IOSR Journal of Engineering (IOSRJEN) www.iosrjen.org
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 11 (November. 2018), ||V (III) || PP 14-25

CARD-Utility Guided Clustered Anonymization of Relational Data with Minimum Information Loss and Optimal Re-Identification Risk

Kishore Verma S[1]Rajesh A[2]Adeline Johnsana J S[3]

[1]*Research Scholar, Department of Computer Science and Engineering, Sri Chandra-sekharendraSaraswathiViswaMahavidyalaya University, Kanchipuram, Tamilnadu, India.*

[2] *Principal, C.Abdul Hakeem College of Engineering and Technology, Melvisharam, Tamilnadu, India.*

[3]*Research Scholar, Department of Computer Science and Engineering, St.Peter's Institute of Higher Education and Research, Avadi, Tamilnadu, India.*

Corresponding Author: Kishore Verma S

Abstract:Accumulation of relational data about the individual are exponentially grooming to support various organisations in maintaining financial records, manufacturing and logistical information and personnel data. These organisations are required to publish their data for analytical purposes. However this may cause privacy breach, if an invader deeds to link the possibly identifiable information to the records present in the published databases. Many data anonymization procedures that averts this threat by altering relational data before release have proposed in recent years, however these procedures suffers from information loss and utility loss. To address this issue, we propose a new utility guided clustered anonymization framework to anonymize relational data with high utility and less information loss. Based on this frame work, we developed three approaches varies on clustering process adopted for Record oriented Anonymization (RoA), which explores large solution space than existing methods and satisfies an extensive variety of privacy necessities. Furthermore, a classification analysis is performed on the anonymized datasets to ensure utility necessities. Experimental results on the benchmark data sets confirm that our framework expressively outperforms from 18% to 60% in terms of information loss, re-identification risk and classification accuracy than the existing procedures.

Key Terms*:Clustered Anonymization, Relational data, k-anonymization, Classification, Privacy and Utility.*

Date of Submission: 11-11-2018 Date of acceptance: 23-11-2018

I. INTRODUCTION

For every database, particularly where health information of individuals are accumulated with the aid of hospitals or government sectors, anonymity has a widespread role in protecting the privacy of the individual records when being linked to publicly available data. In business data bases, where corporations would really like to reveal an individual statistics to third parties (e.g. Outside agencies), anonymity could be used to guard the privacy of the people as in such instances a person's privacy may not be treasured. Thus inside the corporations, people's statistics need to be confined in phrases of access and anonymous, via getting rid of all data which can at once link records to persons by means of generalization or suppression formerly revealing in way that privacy is not broken. This procedure is denoted as data anonymization.

A significant approach addressing relational data privacy is k-anonymization [1,2,3,4], in this method data privacy is ensured in such a way that every record in the published record set is indistinguishable from at least (k-1) other records. K anonymizing of the quasi identifiers are done with respect to the each attributes generalization hierarchy as shown in Figure 1 to Figure 4. For instance Table 1 represents the patient details retrieve from the hospital to do some medical research. The micro data is free from identity details, however if it is exposed to the linkage attack by intruder i.e. the hospital data set is subjected to be intersected with publicly available voters list as shown in Table 2. Some of the records may seems to be exposed as a result possessing common quasi identifier (highlighted in table 2). K anonymization effectively protects the dataset from these type of attacks.By means making each record can be seen only in at least k records i.e. if the intruder want to locate a particular record in the data set, he can do this only on at least k records as shown in Table 3. Though k-anonymization have gained creditable privacy protection on relational data, there is a severe issue, that the

anonymized data are not much useful for the data mining tasks. This would have great impact on the effectiveness of data mining results. To make sure the performance of the data mining process, utility need to be taken into account during anonymization. Usually all of the anonymization procedures takes information loss as the straight forward measure to assess the utility. Generally, anonymity procedure that possess less information loss means to support good utility on data mining tasks.In order to gain minimum information loss for k anonymization, we have formulated the problem as CARD (Clustered Anonymization of Relational Data). According to CARD framework, the given records are grouped as clusters, in such a way that the records within the cluster are closer to each other (in terms of minimum distance between the attributes) and farer with records of other clusters. And these formed clusters of records are anonymized distinctly and published as single anonymized data set. Since the records with each cluster possess minimum distance on the values of the quasi identifiers, which is the significant factor that makes each cluster's anonymization to have less generalised value and reduces the information loss. For instance, the original dataset shown in Table 1 is clustered with cluster count =3, where the records with small distances with respect to their quasi identifiers are placed under same clusters. Anonymization and publication results in where the records are anonymized with less generalized values (as shown in Table 4 and Table 5) when compared to the conventional k anonymization (shown in Table 3). This motivated our work to get developed and have productive results.

Table 1 Hospital Patient details

Pincode	Age	Gender	Education	Disease	Expense
632401	25	Male	Bachelors	Flu	2000
635104	52	Male	Masters	Cancer	10000
600050	36	Female	Assoc-acdm	HIV+	15000
625818	41	Male	9^{th}	Diabetes	3000
627108	65	Female	12^{th}	Diabetes	3500
632402	28	Female	Some College	Flu	1800
635517	57	Male	Doctorate	Pneumonia	4200
625051	48	Male	12^{th}	Cancer	9800
600010	33	Male	Prof School	Flu	1600
627812	61	Female	10^{th}	Pneumonia	4700

Table 2 Voter list database

Name	Age	Gender	Pincode
Joe	25	Male	632401
John	27	Male	635103
Sara	41	Female	625819
Rani	35	Female	627109
Fathima	28	Female	632402
Swapna	36	Female	632402
Venus	53	Female	625051
Antony	33	Male	600010
Naveen	61	Female	600125

