International Journal of Engineering and Manufacturing Science.
ISSN 2249-3115 Volume 8, Number 1 (2018) pp. 23-32

http://www.ripublication.com

Organized Clustering Method for Privacy Preserving Data Publishing

Ashoka K[1*] and Poornima B[2]

[1]Assistant Professor, [2]Prof. and Head
Computer Science and Engineering Department,
BIET, Davangere, Karnataka, 577004, India.

**E-mail: ashoka_kkd1@yahoo.com, poornimateju@gmail.com*

Abstract: Privacy preservation is a substantial concern for the organizations that publish/share personal data for informal analysis. In this paper we present a clustering based k-anonymization technique to reduce the information loss while at the same time ensuring data utility. In privacy preserving data mining, anonymization based approaches have been used to preserve the privacy of an individual. But, the anonymization based approaches suffer from the issue of information loss. To minimize the information loss and ensure data quality we presented new approach called organized clustering along with equal combination of quasi-identifier and sensitive attributes. We also evaluate the proposed approach empirically focusing on the information loss and execution time as vital metrics.

Keywords: data privacy, sensitive attribute, sub-databases, systematic clustering, anonymization, PPDM.

1. Introduction

Now a days, the volumes of generated data increases exponentially every year. Among this data, there is a growing amount of personal information contained within. This sensitive data has attracted the attention of those fascinated in creating more tailored and personalized applications. This violates the privacy of the individuals and leads to the apprehensions that personal data may be breached and distorted. As a result, this phenomenon has brought new challenges to protect the privacy of the people as a key issue in privacy preserving data mining [1]. To resolve these privacy concern, several Privacy Preserving Data Publishing (PPDP) algorithms are

developed by researchers. Some of the well-accepted techniques are, the *k*-anonymity model [3, 8]. It has been considerably used in privacy preserving data mining because of its easiness and effectiveness. However, information loss and data utility are the major issues in the anonymization approaches as discussed in [2, 8]. The *k*-anonymity model provides privacy and generates an anonymous database via generalization and/or suppression. In the case of generalization, the values in a database are replaced with some related values. Though, the anonymous database generated via generalization and/or suppression bring about *information loss*. The *k*-anonymity model works by ensuring that each record of a table is indistinguishable to at least ($k - 1$) other records with respect to a set of privacy-related features, called quasi-identifiers, that could be potentially used to identify individuals by linking these attributes to externally available data sets [7]. For illustration, consider the patient diagnosis records in a hospital in Table-1, consists of three types of attributes. Identifier (ID), quasi-identifier (QI) and sensitive attributes (SA). The attributes *ZipCode*, *Age* and *Gender* are treated as quasi-identifiers, *Name* as identifiers and *Disease* as sensitive attributes. If the hospital simply publishes the table to other organizations for classifier development, those organizations might extract patient's disease histories by joining this table with other tables. By contrast, Table-2 is a 3-anonymization version of the data values of Table-1. Here the attributes *Gender* and *Age* have been generalized and *Zipcode* is suppressed as common values. To reduce the information loss due to *k*-anonymization, all records are partitioned into several groups such that each group has a minimum of *k* similar records with respect to the quasi-identifiers and then the records in each group are generalized or suppressed such that the values of each quasi-identifier are the same.

Such similar groups are known as clusters [7]. Therefore, the *k*-anonymity model can be addressed from the perspective of clustering.

Table-1: Original Patient Table

Identifier	**Quasi-Identifier**			**Sensitive Attribute**
Name	**Age**	**Gender**	**Zipcode**	**Disease**
Akul	31	M	55441	Cholera
Balaji	35	M	55440	Typhoid
Chitra	37	F	55442	Cancer
Divina	26	F	55440	Cancer
Emili	29	F	55440	Arthritis
Federer	23	M	55443	Flu

It is highlighted that the '*information loss'* and the '*data utility'* are two contradicting goals of privacy preserving data publishing. The *information loss* increases by hiding more data, but reduces the *data utility*. Contrarily, *information loss* decreases by hiding less data, but increases the *data utility*.

