An Improved Clustering Realized Relational Data Anonymization with Optimal Privacy and Utility Measures

G. Sasirekha, S. Kishore Verma, S. Sheik Faritha Begum, J.S.Adeline Johnsana

Abstract*: Massive growth of technology results the increased usage of computer in day to day life. Every user feeds millions of data for every minute. The process of converting this raw data into useful information is called data mining. The need for preservation of data for its privacy is called Privacy Preserving Data Mining (PPDM). In recent years privacy preserving data mining has become more crucial because of increased storage of digital collection of data about users in many of government sectors, corporate, hospitals, banks, etc., This collection of data contains many sensitive attributes, which reveals their identity of the users by combining the data's with publicly available data's, which had been stolen by hackers. To prevent from this, a protection model called k-anonymization is introduced. This k-anonymity model preserves the individual identity through generalization and suppression. Privacy and utility measures are inversely proportional to each other. The need to maintain a tradeoff between privacy and utility is a vital factor in PPDM. In this paper, CARD (Clustered Anonymization of Relational Data) is presented to reduce the information loss of utility aware anonymization. The utility aware anonymization means k-anonymizing the dataset by accounting the two novel factors, transformation pattern loss (tpl) and null value count having minimum values. This utility aware anonymization is done for Cell oriented Anonymization (CoA), Attribute oriented Anonymization (AoA) and Record oriented Anonymization (RoA). CARD proceeds in clustering the given dataset with various benchmarked clustering algorithms like Simple K-Means (KMeans), Farthest First (FF), Expectation Maximization (EM), Partition around Medoids (PAM) and Gower method then this clustered data set are subjected to utility aware CoA, AoA and RoA anonymization approaches. Classification analysis like logistic regression, naïve bayes and random forest are done on clustered anonymized data set to assess and prove the privacy and utility of the proposed approach based on Information Loss, Re-Identification Risk and Classification Accuracy of the clustered dataset before publishing them. Our experimental results prove to be better than the non-clustered anonymization procedures. Among the five clustering algorithms, In our analysis Gower and Partition around Medoids (PAM) results give better solution in terms of privacy and utility since PAM and Gower approaches are the best clustering methods that are capable of clustering mixed data type (numerical and categorical).

***Index Terms*: K Anonymization, Cell Oriented Anonymization, Attribute Oriented Anonymization, Record Oriented Anonymization, Partition around Medoids, Gower**

Revised Manuscript Received on December 22, 2018.

G. Sasirekha, Computer science and Engineering, C.Abdul Hakeem College of Engineering and Technology, Vellore, India.

S. Kishore Verma, Computer science and Engineering, C.Abdul Hakeem College of Engineering and Technology, Vellore, India.

S. Sheik Faritha Begum, Computer science and Engineering, C.Abdul Hakeem College of Engineering and Technology, Vellore, India.

J.S.Adeline Johnsana, Information Technology, Adhiparasakthi College of Engineering, Kalavai,India

I. INTRODUCTION

In recent years due to the large usage of computers, data publishing became an everyday activity of government sectors, corporate, hospitals, banks, and many organizations. During the entire cycle of storing these records, which typically contain sensitive men or woman information with personal, health and financial data that frequently get reveled to numerous parties which includes hackers, miners, etc [1]. Privacy preserving data mining (PPDM) has originated as an essential concern to interrelate the collection, analysis, and distribution of data and aims to protect the exposure of individuals' private and sensitive information from the published data [2]. Privacy preserving data mining (PPDM) deals with privacy of individual records without sacrificing the utility of the records [3]. To protect from information leakage, privacy preservation methods have been developed to protect owner's exposure. PPDM methodologies are designed to guarantee a certain level of privacy, while maximizing the data utility, so that data mining can still be performed on the transformed data efficiently [4]. One of the main methodologies to keep the identity of individuals when releasing or distributing the sensitive data is to anonymize it [5]. A famous methodology for data anonymization is k-anonymity [6,7]. With k-anonymity an original data set containing personal sensitive information can be changed so that it is hard for a trespasser to conclude the identity of the individuals in that dataset [7]. A k-anonymized data set has the property that each record is similar to at least another k-1 other records on the potentially identifying variables. For instance, if k = 5 and the potentially identifying factors like age and gender, then a k-anonymized dataset has at least 5 records for each value grouping of age and gender. The most widely used implementations of k-anonymity use transformation techniques such as generalization and suppression [8]. In [9] Jun-Lin Lin et al proposed a clustering-based k-anonymization technique called One-pass K-means Algorithm (OKA) which is derived from the K-Means algorithm, difference in OKA from K-means is it run for only one iteration. It reduces the information loss which was running in O(n2/k) time instead of O(n2) but it improves the resilience to outliers when number of iteration is increased during clustering stage. In [10] Chitra et al proposed a clustering based k-anonymization called Fuzzy C-Means (FCM) which is optimized under Particle Swarm Optimization (PSO) algorithm. Even though it provides less information loss but they carried out the experimental results by eliminating few attributes in adult dataset and

also they separated the dataset into numerical and categorical data type which is not examined as whole set. In [11] Ji-Won Byun et al proposed a new specific clustering algorithm called greedy k-member clustering problem which organize the data values that are obviously like to one another and also should be part of the same equivalence class but the problem is execution time for the greedy k-member algorithm is higher than the partitioning algorithm.