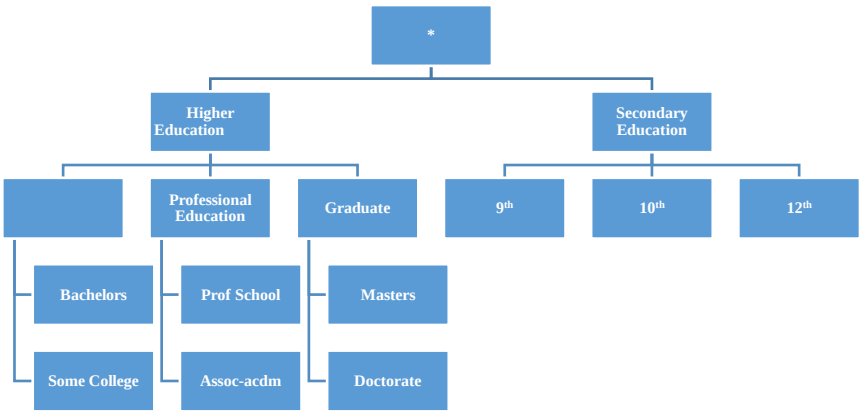

Figure 1 Generalization Hierarchy for Education attribute

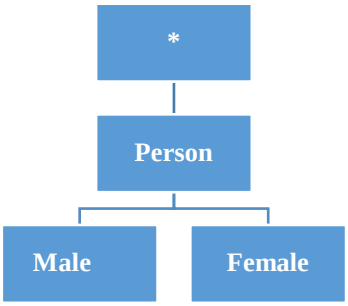

Figure 2 Generalization Hierarchy for Sex Attribute

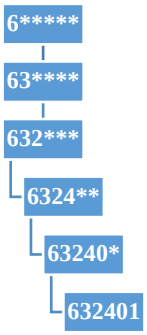

Figure 3 Generalization Hierarchy for Pin code Attribute

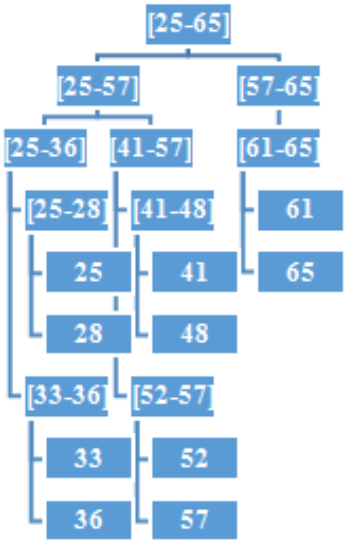

Figure 4 Generalization Hierarchy for Age Attribute

Table 3 K-anonymous Hospital Patient details (k=2)

Pincode	Age	Gender	Education	Disease	Expense
63****	25-52	Male	Higher Education	Flu	2000
63****	25-52	Male	Higher Education	Cancer	10000
6*****	36-41	Female	*	HIV+	15000
6*****	36-41	Male	*	Diabetes	3000
6*****	28-65	Female	*	Diabetes	3500
6*****	28-65	Female	*	Flu	1800
6*****	48-57	Male	*	Pneumonia	4200
6*****	48-57	Male	*	Cancer	9800
6*****	33-61	Male	*	Flu	1600
6*****	33-61	Female	*	Pneumonia	4700

Table 4 Clustered dataset

Pincode	Age	Gender	Education	Disease	Expense	Cluster
632401	25	Male	Bachelors	Flu	2000	Cluster0
632402	28	Female	Some College	Flu	1800	Cluster0
600010	33	Male	Prof School	Flu	1600	Cluster0
600050	36	Female	Assoc-acdm	HIV+	15000	Cluster0
625818	41	Male	9^{th}	Diabetes	3000	Cluster1
625051	48	Male	12^{th}	Cancer	9800	Cluster1

635104	52	Male	Masters	Cancer	10000	Cluster1
635517	57	Male	Doctorate	Pneumonia	4200	Cluster1
627812	61	Female	10^{th}	Pneumonia	4700	Cluster2
627108	65	Female	12^{th}	Diabetes	3500	Cluster2

Table 5 Clustered Anonymized Dataset

Pincode	**Age**	**Gender**	**Education**	**Disease**	**Expense**	**Cluster**
632***	25-28	Person	Undergraduate	Flu	2000	Cluster0
632***	25-28	Person	Undergraduate	Flu	1800	Cluster0
600***	33-36	Person	Professional Education	Flu	1600	Cluster0
600***	33-36	Person	Professional Education	HIV+	15000	Cluster0
625***	41-48	Person	High School	Diabetes	3000	Cluster1
625***	41-48	Person	High School	Cancer	9800	Cluster1
635***	52-57	Person	Graduate	Cancer	10000	Cluster1
635***	52-27	Person	Graduate	Pneumonia	4200	Cluster1
627***	61-65	Person	High School	Pneumonia	4700	Cluster2
627***	61-65	Person	High School	Diabetes	3500	Cluster2

II. BASIC CONCEPTS AND SURVEY ON RELATED APPROACHES

This section discusses the concepts and methods that form as the fundamental for our work.

1.1 K anonymization

A table T is said to be k anonymized tableT^*, if the all the quasi identifiers of each records in the table T are generalized or suppressed up until each record is indistinguishable with at least k-1 other records in the table T^*

1.2 Utility guided K-anonymization

A table T is said to be utility guided k anonymized tableT^{**}, where all the quasi identifiers are subjected to k anonymization in such a way to possess minimum null value count and transformation pattern loss.

1.3 Clustered Anonymization

Clustered anonymization of relational data problem is defined to find a solution that creates a clusters (C_1, C_2…, C_n) from a given relational data records R, in a such a way that every clusters contains at least ($k \leq n$) records which are closer enough with all other records of the same clusters and farther from the records of other clusters. These clusters C_1, C_2…, $C_n$are anonymized, that each and every records of the clusters are indistinguishable from at least k-1 other records abiding utility factors possessing minimum transformation loss and null value count.