Table-2: A 3-Anonymous Table

Equivalent Class	**Age**	**Gender**	**Zipcode**	**Disease**
1	[31-40]	Person	5544*	Cholera
	[31-40]	Person	5544*	Typhoid
	[31-40]	Person	5544*	Cancer
2	[21-30]	Person	5544*	Cancer
	[21-30]	Person	5544*	Arthritis
	[21-30]	Person	5544*	Flu

2. Related Work

The problem with preserving the privacy of an individual when data mining has gained much importance in recent years and due to this several algorithms have been proposed [2, 8, 10]. In privacy preserving data mining, preserving the privacy of an individual has been a prime research issue. In order to preserve the privacy, various anonymization based approaches were proposed in the literature [2, 8, 10, 11]. The *k*-anonymity model [3] is one of the simple technique used for the privacy preservation. Extending the idea of k-anonymity, a number of anonymization based clustering approaches have been proposed in [4, 8, 10]. It includes Byun et al. Greedy k-member clustering algorithm [5], Loukides et al. Clustering algorithm [6], Lin et al. One passes *k*-means clustering algorithm [8] and Kabir et al. Systematic clustering algorithm [9]. Byun et al. [4] proposed a greedy *k*-member clustering algorithm. The greedy *k*-member clustering algorithm is sensitive to outlier records. With the presence of outlier records in the cluster, the *information loss* of the cluster is also increases. Loukides et al. [5] proposed a clustering algorithm, which produce one cluster at a time. This algorithm builds a cluster with a user defined threshold value. Based on the user defined threshold value, the records are inserted and deleted in a cluster. The *information loss* of the generated cluster should not exceed the user defined threshold value. If the number of records in a particular cluster is less than user defined threshold, the cluster is deleted. Thus, with the use of user defined threshold, this algorithm is less sensitive to outlier records. In addition, this algorithm deletes records, and therefore, generates huge *information loss*. Lin et al. [6] proposed

one pass *k*-means clustering algorithm. This algorithm builds a cluster with lesser *information loss* and execution time as compared with the greedy *k*-member clustering algorithm [4].

Kabir et al. presents a systematic clustering technique [7]. This algorithm has lesser information loss as compared to Byun et al. Greedy *k*-member clustering algorithm [4]. The systematic clustering algorithm makes a cluster of similar records. With the presence of similar kinds of records, it leads to the lesser generalization and/or suppression and hence results in minimum *information loss*.

3. Proposed Approach

In this section, we present organized clustering approach with equal combination of QI and SA using systematic clustering algorithm [7] that partition the database using a combination of quasi-identifier and sensitive attribute then produces the anonymized sub-databases by hiding the private information. Thus, publishing such database would preserve the sensitive information of a person when joining with the external available databases.