Our proposed work Clustered Anonymization of Relational Data (CARD) Model works with distinct perspective contributions as follow.

i. CARD model is mainly framed to reduce the information loss, Re-identification Risk and increased accuracy of all the data quality models in terms of clustering the dataset.

ii. We experimented the dataset mainly with Partition around Medoids (PAM) and Gower method because that algorithm goes well for mixed data types (Numerical and Categorical) without separating individually.

The rest of the paper is organized as Section 2 discusses the related works; Section 3 introduces the basic definitions, concepts that are required to have better understanding. Section 4 presents the proposed approach and discusses its importance. Section 5 explains the methodology of the proposed approach .Section 6 shows the implementation and experimental analysis of the proposed approach. Section 6 concludes the benefits for our proposed approach.

II. RELATED WORKS

Samarati and Sweeney [12] proposed the k-anonymity. If one record in the table has some value qid, at least k − 1 other records also have the value qid. A table is called as k-anonymous table in which each record is indistinguishable from at least k-1 other records with respect to QID. This method checks for the possible k-anonymous solutions in different levels of Domain Generalization Hierarchy [13]. It looks for solution with least generalization and also output has a chance to be in optimal. It produces the great result when compared to Datafly [12]. But this as a chance of getting the optimal solution dramatically varies with k-Anonymity model protects the identification information but not protecting the sensitive relationship in a dataset [14].According to Sweeney, the best solutions are attained after generalizing the variables with the unique values. This approach only goes through a very small number of nodes in the lattice to find its solution. Thus, from a time perspective, this approach becomes very efficient. The algorithm called MinGen [12] proposed by Sweeney is to check only very few nodes for k-anonymity which able to give result very fast. But this method is to skips many nodes, therefore, the resulting data is much generalized and sometimes this released data may not be suitable for research purpose as it provides very little information. K-Anonymity protects the identification information but not sensitive relationship in data [14].Incognito [15] algorithm produces all the possible k-anonymous full-domain generalizations of a relation with an optional tuple suppression threshold. It starts by checking single-attribute subsets of the quasi-identifier, and then iterates, checking k-anonymity with respect to larger subsets of quasi-identifiers. This method is to find all the k-anonymous full domain generalizations and the optimal solution can be selected according to different criteria. But this method uses breadth first search method that takes a lot of time to traverse the solution. To overcome this new privacy model l-diversity [16] is introduced. It requires the distribution of a sensitive attribute in each equivalence class with at least l "well- represented" values. l-diversity provides privacy even when the data publisher does not know what kind of knowledge is possessed by the adversary. The main goal of l-diversity is the value of the sensitive attributes is well represented in each group [16] by using equivalence class has at least l well represented values for sensitive attributes. But this method is it is difficult and unnecessary to achieve. It is insufficient to prevent attribute disclosure. To address the limitations of l-diversity, another privacy model t-closeness is proposed [17]. It formalizes the idea of global background knowledge by requiring that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). This effectively limits the amount of individual-specific information an observer can learn [17]. But using of this method is it is not perfect and information gain is not clear.(α, k) Anonymity method is proposed to overcome the limitation of Incognito algorithm. In (α, k) Anonymity α is a fraction and k is an integer. After anonymization of the equivalence class the frequency of the sensitive values should not more than α [14]. The importance of this method is to protect both identification and relationships in data. It limits the confidence of implications of the quasi-identifier to a sensitive value within α. The drawback of this proposal is achieving optimal (α, k) anonymity is NP-hard. Algorithm is not scalable and has lot of distortions due to global based recording.

Our proposed work concentrated on the anonymization of dataset is an essential concern to maintain the privacy of the dataset when publishing. For anonymization of the dataset clustering technique is applied. Clustering is done by implementing Expected Maximization (EM), Farthest First (FF), Simple K-Means (KMeans), Partition around Medoids (PAM), Gower algorithms are introduced. Generalization and Suppression are used to maintain secrecy of the dataset and get the anonymized view of the dataset..

III. DEFINITIONS AND CONCEPTS

A. k-Anonymization:

It is carried out on partial identifiable attributes and it frames data attributes in this sort of manner that the possibility of finding out an individual from an attribute at maximum by means of k or in different words we are able to say that it offers at least k-1 similar entities for an attribute set.