1.4 Literature survey

In this section we discuss various data anonymization procedures that are proposed in past few years by adopting clustering strategy as key tactics to reduce information loss and increase data utility. All the procedures proposed in recent literatures attempted to show their effectiveness in delivering reduced information loss by employing novelty in these aspects i) methodology ii) metrics ii) different type of data like transaction data, social network data, image data set, sequence data set and time series data and iii) Utility analysis. [5] authors proposed a anonymization procedures which works similar to clustering process. Here the records are grouped as clusters with respect to the distance calculation of quasi identifiers and the quasi identifiers of the clusters are generalized with the computed centroid value, however the calculated centroid value for all attributes will not possess the righteous centroid value with respect to domain hierarchy.[6] presented a method that creates clusters on the given data set based on the quasi identifiers distance factor and the clusters are anonymized individually. This method possess good impact with respect to information loss reduction but not at the convincing level for run time complexity which is about$O(n^2)$. The authors of [7] viewed clustered anonymization in different perspective and proposed two metrics tuple diversity and attribute diversity to compute information loss measure through which utility of the data is analysed to be enhanced.[8] proposed method that clusters and anonymize the record set simultaneously through which they attained good

improvement in the run time as $O(n^2/k)$, however simultaneous creation of clusters will give way to outlier formation.[9] projected a method that clusters the records twice i.e. k member and one pass k means ,based on integrity of the associated attributes to generate reduced information loss but suffers from increased execution time [10] proposed systematic clustering method to anonymize the data , where the records are initially sorted and the sorted total number of records are divided into partitions and each partition's records values adjusted in such a way that the attribute values of the records that are closer to each other is considered. This method reduces the information loss and capable of capturing extreme values of the records in cluster formation. However this method suffers in time spent on initial sorting. [11] proposed genetic algorithm for gene type of data sets, which clusters and anonymize the records efficiently.[12] presented method to improvise the anonymization process by means of including clustering in the framework. Micro aggregation has been used as the core anonymization process, where the records are formed with respect to centroid value computed randomly on the partitions. A bottom up clustered anonymization procedure [13] were employed to improve the cluster quality in delivering optimal trade-off, the methods are applicable to both numeric and categorical attributes. But this similarity based clustering is sensitive to outliers of the dataset.Some of the methods extend their perception in handling different dimensional data, as done in [14], the authors proposed clustering based anonymization for multidimensional data. Here the authors have eliminated the suppression factor to gain better results in terms of information loss. In this approach the clusters are formed using centroid value and information loss value is calculated with respect to each records in the cluster. Though this method attains high data utility but privacy remains trivial one because of suppression factor elimination. The authors of [15] performed clustering based anonymization in both directive options i.e. vertical and horizontal partitioning and formed clusters yields minimum information loss. [16]proposed agglomerative clustering based anoymization framework for transaction dataset and attained reduced information loss on transaction data anonymization. The authors of [17] applied the same agglomerative clustering for transaction data but in different perspective i.e. rather than applying the clustering on the micro data to form the clusters, they have applied the clustering process over the generalization hierarchy to build the clusters. [18] proposed a clustered anonymization method to the special type of data that possess of both relational and transaction data characteristics and succeed in attaining minimum information loss. [19] presented clustering based anonymization method that are applicable to internet of things(IOT) data under distributed environment, well enough in preventing the data's from similarity and probabilistic attacks with reduced information loss. The supportability of the technique in terms of data mining task is analysed by executing different classifiers on the anonymized data set and accounted the assessment in terms of accuracy and f-measure of the classifiers. The security and openness EGO data in IOT are effectively handled by [20] by means of fuzzy clustering based anonymization.[21] proposed a method to anonymize the graphical structure of the nodes in social networks and attained minimum information loss for the reason that anonymization is done through clustering.The authors of proposed k-means clustering based anonymization for social networks that are applicable on raw social network data. Then there arise some of the procedures like [22, 23] that anonymizes spatial and image data through clustering to increase the data utility in spatial data mining and image processing. The other types of data like sequence and times series data are also subjected to cluster based anonymization by the recent well known procedures like [24,25]. In [24] the author have employed agglomerative clustered anonymization for multi-dimensional sequence data. The authors [25] have extended their method of time series anonymization in finding out the optimum privacy level of anonymization with respect to re- identification risk. Later [26] proposed cluster based anonymization which is capable enough in handling mixed data i.e. separate information loss measure have been incorporated to process numeric, categorical and structural data with reduced information loss. From experimental section of [27, 28] the importance of the utility analysis and attribute correlation that should persists for every data anonymization process is inferred and realized. [29] proposed a method that delivers good trade-off between privacy and utility, the factor of trade-off is measured in term of re-identification risk.[30,31] proposed methods that anonymizes the given data set in cell-level, attribute level and record level and privacy/utility analysis are done with respect to re-identification risk, this had inspired us in proposing a cluster based anonymization analysis for cell-level, attribute level and record level anonymizations by incorporating reidientification risk as one of the key factor.