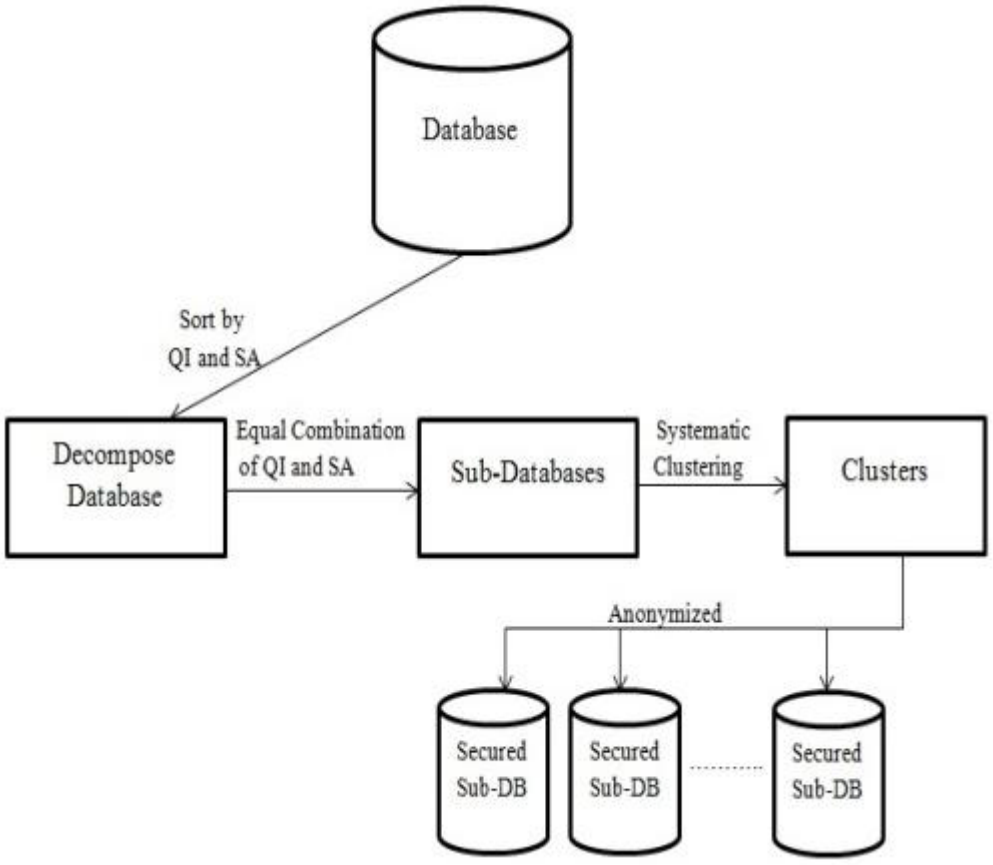

Figure-1*:* Proposed System Architecture

The architecture of the proposed system comprises of the initial database of any organizations of the distinct sectors such as Stores, Hospitals, and Banking, where privacy is needed for their customer information. These databases are decomposed by using equal combination of QI and SA in the database, then sub-databases is constructed. Later clustering and anonymization is performed on these sub-databases and lastly it is published.

Let B be a database with attributes viz. quasi-identifier (QI) and sensitive attribute (SA) as shown in Equation. (1).

$$D = \{QI, SA\} \quad (1)$$

Let QI and SA be the set of possible attribute. The possible values of quasi-identifier and sensitive attributes are represented in equations. (2) and (3).

$$QI = \{QI_1, QI_2, \ldots, QI_n\} \quad (2)$$

$$SA = \{SA_1, SA_2, \ldots, SA_n\} \quad (3)$$

Then, the resulting original database is shown in equation. (4).

$$D = \{QI_1, QI_2 \ldots, QI_n, \ldots, SA_1, SA_2, \ldots, SA_n\} \quad (4)$$

Let B_1 be a sub-database produsedfrom the database B. The sub-database B1 is a combination of quasi-identifier (QI) and sensitive attribute (SA). It is shown in equation. (5).

$$B_1 = \{QI_1, SA_1\} \quad (5)$$

In the similar way, other possible combinations $\{B_1, B_2, \ldots, B_n\}$ of QI and SA are constructed for the original database B.

We first identify and classify the attributes such as identifier, quasi-identifier and sensitive attributes in a database (step 1). Subsequently, we remove the identifier attribute from the database and sort all records using the quasi-identifiers (steps 2 and 3). Then, we find out the number of groups and clusters such that $\sigma=r/k$, where *r* is the number of records in a database and *k* is the anonymization factor (step 4). After identifying the groups and clusters in an original database, we generate a sub-database using a combination of quasi-identifier and sensitive attribute (step 5). In our Approach, we create an equal combination of quasi-identifier (QI) and sensitive attribute (SA). From each generated sub-databases, we make a partition of all records into *k* groups (step 6). Then, we used Systematic clustering algorithm in order to generate the clusters. According to the Systematic clustering algorithm, we randomly select a record from the first group for the creation of the first cluster (step 7). Similarly, we create the remaining cluster by randomly selecting the records from the remaining groups (step 8). Subsequently, we calculate the information loss of each

cluster (step 9). Now, we select other records from the first group and add records in a cluster whose information loss is the lowest (step 10).

Algorithm: Organized_Clustering ( )

Input: Database B with r records

Output: $\gamma = \{\sigma 1, \sigma 2, \ldots. \sigma p\}$ be a partitioning of r // B is original database

// r is the number of records in the database // γ is a partitioning of r records

// σ is a cluster

Begin

1. Identify the attributes such as identifier , quasi-identifier (QI) and sensitive attribute (SA)
2. Remove the identifier attribute and replace it with ID
3. Sort all records by their quasi-identifiers
4. Identify the number of clusters
5. Make an equal combination of QI and SA to construct the sub-database
6. Make a partition of all records into k groups
7. Select a record ri randomly from the first partition of k records
8. Similarly select another records rj from the other partition of k records
9. Calculate information loss $IL(\gamma) = IL(\sigma i)pi-1$
10. Move the records in a cluster with lowest information loss
11. Find extra element in a cluster those who exceed the k size
12. Add extra element in a cluster whose information loss is lowest

End

In the same way, we select and add other records in a cluster whose information loss is the lowest. In our proposed algorithms, we assume *r*, *k* and σ as the total number of records, the *k*-anonymity parameters and the number of clusters, respectively. The proposed algorithm takes O (nlogn) time to sort the records once in the database. The number of clusters are calculated as $\sigma = r/k$. As a result, the time complexity of our proposed clustering algorithms is a product of number of records (r) and the number of clusters (σ). Therefore, the total time complexity of our proposed algorithm is $O = (r2/k)$.