1) Generalization

Generalization replaces some values with a parent value in the taxonomy of an attribute. The reverse

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

operation of generalization is called specialization. Generalization consists of substituting attribute values with semantically consistent but less precise values. For example, the month of birth can be replaced by the year of birth which occurs in more records so that the identification of a specific individual is more difficult.

2) Suppression

Suppression replaces some values with a special value, indicating that the replaced values are not disclosed. The reverse operation of suppression is called disclosure. Suppression can be thought of as a special kind of generalization. Suppression can drastically reduce the quality of the data if not properly used. This is the reason why most k-anonymity related studies have focused on generalization.

For example, data contain Quasi-Identifier attributes such as gender, age and pin code. Gender = 'male', age = 32, pin code = 632014, may also constitute all of the men and women of elderly 32 in the area 632014. Anonymization is carried out via the generalization of the information. This generalization is carried by the suppressing and displaying the data by using the symbol '*' values. After anonymization the result would be gender = '*', age = '30-60', and pin code = '632***'. This generalization is achieved on the premise of cardinality or range of entities in the hierarchy. If the cardinality of an attribute is greater, then it follows the lesser suppression on data and vice versa.

According to our k-anonymization strategy the attributes of the dataset is classified as

(i) Qid – Quasi Identifier, are the attributes which are considered as the linking attributes that are exposed to linking attacks.

(ii) Sa- Sensitive attributes are the attributes which should not be correlated with the particular individual as account of linking attacks.

(iii) Ia-Identifying attributes are the direct signifiers of the records i.e. explicitly reveal the identity of the individual.

B. Data quality models

Three kinds of data quality models as mentioned below are experimented

1) Cell oriented anonymization (CoA)

Cell oriented Anonymization procedure works on the principle of generalization that is taking place with respect to cell values.

2) Attribute oriented anonymization (AoA)

This procedure works on the principle of generalization done with respect to each column/attribute of a given dataset/recordset.

3) Record oriented anonymization (RoA)

Record Oriented Anonymization (RoA) works on the principle of applying anonymization procedure with respect to all quasi identifier's domain hierarchy.

IV. PROPOSED WORK

CARD (Clustered Anonymization of Relational Data) aims to reduce the information loss and risk factor with increased accuracy for utility aware anonymization. The utility aware anonymization means k-anonymizing the dataset with reduced two novel factors transformation pattern loss and null value count. CARD starts with clustering the given dataset with different clustering algorithms approaches like k-means, farthest first and expectation maximization, partition around Medoids, gower methods and then this clustered data set are undergoes to utility aware CoA, AoA, RoA anonymization approaches. Classification analysis like logistic regression, naive bayes and random forest are done on CARD's anonymized data set to assess and prove the utility of the proposed approach based on accuracy for the clustered data set before publishing. Fig 1 explains the diagrammatic representation of the data flow among the modules.

Benefits of the Proposed System:

1. Attains decreased information loss for the clustered data set when compared with non-clustered dataset anonymization for CoA, AoA, RoA.

2. Attains decreased Re-identification Risk for the clustered data set when compared with non-clustered dataset anonymization for CoA, AoA, RoA.

3. Increased Accuracy for the clustered dataset anonymization for CoA, AoA, RoA.

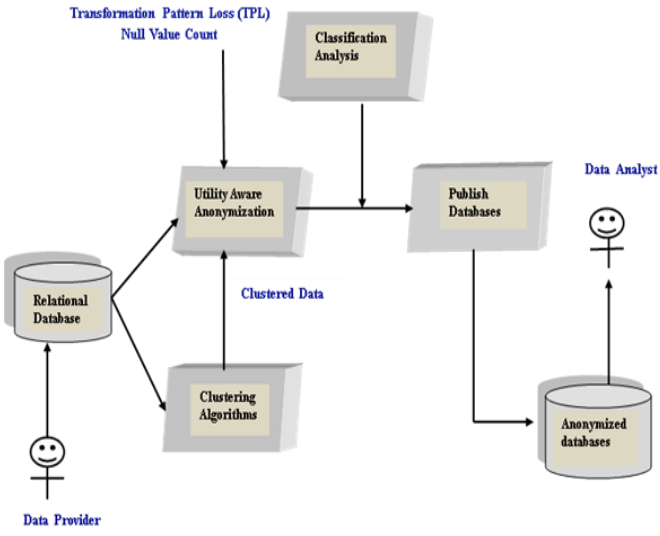

Fig. 1: Architecture Diagram

V. METHODOLOGY

A. Clustering

Clustering is the process of grouping of similar objects together. Clustering is carried out for Cell Oriented Anonymization (CoA), Attribute Oriented Anonymization (AoA) and Record Oriented Anonymization (RoA). CARD proceeds in clustering the given dataset with various benchmarked clustering algorithms.

a) Expectation Maximization
b) Farthest First
c) Simple K Means
d) Partition Around Medoids
e) Gower