III. CARD- THE METHODOLOGICAL EXPLORATION

3.1 Architecture of CARD

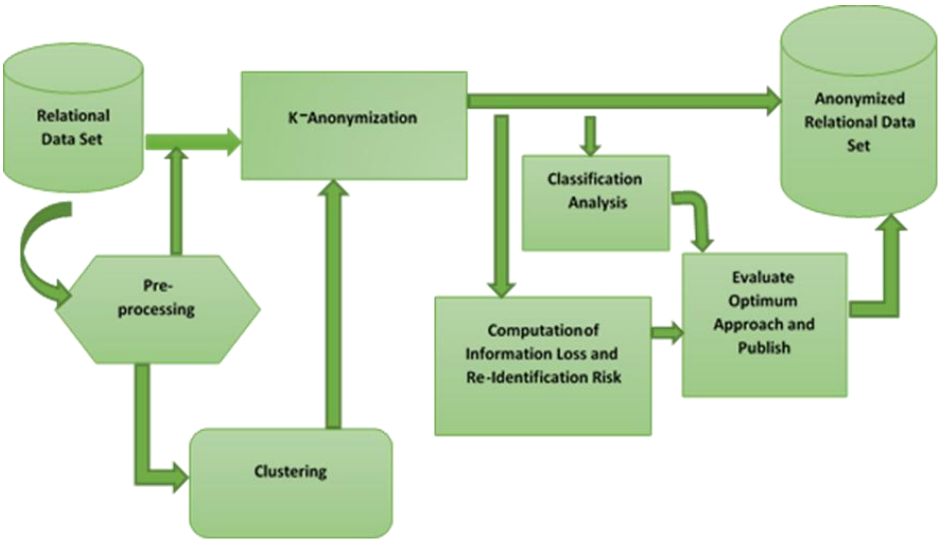

Figure 2 Architecture Diagram of CARD

3.2 Components of CARD

3.2.1 Pre-processing

This part of the work pre-process the given data/record set according to frame that can be applicable for further processing. Here the data cleansing is performed i.e. the records that consists of null values are processed as in [32]. Since null values present in the record set may mislead the clustering and anonymization process. In clustering process the output will be subjected to more number of outliers whereas in anonymization, can have more information loss. Thus to have proper utility the records containing null values are incredibly processed as an activity of pre-processing.

3.2.2 Clustering

The pre-processed data set are clustered using three well known clustering methodology i.e. k-means, farthest first and expectation maximization clustering implemented in most of the approaches in the literature. Then the each cluster is subjected to utility guided anonymization individually. The procedural explanation of clustering algorithms used in this work are mentioned below.

K- Means Clustering:

This method is a well-known widely used clustering that attempts to find random centroids for k clusters from the given record set. Then the records are assigned to each clusters in a way they possess minimum mean with the centroids of each clusters. On each iteration the centroids are adjusted in order make clusters possessing minimum intra cluster distance. This method seems to be NP-hard problem, since the initial guess on choosing the number of clusters is trivial one. This may lead to have bad cluster formation. Though it is fast and simple enough in generating clusters, the advisable factor is to avoid initial guesses. The only best way in generating optimal number of clusters is to run the algorithm with numerous random initial guesses on the number of clusters and choose the k value that are able to possess least variance i.e. at a specific point of execution, there will be slow decrease in the variance values. These value of k seems to generate best number of clusters.

Farthest First Clustering:

This method is variant of k-means clustering, it picks cluster randomly from the given data set/ record set in such a way that each of the cluster centres are farther among themselves and must be within the data/ record space. Then assigns each clusters with records that are closet to the chosen cluster centers. This method viewed to be the greedy approximation algorithm or greedy permutation. It requires only less adjustments comparatively with k-means during cluster assignment and able to perform the execution in polynomial time even for large and multidimensional data/record sets. Farthest First methodology attempts to give heuristic

solutions grounded on two factors i) pigeon hole principle i.e. two records with optimal solution must both be within r- distance of the same point from k-1 records in the cluster and ii) triangle inequality within 2r of each other. These factors of farthest first methodology have made it to work robust in terms of cluster quality and run time complexity.

Expectation- Maximization Clustering:

This method is predominant unsupervised clustering extension of k means methods, which does not requires training phase. It is a repetitive methodology, which attempts to find highest possibility of attributes in the record set. It is well exposed enough in handling the hypothetical constructs of the data set / record set. This EM clustering continues the repetitions in two steps i) E –Step, where a log-possibility are computed using the current calculation of the parameters i.e. from the obtained values of the attributes in the record set. ii) M- Step , in this step computations are done by increasing the expected log- possibility of the parameters that are found in E-step. It is better in creating non-overlapping clusters, the records of one clusters do not intersect with clusters of other cluster. Thus this may lead to have good cluster formation. Here the number of cluster that need to be created are analysed through Bayesian information criterion (bic), which chooses the right cluster count by varying it from 1 -9 and led to have increased cluster quality.

3.2.3 Utility Guided K-Anonymization

Here in this part of the work, each individual clusters drawn from the clustering process are K-anonymized with utility guided properties. Utility guided K-anonymization is the process of anonymizing the given record set R to R^* according to k- anonymization strategy, where for every K-value execution of R toR^*, R^* seems to hold minimum factors i) Number of null Values(nV_C) and ii) Transformation pattern Loss $(T.p.L)$[32]. Thus this make the anonymization process to possess minimum information loss and yields maximum data utility. Accordingly as done in [32], we have adopted Record Oriented Anonymization's discernibility data quality model to anonymize the clusters, which are received from different clustering process and also compared the effectiveness of the proposed work with and without clustered anonymization.

3.2.4 Computation of Information Loss and Re-Identification Risk

This part of the work computes the information loss and Re-Identification Risk incurred on account of k anonymizationfor with and without clustered data sets. Especially for clustered data set the information loss and Re-Identification Risk calculation are done for each individual clusters and summed up. These two factors are meant with tool implementation and can be referred on necessity [33].

3.2.5 Classification Analysis

This part of the work, a classification analysis is done on the clustered and non-clustered anonymized data to gain the utility effectiveness of our proposed CARD approach. Thus on part of the execution three widely used classifiers are chosen from the literature i) Logistic regression ii) Bayes and iii) Random forest. The classification process is analysed in terms of classification accuracy of the input data set and output data set, which is the novel part of analyses adopted in this work.

