserving Data Analytics. International World Wide Web Conference Committee, 18–22.

[12] Lei Xu, Chunxiao Jiang Yan Chen, Yong Ren and Ray Liu 2015 Privacy or Utility in Data Collection? A Contract Theoretic Approach. IEEE Journal of selected topics in signal processing, Vol. 9, No. 7.

[13] Yang Xu, Tinghuai Ma, Meili Tang and Wei Tian 2014 A Survey of Privacy Preserving Data Publishing using Generalization and Suppression. International Journal of applied Mathematics and Information Sciences, 8, No. 3, 1103-1116.

[14] Aggarwal C 2005 On k-Anonymity and the Curse of Dimensionality. Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909.

[15] Kifer D and Gehrke J 2006 Injecting Utility into Anonymized Data Sets. Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 217-228.

[16] Xiao X and Tao Y 2006 Anatomy: Simple and Effective Privacy Preservation. Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150.

[17] Nergiz M.E, Atzori M and Clifton C 2007 Hiding the Presence of Individuals from Shared Databases. Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 665-676.

[18] Sweeney L 2002 k-Anonymity: A Model for Protecting Privacy. Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570.

[19] Muthulakshmi N.V and Sandyarani K 2012 Privacy preserving association rules mining in vertically partitioned databases. Proceedings of IJCA, VoI.39, No. 13 pp 29-35.

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

[21] Kifer D and Gehrke J 2006 Injecting Utility into Anonymized Data Sets. Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 217-228.

[22] Machanavajjhala A, Gehrke J, Kifer D, and Venkitasubramaniam M 2006 Diversity: Privacy beyond k-Anonymity. Proc. Int'l Conf. Data Eng. (ICDE), p. 24.

[23] Emad Elabd, Hatem Abdulkader and Ahmed Mubark 2015 L–Diversity-Based Semantic Anonymization for Data Publishing. I.J. Information Technology and Computer Science, 10, 1-7.

[24] Tianqing Zhu, Ping Xiong and Gang Li 2015 Correlated differential privacy: Hiding information in non- IID data set. IEEE Transactions on Information Forensics and Security, Vol.10, No.2.

[25] Jordi Soria-Comas, Joseph Domingo-Ferrer, David Sanchez and Sergio Mart 2015 t-Closeness through Micro aggregation: Strict Privacy with Enhanced Utility Preservation. IEEE Transactions on Knowledge and Data Engineering, Vol.27, No.11.

[26] Christy Thomas and Diya Thomas 2013 A survey on Privacy Preservation in Data Publishing. International Journal of Computer Science and Engineering Technology, Vol.4, No.5.

[27] Javier Herranz, Jordi Nin, Pablo Rodriguez and Tamir Tassa 2015 Revisiting distance based record linkage for privacy preserving release of statistical datasets. Data & Knowledge Engineering, 100, 78-93.

[28] Jaideep Vaidya, Basit Shafiq and Wei Fan 2014 A random decision tree framework for Privacy Preserving Data Mining. IEEE Transaction of Dependables and Secure Computing, Vol 11, No.5.

[29] Masoud Rahimi, Mehdi Bateni and Hosein Mohammadinejad 2015 Extended K-Anonymity Model for Privacy Preserving on Micro Data. I.J. Information Technology and Computer Science, 12, 42-51.

[30] Salvatore Ruggieri 2014 Using t-closeness anonymity to control for non-discrimination. Transactions on Data Privacy 7, 99–129.