1) Expectation Maximization (EM) Algorithm

This is basic and straight forward to implement and iterative refinement. EM algorithm repeats and advances the likelihood of seeing observed data while evaluating the parameters of a statistical model with unobserved variables. It is an expansion of the k-means algorithm, which assigns an object to the cluster with which it is most comparative, in light of the cluster mean dedicated cluster. EM allots each object to a cluster according to a weight speaking to the likelihood of participation. In other words, there are no strict confinements between clusters. In this manner, new earnings are figured dependent on weighted measures. EM starts by making a guess at the model parameters. At that point it pursues an iterative 3-step process:

Algorithm for Expected Maximization:
1. **E-step:** Based on the model parameters, it ascertains the probabilities for assignments of every data point to a cluster.
2. **M-step:** Update the model parameters dependent on the cluster assignments from the E-step.
3. Repeat until the model parameters and cluster assignments balance out.

2) Farthest First (FF) Algorithm

The farthest-first is a quick and greedy algorithm. In this algorithm k points are first chosen as cluster centers. The primary center is select arbitrarily. The second center is greedily select as the point farthest from the first. Each staying center is controlled by greedily choosing the point most distant from the set of already chosen centers, and the rest of the points are added to the cluster whose middle is the nearest.

Algorithm for Farthest First:
1. Farthest first (d: informational index, k: integer) {
2. Randomly select first point;
3. //select centers
4. For (i= 2,… ,k) {
5. For (each outstanding point) {compute distance to the current center set;}
6. Select the point with greatest separation as new center;}
7. //allot remaining point for (each outstanding point) {
8. Calculate the distance to each cluster center;

3) Simple K-Means Algorithm

K-Means clustering is a method for cluster analysis which intends to segment n perceptions into k clusters in which every announcement has a place with the cluster with the closest mean. The principle thought is to characterize centroids for each cluster. These centroids must be situated calculatingly as a result of various position causes distinctive outcome. In this way, the best decision is to put them however much as could reasonably be expected far from one another. Each point having a place toward a given dataset is associate to the closest centroid. When no point is anticipating, we have to recalculate k new centroids of the clusters coming about because of the past advance. It then uses an iterative relocation technique that endeavors to enhance the partitioning by moving object starting with one gathering then onto the next. The general proportion of a good partitioning is that objects in a similar cluster are "close" or identified with one another, while distinctive cluster's objects are "far separated" or altogether different.

Algorithm for Simple K-Means:
1. Place k points into the area represented by the objects that are being clustered. These points represented as starting centroids.
2. Assign each object to the group that has the close-by centroid.
3. When all objects have been assigned, recalculate the locations of the k centroids.
4. Repeat Steps 2 and 3 until the centroids never again move. This makes a division of the objects into groups from which the metric to be limited can be computed.

4) Partition around Medoids (PAM) and Gower distance Function

The working of K-Medoids clustering algorithm is similar to K-Means clustering. It also begins with randomly selecting k data items as initial medoids to represent the k clusters. All the other remaining items are included in a cluster which has its medoid closest to them. Thereafter a new medoid is determined which can represent the cluster better. All the remaining data items are yet again assigned to the clusters having closest medoid. Consider in every iteration, the medoids alter their location. The method minimizes the sum of the dissimilarities between each data item and its corresponding medoid. This cycle is repeated till no medoid changes its placement. This marks the end of the process and we have the resultant final clusters with their medoids defined. K clusters are formed which are centered on the medoids and all the data members are placed in the appropriate cluster based on nearest medoid. The Gower function computes the dissimilarity among units in a dataset or among observations in two distinct datasets. It first computes distances between pairs of variables over two data sets and then combines those distances to a single value per record-pair.

Algorithm for PAM/ Gower:
Input:k: number of clusters; D: the data set containing n items
Output:A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoids.
Steps:
1:*Calculate the Gower distance method by using daisy function and export the newly calculated dataset.
2: Arbitrarily choose k data items as the initial medoids.
3: Assign each remaining data item to a cluster with the nearest medoid.

B. Utility Enhanced

Anonymization

1) Null Value Impact

In [18, 19] for a recordset S and the generalized form S of S*, if S* is accomplished by k-anonymized. At that point if there exist N (S*) = *, means the original values are replaced with * null value. Null value impact means the presence of null values (null Values count) the anonymized recordset S* will terribly influence the accuracy of the data mining task. So, it is prudent to have a strategy that has a smaller number of null values after anonymization. Hence null Values count nVC is indirectly proportional to accuracy in equation (1) and directly proportional to Information Loss (IL) in equation (2).

$$nVc \propto \frac{1}{classificationAccuracy} \quad (1)$$

$$nVc \propto IL. \quad (2)$$

On these two perceptions, the algorithms are analyzed

2) Transformation pattern Loss (TPl)

Transformation pattern is the reference string represented as level of generalization hierarchy adopted by each attribute with anonymization in equation (3).