3.2.6 Evaluate Optimum Approach and Publish

Here the RoA's best data quality model with respect clustered anonymization are analysed in terms of information loss difference between clustered and non-clustered anonymization and also the Re-Identification Risk and Classification Accuracy. Based on this analysis the best method of clustering and RoAanonymization are identified and the given data set is clustered, anonymized and published accordingly.

IV. EXPERIMENTAL RESULTS

The significant objective of experiment is to investigate the performance of our CARD in terms of data privacy and data utility. To accurately prove the efficiency of our approach, CARD implementation is compared with the existing K- Anonymization approach.

4.1 Executional Platform

To enrich proposal we used two well-known toolsi) WEKA and ii) ARX Anonymization .Three clustering procedures are implemented and executed in WEKA.K-anonymization is executed by widely used open source anonymization tool ARX available in [33]. The experiments were executed on machine running 64-bit windows 8.1, Intel core i5 processor with 8GB RAM.

4.2 Experimental Setup

In this experimental assessment, UCI Machine learning repository -Adult data set is used. This dataset consists of 30162 records with 9 attributes. According to our k-anonymizationstratergy the dataset are categorised 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 specific individual as account of linking attacks. (iii) Ia-Identifying attributes are the direct signifiers of the records i.e. explicitly reveals the identity of the individual. Each attributes requires special consideration in k-anonymization process .Qid's need to generalized or suppressed to support k-anonymization, Sa's need to protected from correlating with Qid's and Ia's need to be eliminated from publishing. Here in our experimentation from 9 attributes, first eight attributes are taken as Qid's, last attribute is considered as Sa.The data set is clustered with three different clustering approaches K-Means, Farthest First and Expectation Maximization and anonymized with utility guided recordorientedanonymization approach – Discernibility. The clustered data set is anonymized with varying k values as 5, 10,20,25,30 and 35 etc.

4.3 Results

4.3.1 Privacy and Risk Analysis

In this part of the experimentation, it is attempted to prove that our proposed framework CARD(Clustered Anonymization of Relational Data) is well enough in generating minimum information loss and Re-Identification Risk than that of non-clustered anonymization. Privacy is measured in terms of pertaining minimum Re-Identification Risk i.e. the CARD approach that possess minimum Re-Identification Risk is mean to be highly protected. Whereas Utility is measured in terms of pertaining minimum Information Loss i.e. CARD approach that possess minimum information loss. The results obtained for information loss and Re-Identification Risk pertaining to different K values are discussed in following sections.

4.3.1.1 Comparative Analysis of Information Loss and K-Value

The information loss obtained on applying K-Anonymization on the three kinds of CARD for different values of K are representedandalso the proposed CARD approach is compared with the non- clustered anonymization is shown in Figure 6. From the figure it clearly inferable, all the three CARD approaches are generating highly distinguishable minimum information loss when compared with non-clustered anonymization. By analysing the CARD's three approaches CARD-Expectation Maximization seems to perform the best by generating lowest information loss than the other two approaches of CARD. By analysing the results it is inferred that our approach can able to attain 38% to 55% reduction rate in information loss.

Typically information loss pertaining to anonymization has a great impact in the data mining utility on a particular data set. Our approach proves to be the optimum one delivering the lowest information loss and facilitates the data mining process with greater utility.

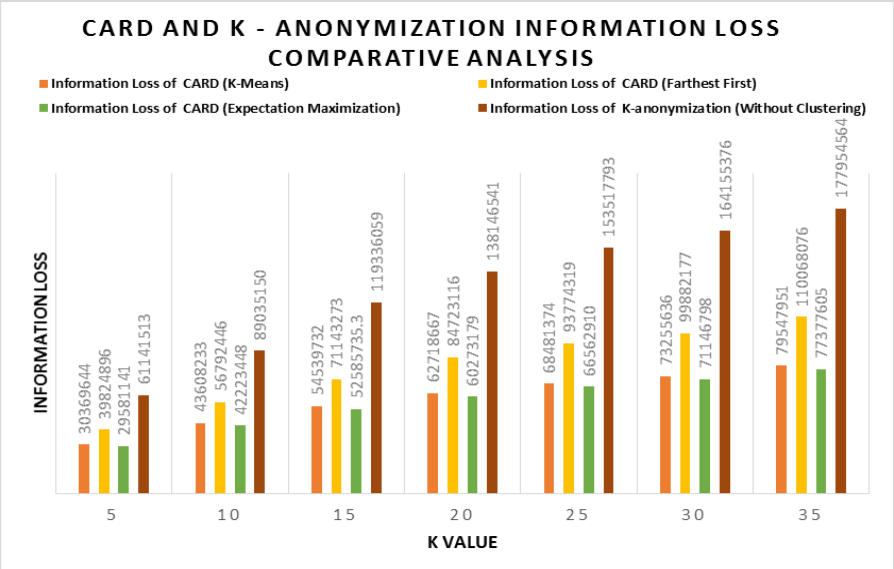

Figure 3 CARD and K-Anonymization Information Loss Analysis

4.3.1.2 Comparative Analysis of Re-Identification Risk and K-Value

The Re-Identification Risk is the factor that enables us to know the uncertainty pertaining with anonymized data set. Since most of the experimental results only concentrates in delivering the good privacy with greater utility where they concentrate on measuring the information loss. Whereas our approach

distinguishes in measuring the re-identification risk on the anonymized and accounting this factor in deciding the optimum one. From the Figure 7 pictures CARD approaches presents minimum re-identifaction risk of about 12% to 60% than nonclusteredanonymization approach. All the three approaches seems to be better in terms of Re-Identification Risk, varies in close proximity, anyhow among three approaches CARD(K-means) seems to be the optimum.