Information loss for Numerical attributes (Ln): Let the minimum and maximum records in a cluster σ be Ri max and Ri min respectively. Let Di max and Di min be the maximum and minimum values of the records in a database D.

$$Ln = (\sum Ri\ max - Ri\ min / Di\ max - Di\ min) \quad (\forall 1 \leq i \geq n)$$

Information loss for Categorical attributes (Lc): Let Cj (j=1,2,..c) be a set of categorical attributes.

$$Lc = \sum H(A\ UCj)/ H(TCj) \qquad (\forall 1 \leq j \geq c)$$

Where, H(A(UCj)) be the sub tree rooted at the lowest common ancestor for each value of categorical attribute and H(TCj) is the height of the taxonomy tree. Finally, the total information loss IL can be represented as

$$IL= (| r | (\sum Ri\ max - Ri\ min / Di\ max - Di\ min) + \sum H(A\ UCj)/ H(TCj))$$

Where, *r* is the total number of records in cluster σ.

4. Experimental Evaluation

In this section, we present the effectiveness of our with respect to the parameters such as information loss and execution time. The experiment is implemented in Java with JDK 1.6 in a system configured with Intel core i5 processor, 4 GB RAM and 500GB hard disk.

We use the ADULT database from the UCI Machine Learning Repository [13] for experimentation. The ADULT database contains 32561 records and 15 attributes. We considered numerical attributes *Age* and *fnlwt* and categorical attributes *Race*, *Marital-status*, *Sex* and *Occupation*. The attribute *Occupation* is taken as a sensitive attribute in the database. The experiment executed for the various *k*-values such as 20, 40, 60, 80 and 100. In Figure-2, we show that our Approach equal combination of QI and SA achieves lesser information loss as compared to state-of-the-art clustering approaches viz. Greedy k-member algorithm and Systematic clustering algorithm.

Additionally in Figure-3, we show that our Approach achieves lesser execution time compared with the greedy k-member algorithm [4] and systematic clustering algorithm [7]. This is because; we use the minimum number of attributes in the generated sub-databases. Subsequently, we add the records in a cluster in a systematic way using systematic clustering algorithm [7] for the production of an anonymized database.

5. Conclusion

In this paper, we present Approach: Equal combination of quasi-identifier and sensitive attribute. Our approach generates anonymized sub-databases with a minimum number of attribute to reduce the risk of disclosure of sensitive attribute.