Transformation pattern Loss is the cosine distance between expected transformation pattern and actual transformation pattern in equation (5).

$$P = \coprod_{i=1}^{n} (Ae) \quad (3)$$

Transformation pattern is formed by combining each attributes transformation level string after anonymization. Let P be the vector of the expected Transformation pattern representation. Ae is each attribute's level of transformation in expected form.

$$P = \coprod_{i=1}^{n} (Aa) \quad (4)$$

Q be the vector of actual transformation pattern representation... Aa is each attribute's levels of transformation in actual form.

Transformation pattern Loss (TpL) =$Cosinedista$nce (P,Q) (5)

C. Classification Analysis

Classification analysis is used to find the accuracy of the clustered data. The classification accuracy is calculated by using below algorithms:

i. Logistic Regression

ii. Naive Bayes

iii. Random Forest

1) Logistic Regression

Logistic regression is used for classification tasks. Logistic regression predicts the likelihood of a result that can only have two values (i.e. a dichotomy). The expectation is based on the use of one or several predictors (numerical and categorical).

2) Naive Bayes

The Naive Bayesian classifier depends on Bayes' hypothesis with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it especially useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is broadly used because it often outperforms more sophisticated classification methods.

$$P\left(\frac{c}{x}\right) = \frac{P\left(\frac{x}{c}\right)P(c)}{P(x)} \quad (6)$$

•P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
•P(c) is the prior probability of class.
•P(x|c) is the probability which is the likelihood of predictor given class.
•P(x) is the prior probability of predictor.

3) Random Forest

Random Forest algorithm is one of the best among classification algorithms - able to classify large amounts of data with accuracy.

1. If the number of chances in the training set is N, sample N cases at random - but with substitution, from the original data. This example will be the training set for developing the tree.

2. If there are M input factors, a number mM is specified such that at each node, m factors are chosen at random out of the M and the best split on these m is utilized to split the node. The value of m is held constant during the forest growing.

3. Each tree is grown to the largest extent possible. There is no pruning.

Once the data is done with classification analysis we can publish the data and submit it to data analyst. The clustered dataset has increased privacy without compromising the utility measure. The information loss is decreased with increased accuracy. We can also see there is a decrease in re-identification risk for the clustered data.

VI. EXPERIMENTAL EVALUATION

In CARD, Clustering is carried out by using open source tool WEKA (EM, FF, and Simple K-Means). K-anonymization is executed by open source anonymization tool ARX. The experiments were executed on machine running 64-bit windows 8.1 and above, Intel core i3 processor with 4GB.In this experimental evaluation, UCI Machine learning repository -Adult dataset is utilized. This dataset comprises of 30162 records with 9 attributes. Hierarchy level has set for each attribute during anonymization. Each attribute has different set of hierarchy levels which is explained in the table 1 Dataset Description. Values are recorded based on the generalization hierarchy levels.

Table 1: Dataset Description

S.No	Name of the Attribute	Attribute Category	Generalization Levels in Hierarchy
1.	Sex	Qid	0
2.	Age	Qid	1-3
3.	Race	Qid	1
4.	Marital Status	Qid	1
5.	Education	Qid	1-2
6.	Native Country	Qid	1
7.	Work Class	Qid	1
8.	Occupation	Qid	1
9.	Salary	Sa	0

We compute the information loss, re-identification risk and accuracy for CoA (Loss-Geometric mean), AoA (Normalized non-inform Entropy-Rank) and RoA (Discernability) measures [20]. Decreased information loss and re-identification risk, increased accuracy set up the good throughput on the clustered anonymized record set. We measured these three metrics in various parametrical setup and the results are demonstrated.

A. *Cell oriented Anonymization (CoA)*

Cell oriented Anonymization procedure works on the principle of generalization that is taking place with respect to cell values. In CoA we calculated the information loss, Re-identification Risk and accuracy by using logistic regression, Naïve bayes and Random Forest algorithm in loss-geometric mean measurement of all attributes. Comparative analysis of all the three factors is taken by using Expectation Maximization (EM), Farthest First (FF) and Simple K-Means, PAM and Gower Algorithm. Set Suppression Limit to maximum level (100%) for each iteration.

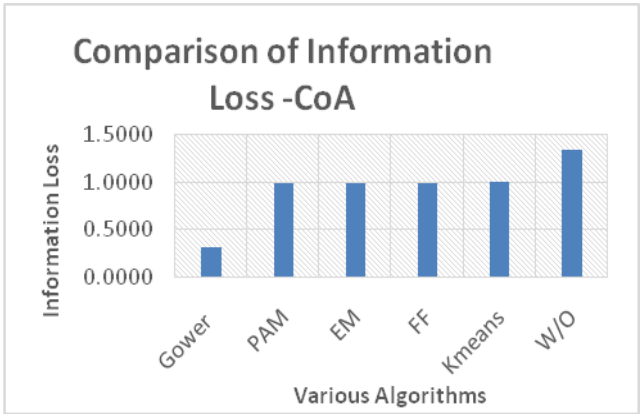

Fig. 2: Comparison of information loss across various algorithms

	Logistic Output		Naïve Bayes Output		Random Forest Output	
Algorithm	W/O	Clustered	W/O	Clustered	W/O	Clustered
EM	80.6057	85.4502	75.0599	81.5296	75.2323	85.6378
FF	80.6057	80.9550	75.0599	79.1676	75.2252	76.2163
Kmeans	80.6057	84.4429	75.0599	82.0291	75.2200	81.7148
PAM	80.6062	81.3662	75.0348	76.4407	75.1622	76.2241
Gower	80.6109	92.8998	75.0358	88.5187	75.2053	92.9205