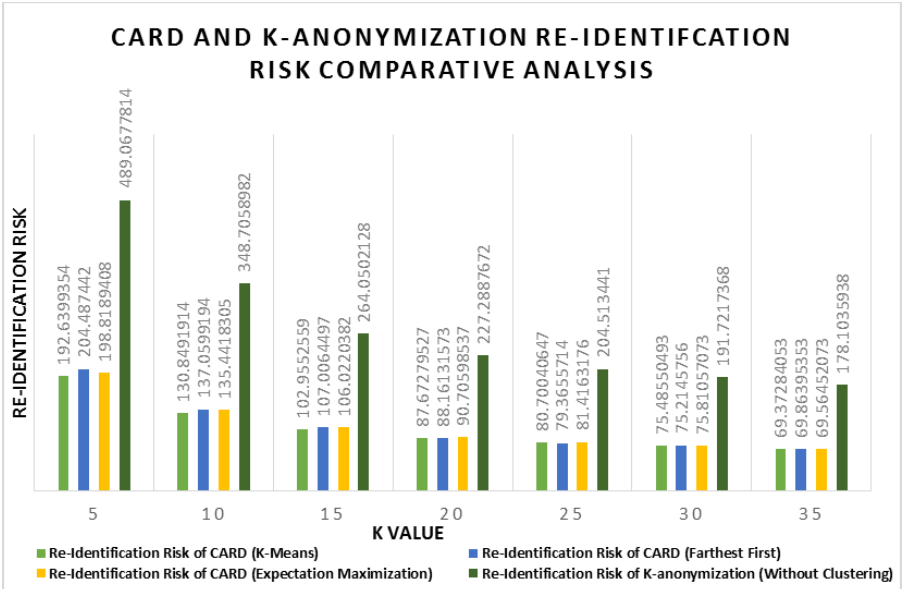

Figure 4 CARD and K-Anonymization Re-Identification Risk Analysis

4.3.2 Utility Analysis

In this part of the experimentation, it is attempted to prove that our proposed framework CARD (Clustered Anonymization of Relational Data) is well enough in generating better classification accuracythan that of non-clustered anonymization.The anonymized dataset of CARD's approach is subjected to three classifiers i.e. Logistic Regression, Naïve Bayes and Random Forest. Utility is measured in terms of pertaining Higher Classification Accuracy i.e. the CARD approach possessing higher Classification Accuracy is mean to be extremely utilizable. The result obtained for CARD's three approaches, Classification Accuracy pertaining to different K values are discussed in following sections.

4.3.2.1 CARD(K-Means) Classification Analysis

Figure 8 shows the classification accuracy's comparative analysis of three classifiers Logistic regression, Naïve Bayes and Random forest that are applied on CARD's k-means K-anonymized dataset. From the obtained results it evidently provable that non-clustered K-anonymized data set holds less classification accuracy than that of the CARD's k-means approach. On part of CARD's three classification analysis, Logistic Regression is capable of generating higher accuracy than that of other approaches.

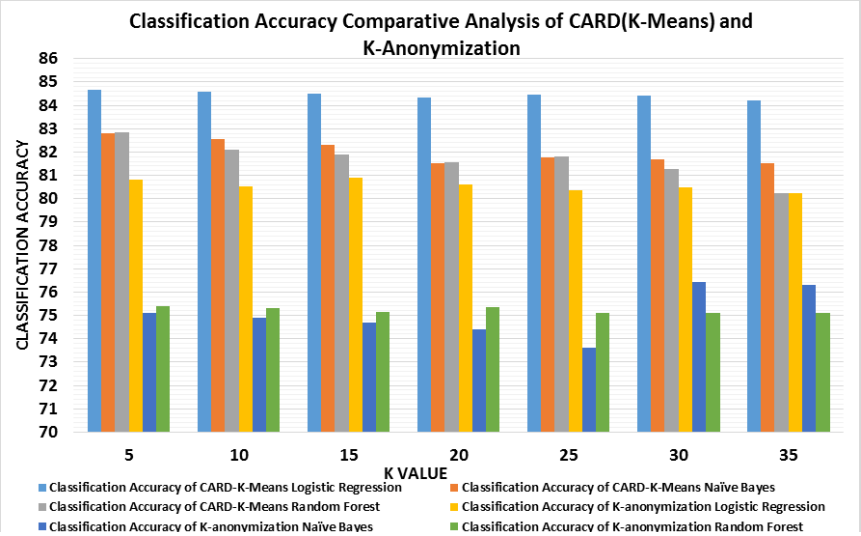

Figure 5 CARD (K-Means) Classification Accuracy Analysis

4.3.2.2 CARD(Farthest First) Classification Analysis

Figure 9 shows the classification accuracy's comparative analysis of three classifiers Logistic regression, Naïve Bayes and Random forest that are applied on CARD's Farthest First and non- clustered K-anonymized dataset. From the obtained results it evidently provable that non-clustered K-anonymized data set holds less classification accuracy than that of the CARD's k-means approach. On part of CARD's three classification analysis, Logistic Regression is capable of generating higher accuracy than that of other approaches and produces best classification accuracy even on non-clustered K-anonymization data set. Thus Naïve Bayes and Random forest classifiers built on our CARD anonymized data set is producing less accuracy when compared to the logistic regression classifier built on non-clustered k-anonymized data set.