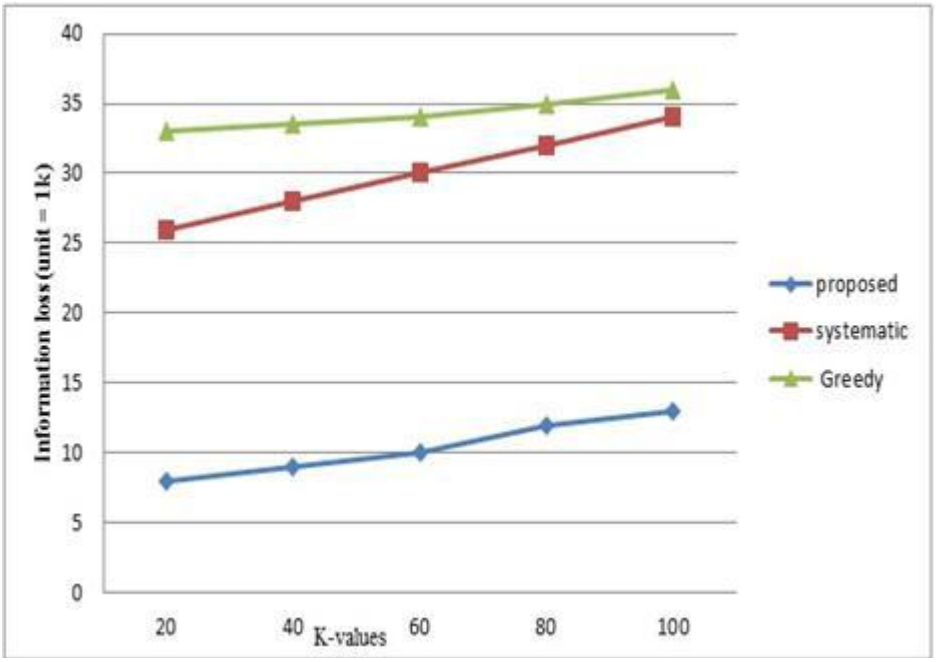

Figure-2: Information loss for Adult Dataset

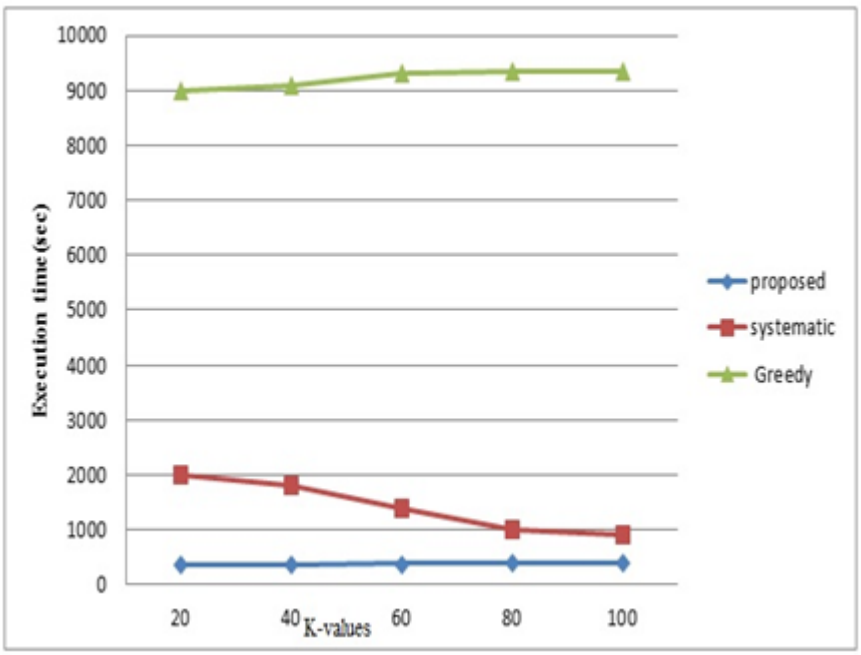

Figure-3: Execution Time for Adult Dataset

The proposed approaches use a concept of organized clustering algorithm for the generation of clusters to achieve minimal information loss and execution time. The experimental result shows that our proposed approach generate lesser information loss and execution time compared to Greedy k-member algorithm and Systematic clustering algorithm.

In addition, our Approach considers one SA for generating the sub-databases with QI attributes. Thus, our further work would be to investigate the performance on a combination of multiple SA and QI attributes.

REFERENCES

1. Y. Lindell and B. Pinkas, 2002, "Privacy preserving data mining," Journal of Cryptology,Vol. 15, pp. 177-206.

2. M. Upmanyu, A. M. Namboodiri, K. Srinathan, and C. V. Jawahar, 2010, "Efficient privacy preserving k-means clustering," Intelligent and Security Informatics, LNCS,Vol. 6122, pp. 154-166.

3. L. Sweeney, 2002, "k-Anonymity: a model for protecting privacy," International Journalon Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, pp. 557-570.

4. J. W. Byun, A. Kamra, E. Bertino, and N. Li, 2007, "Efficient k-anonymization usingclustering techniques," in Proceedings of International Conference on DatabaseSystems for Advanced Applications, pp. 188-200.

5. G. Loukides and J. Shao, 2007 "Capturing data usefulness and privacy protection in anonymization," in Proceedings of ACM Symposium on Applied Computing, pp. 370-374.

6. J.-L. Lin and M.-C. Wei, 2008, "An Efficient clustering method for k-anonymization," in Proceeding of International Workshop on Privacy and Anonymity in Information Society, pp. 46-50.

7. M. E. Kabir, H. Wang and E. Bertino, 2011 "Efficient systematic clustering method for k-anonymization," Acta Informatica, Vol. 48, pp. 51-66.

8. Pawan R. Bhaladhare and Devesh C. Jinwala, 2016 "Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model," Journal of Information Science and Engineering 32, 63-78.

9. X. Xiao and Y. Tao, 2006 "Anatomy: simple and effective privacy preservation," in Proceedings of the 32nd Intl. Conference on Very Large Data Bases, pp. 139-150.

10. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, 2006 “l-diversity: privacy beyond k-anonymity,” in Proceedings of the 22nd International Conference on Data Engineering, pp. 1-12.

11. N. Li, T. Li and S. Venkatasubramanian, 2007 “t-closeness: privacy beyond k-anonymity and l-diversity,” International Conference on Data Engineering, pp. 106-115.

12. R. C.-W. Wong, J. Li, A.W.-C. Fu, and K. Wang, 2006 “(α, k) anonymity: an enhanced k-anonymity model for privacy preserving data publishing,” in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 754-759.

13. UCI machine learning repository, http://archive.ics.uci.edu/ml/datasets.html.