Fig. 3: Comparison of output accuracy among various algorithms

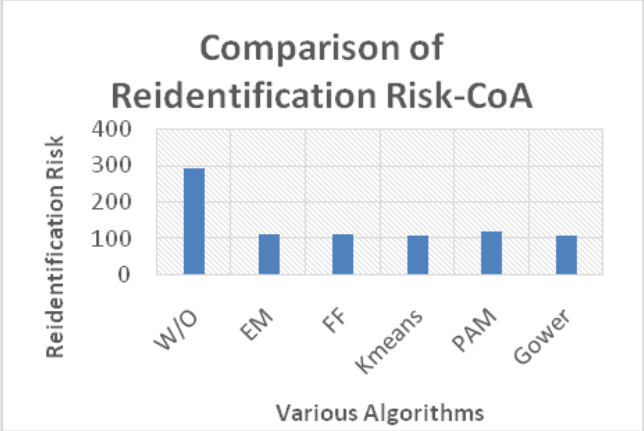

Fig. 4: Comparison of Re-identification risk across various algorithms

From the figure 2 we can able to see the information loss is lower to Gower method compared to all other algorithms and non-clustered dataset. Figure43 shows non-clustered dataset having the higher possibility of re-identification risk across different algorithms. Figure 3 shows non-clustered dataset output accuracy of all the three methods logistic regression, naïve bayes and random forest is lower than clustered(EM, FF, KMeans, PAM, Gower) algorithms accuracy and also gower method has a best accuracy result when compared with other four clustered algorithms.

B. *Attribute oriented Anonymization (AoA)*

This procedure works on the principle of generalization done with respect to each column/attribute of a given dataset/recordset. This model quantifies loss of information based on mutual information, which measures the amount of information that can be obtained about the original values of variables in the input dataset by observing the values of variables in the output dataset. We calculated Non-Uniform entropy-Rank measurement by using logistic regression, Naïve bayes and Random Forest algorithm with geometric mean measurement of all attributes. Comparative analysis of all the three factors is taken by using Expectation Maximization (EM), Farthest First (FF) and Simple K-Means, PAM and Gower Algorithm. Set Suppression Limit to maximum level (100%) for each iteration.

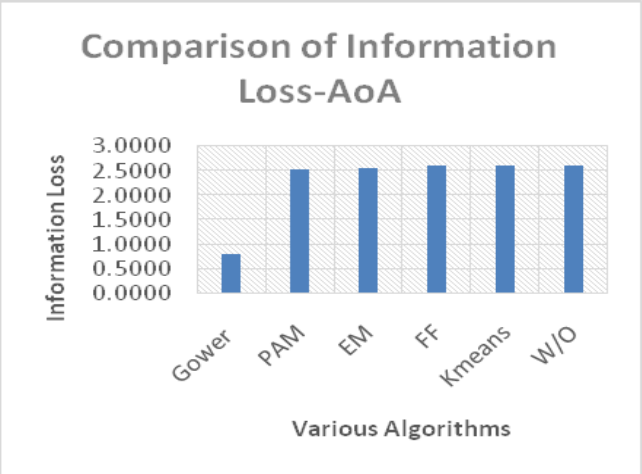

Fig. 5: Comparison of information loss across various algorithms

	Logistic Output		Naïve Bayes Output		Random Forest Output	
Algorithms	W/O	Clustered	W/O	Clustered	W/O	Clustered
EM	80.6568	85.4104	75.3517	81.5321	75.1816	85.6378
FF	80.6668	80.6477	75.3531	78.2268	75.1617	74.7386
Kmeans	80.6611	84.4442	75.3569	82.0348	75.2176	81.5212
PAM	80.6616	81.2662	75.3375	76.4451	75.1859	76.4836
Gower	80.6559	93.0949	75.3389	88.7258	75.2086	92.9205