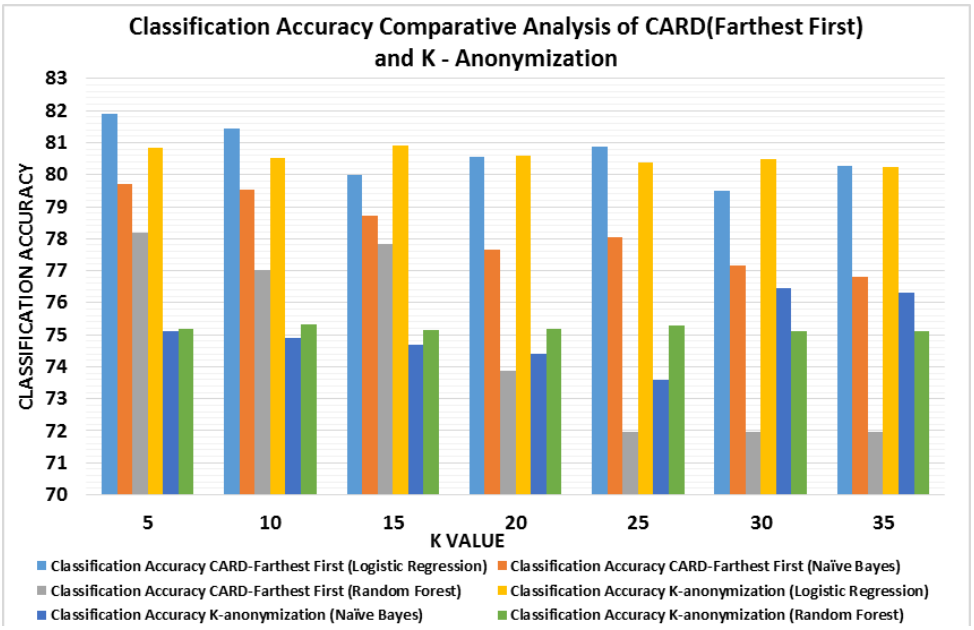

Figure 6 CARD (Farthest First) Classification Accuracy Analysis

4.3.2.3 CARD(Expectation Maximization) Classification Analysis

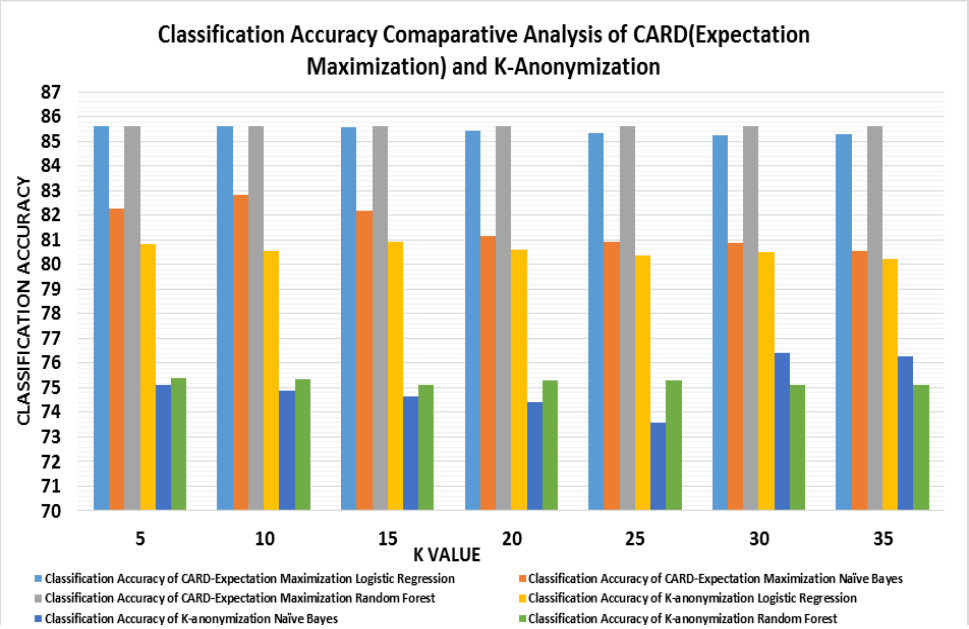

Figure 7 CARD (Expectation Maximization) Classification Accuracy Analysis

Figure 10 shows the classification accuracy's comparative analysis of three classifiers Logistic regression, Naïve Bayes and Random forest that are applied on CARD's Expectation Maximization and K-anonymized dataset. From the obtained results it evidently provable that non-clustered K-anonymized data set holds less classification accuracy than that of the CARD's Expectation Maximization approach. On part of CARD's three classification analysis, Logistic Regression and Expectation Maximization is capable of generating higher accuracy than that of other approaches. Logistic Regression classifier produces optimum classification accuracy i.e. 80% approximately even on non-clustered K-anonymization data set.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an efficient framework of k-anonymization for relational data, which anonymizes the dataset with minimum information loss by having greater utility. Our approach CARD includes the theme of clustering the record set before anonymizing and we refer this as CARD approach. The fundamentals that supports our proposal were discussed with sufficient properties and examples. We have implemented CARD framework with three different clustering approaches that mean to be strong proof of effectiveness of our proposed theme. Any of the Anonymization procedure concentrates much on privacy and utility, our results confirms that this CARD is well efficient in generating good anonymized data set with 38% to 50% reduction rate in information loss and at the same time maintains the re-identification risk % to be less ranging from 13% to 60% than non-clustered anonymization and 2% to % 4 % increased rate in classification accuracy. These CARD approaches works effectively for utility guided Record Oriented Anonymization RoA, whereas still being a trivial for other approaches like Cell Oriented Anonymizations and Attribute oriented Anonymization. This would be our future aspect working towards right clustering approach by employing other kind of clustering that clusters the record set effectivelyandgenerates minimum information loss.

REFERENCES

[1]. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

[2]. LeFevre, K., DeWitt, D. J., &Ramakrishnan, R. (2006, April). Mondrian multidimensional k-anonymity. In Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on (pp. 25-25). IEEE

[3]. Ciriani, V., di Vimercati, S. D. C., Foresti, S., &Samarati, P. (2007). K-Anonymity. In Security in decentralized data management. Springer-Verlag.

[4]. El Emam, K., &Dankar, F. K. (2008). Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association, 15(5), 627-637.

[5]. Loukides, G., & Shao, J. (2007, March). Capturing data usefulness and privacy protection in k-anonymisation. In Proceedings of the 2007 ACM symposium on Applied computing (pp. 370-374). ACM

[6]. Byun, J. W., Kamra, A., Bertino, E., & Li, N. (2007, April). Efficient k-anonymization using clustering techniques. In *International Conference on Database Systems for Advanced Applications* (pp. 188-200). Springer, Berlin, Heidelberg.
[7]. Loukides, G., & Shao, J. (2008, March). Data utility and privacy protection trade-off in k-anonymisation. In Proceedings of the 2008 international workshop on Privacy and anonymity in information society (pp. 36-45). ACM.
[8]. Lin, J. L., & Wei, M. C. (2008, March). An efficient clustering method for k-anonymization. In *Proceedings of the 2008 international workshop on Privacy and anonymity in information society* (pp. 46-50). ACM.
[9]. Lin, J. L., Wei, M. C., Li, C. W., & Hsieh, K. C. (2008, December). A hybrid method for k-anonymization. In Asia-Pacific Services Computing Conference, 2008. APSCC'08. IEEE (pp. 385-390). IEEE.
[10]. Kabir, M. E., Wang, H., &Bertino, E. (2011). Efficient systematic clustering method for k-anonymization. ActaInformatica, 48(1), 51-66.
[11]. Lin, J. L., & Wei, M. C. (2009). Genetic algorithm-based clustering approach for k-anonymization. Expert Systems with Applications, 36(6), 9784-9792.
[12]. Thaeter, F., &Reischuk, R. (2018). Improving Anonymization Clustering. SICHERHEIT 2018.
[13]. Aghdam, M. R. S., &Sonehara, N. (2016). Achieving high data utility K-anonymization using similarity-based clustering model. IEICE TRANSACTIONS on Information and Systems, 99(8), 2069-2078.
[14]. Pramanik, M. I., Lau, R. Y., & Zhang, W. (2016, November). K-anonymity through the enhanced clustering method. In e-Business Engineering (ICEBE), 2016 IEEE 13th International Conference on (pp. 85-91). IEEE.
[15]. ShyamalaSusan,V., Christopher, T.,(2017). An Efficient Anonymization Model (EAM) For Data Publishing Using Optimized Clustering Approach,International Journal of Pure and Applied Mathematics,118(19),2743-2459.
[16]. Gkoulalas-Divanis, A., &Loukides, G. (2011, March). PCTA: privacy-constrained clustering-based transaction data anonymization. In Proceedings of the 4th International Workshop on Privacy and Anonymity in the Information Society (p. 5). ACM.
[17]. Gkoulalas-Divanis, A., &Loukides, G. (2012). Utility-guided Clustering-based Transaction Data Anonymization. Trans. Data Privacy, 5(1), 223-251.
[18]. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., &Skiadopoulos, S. (2013, September). Anonymizing data with relational and transaction attributes. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 353-369). Springer, Berlin.
[19]. Nayahi, J. J. V., &Kavitha, V. (2017). Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop. Future Generation Computer Systems, 74, 393-408.
[20]. Xie, M., Huang, M., Bai, Y., & Hu, Z. (2017). The anonymization protection algorithm based on fuzzy clustering for the ego of data in the Internet of Things. Journal of Electrical and Computer Engineering, 2017.
[21]. Kaveri, V. V., &Maheswari, V. (2015). Cluster Based Anonymization for Privacy Preservation in Social Network Data Community. Journal of Theoretical and Applied Information Technology, 73(2), 269-74.
[22]. Heidelberg. Abul, O., Bonchi, F., &Nanni, M. (2010). Anonymization of moving objects databases by clustering and perturbation. Information Systems, 35(8), 884-910.
[23]. Honda, K., Omori, M., Ubukata, S., &Notsu, A. (2016). Fuzzy clustering-based k-anonymization of eigen-face features for crowd movement analysis with privacy consideration. International Journal of Innovative Computing, Information and Control, 12(4), 1375-1384.
[24]. Sehatkar, M., &Matwin, S. (2014). Clustering-based Multidimensional Sequence Data Anonymization. In EDBT/ICDT Workshops (pp. 385-389).
[25]. Johnsana, J. A., Rajesh, A., &Verma, S. K. (2016). CATs-Clustered k-Anonymization of Time Series Data with Minimal Information Loss and Optimal Re-identification Risk. Indian Journal of Science and Technology, 9(47).
[26]. Xu, X., &Numao, M. (2015, December). An efficient generalized clustering method for achieving k-anonymization. In Computing and Networking (CANDAR), 2015 Third International Symposium on (pp. 499-502). IEEE.
[27]. Iyengar, V. S. (2002, July). Transforming data to satisfy privacy constraints. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 279-288). ACM.
[28]. Kim, J. J., & Winkler, W. E. (1995). Masking microdata files. In Proceedings of the Survey Research Methods Section, American Statistical Association.
[29]. Domingo-Ferrer, J., Mateo-Sanz, J. M., &Torra, V. (2001, May). Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In Pre-proceedings of ETK-NTTS (Vol. 2, pp. 807-826).

[30]. Ciriani, V., di Vimercati, S. D. C., Foresti, S., & K-Anonymity, P. S. O. (2007). In Springer US. Advances in Information Security.
[31]. El Emam, K., &Dankar, F. K. (2008). Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association, 15(5), 627-637.
[32]. Verma, S. K., Rajesh, A., &Johnsana, J. A. (2018). A Systematic Evaluated Recommendation on Performance Enhancement Factors and Procedures of Relational Data Anonymization, International Journal of Pure and Applied Mathematics, 120(5), 1175-1187.
[33]. https://arx.deidentifier.org/downloads/

Kishore Verma S. "CARD-Utility Guided Clustered Anonymization of Relational Data with Minimum Information Loss and Optimal Re-Identification Risk." IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 11, 2018, pp. 14-25.