Fig. 6: Comparison of output accuracy among various algorithms

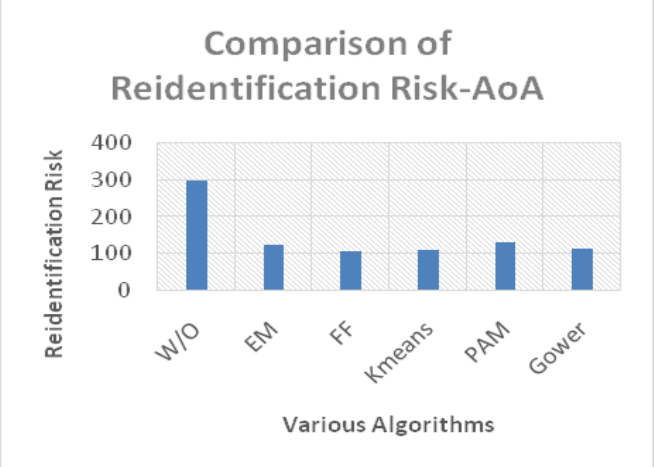

Fig. 7: Comparison of Re-identification risk across various algorithms

From the figure 5 we can able to identify that the information loss is lower to clustered algorithms dataset compared to all non-clustered dataset. Figure 7 shows non-clustered dataset having the higher possibility of re-identification risk across different algorithms. Figure 6 shows non-clustered data output accuracy is lower when compared with EM, KMeans, PAM, and Gower except FF and also Gower method has a best accuracy result when compared with other clustered algorithms.

C. *Record oriented Anonymization (RoA)*

Record Oriented Anonymization (RoA) works on the principle of applying anonymization procedure with respect to all quasi identifier's domain hierarchy. Here the anonymized dataset/record set is obtained by applying the generalization procedure to the vector of quasi identifier value in each record of the recordset. This model also estimates data quality based on the size of the equivalence classes in the output dataset. It does not take into account the actual attribute values in the output dataset. We calculated discernability measurement by using logistic regression, Naïve bayes and Random Forest algorithm with geometric mean measurement of all attributes. Comparative analysis of all the three factors is taken by using Expectation Maximization (EM), Farthest First (FF) and Simple K-Means, PAM and Gower Algorithm. Set Suppression Limit to maximum level (100%) for each iteration.

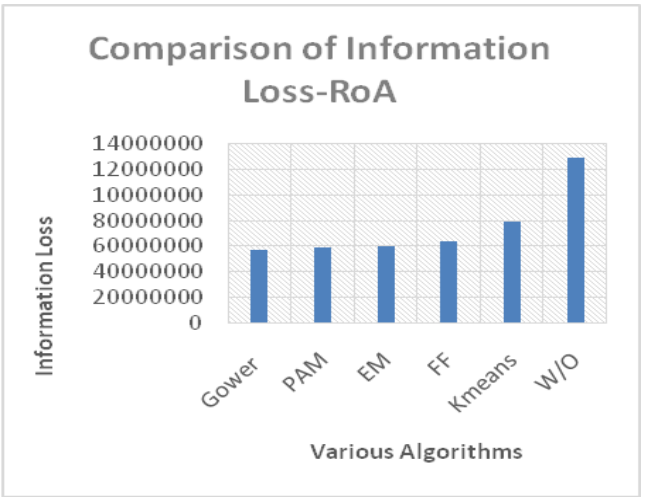

Fig. 8: Comparison of information loss across various algorithms

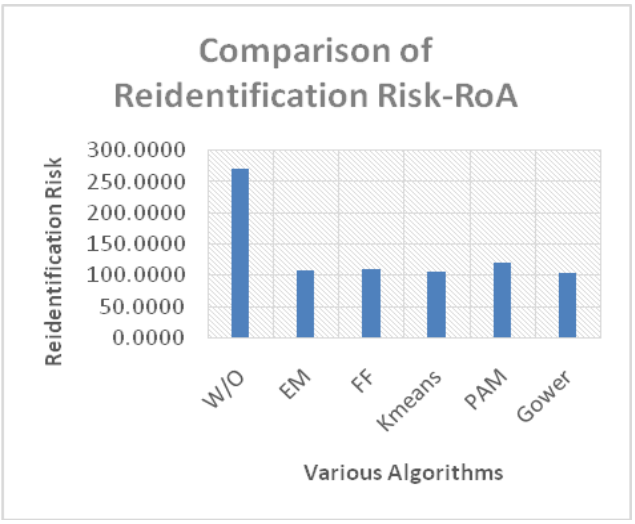

Fig. 9: Comparison of Re-identification risk across various algorithms

	Logistic Output		Naïve Bayes Output		Random Forest Output	
Algorithms	W/O	Clustered	W/O	Clustered	W/O	Clustered
EM	80.5673	85.4525	75.0585	81.5356	75.2371	85.6378
FF	80.5650	80.6477	75.0637	78.2268	75.1958	74.6810
Kmeans	80.5650	84.4513	75.0637	82.0249	75.2247	81.6675
PAM	80.5664	81.3290	75.0613	76.4273	75.1973	76.1672
Gower	80.5796	92.8585	75.0571	88.4550	75.1793	92.9205

Fig. 10: Comparison of output accuracy among various algorithms

From the figure 8 we can able to identify that the information loss is lower to

clustered algorithms dataset compared to all non-clustered dataset. Figure 9 shows non-clustered dataset having the higher possibility of re-identification risk across different algorithms. Figure 10 shows non-clustered data output accuracy is lower when compared with EM, KMeans, PAM, and Gower except FF and also Gower method has a best accuracy result when compared with other clustered algorithms. Gower method provides best minimum loss and accuracy with respect to other algorithms because Gower distance function works well on mixed data types (both numeric and categorical).

VII. CONCLUSION

In the paper, to achieve clustering based anonymization for preserving the privacy of the individual's sensitive data on various databases, CARD is used. Here we successfully implemented the PAM and Gower algorithms that can able to lower information loss and risk and also increases accuracy by using generalization and suppression on the data set when compared with non-clustered data. In future we calculate the information loss by using generalization and suppression to l-diversity and t-closeness measure and conclude the comparative study of different methodologies of Privacy Preserving Data Mining (PPDM).

REFERENCES

1. V. Shyamala Susan and T. Christopher, "An Efficient Anonymization Model (EAM) For Data Publishing Using Optimized Clustering Approach," International Journal of Pure and Applied Mathematics, 2017, 118(19),2743-2459.
2. G. Poulis, G. Loukides, A. Gkoulalas-Divanis, and S. Skiadopoulos, "Anonymizing data with relational and transaction attributes," In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2013, (pp. 353-369), Springer, Berlin, Heidelberg. O. Abul, F. Bonchi, and M. Nanni, "Anonymization of moving objects databases by clustering and perturbation.", Information Systems, 2010, 35(8), 884-910.
3. Hina Vaghashia and Amit Ganatra, "A Survey: Privacy Preservation Techniques in Data Mining," International Journal of Computer Applications, 2015, (0975 – 8887) , Volume 119 – No.4.
4. R. Mendes and J.P. Vilela, " Privacy-Preserving Data Mining: Methods, Metrics, and Applications," IEEE Access, 5, 2017, 10562–10582.doi:10.1109/access.2017.2706947.
5. Vicen¸c Torra and Guillermo Navarro-Arribas, "Big Data Privacy and Anonymization," IFIP Advances in information and communication Technologies, 2017.
6. V. Ciriani, S. D. C. di Vimercati, S. Foresti and P. Samarati, "K-Anonymity. In Security in decentralized data management," Springer-Verlag, 2007.
7. P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," SRI International, 1998
8. Khaled El Emam and Fida Kamal Dankar, "Protecting Privacy Using k-Anonymity," Journal of the American Medical Informatics Association, Volume 15 Number 5, September / October 2008.
9. Jun-Lin Lin and Meng-Cheng Wei, "An Efficient Clustering Method For K-Anonymization," 2008.
10. G.Chitra Ganabathi and P.Uma Maheswari, "Privacy Preserving K-Anonymization Clustering Approach For Reducing Information Loss," Asian Journal of Information Technology, 2016, 15 (10) : 1531-1538.
11. J. W. Byun, A. Kamra, E. Bertino and N. Li, "Efficient k-anonymization using clustering techniques," International Conference on Database Systems for Advanced Applications, (pp. 188-200), Springer, Berlin, Heidelberg, 2007.
12. P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in Proc. 17th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems, 1998, p. 188.
13. Ankit Saroha, "Survey of k-Anonymity", National Institute of Technology, Rourkela, March 2014.
14. R C. Wong, Li J, Fu A W, et a1, "(α, k)-Anonymity: an enhanced k-anonymity model for privacy-preserving datapublishing," Proceedings of the 12th ACM SIGKDD, New York: ACM Press, 2006, pp. 754-759.
15. K. LeFevre et al., "Incognito: Efficient full-domain k-anonymity," in Proc. 2005 ACM SIGMOD Int. Conf. on Management of Data, 2005, pp. 49–60.
16. A. Machanavajjhala et al., "L-diversity: Privacy beyond k-anonymity," ACM Trans. on Knowledge Discovery from Data, vol. 1, no. 1, 2007.
17. N. Li et al., "t-closeness: Privacy beyond k-anonymity and l-diversity," in 23rd Int. Conf. on Data Engineering, 2007, pp. 106–115.
18. Kishore Verma Samraj, Rajesh Appusamy and Ramya Ravi Shankar, "Utility Enhancement of Deficient Relational Recordset Anonymization," International Journal of Intelligent Engineering and Systems, 2018.
19. S. Kishore Verma, A. Rajesh and J.S. Adeline Johnsana, "A Systematic Evaluated Recommendation on Performance Enhancement Factors and Procedures of Relational Data Anonymization," International Journal of Pure and Applied Mathematics, 2018, Volume 120 No. 5 2018, 1175-1188.]. *21(3).* pp. 876—880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar