International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 5 Issue 11, November-2016

Clustering Based L-Diversity Anonymity Model for Privacy Preservation of Data Publishing

A. Malaisamy[1], Dr. G. M. Kadhar Nawaz[2]

[1] Associate Professor, Dept. of Computer Applications, S. S. M. College of Engineering, Komarapalayam, Tamilnadu, India
[2] Director, Dept. of Computer Applications, Sona College of Technology, Salem, Tamilnadu, India

ABSTRACT

Privacy preservation data publication has received great attention in both the database community and theory community in recent years. Data publishing provides easiness for data exchange and data sharing. But, the personal privacy information leakage issues have become progressively more prominent. Anonymous algorithm is key technique to realize the privacy protection in data publishing environment. In addition, the anonymity algorithm of all sensitive attributes values is treated by lacking the sensitivity and specific distribution. Besides, it is vulnerable to similar attacks and deviation of attack. In order to overcome such limitations, Cluster based L-Diversity Privacy Preservation (CLDPP) model is proposed in this paper. The CLDPP model is designed to enhance the privacy preservation of anonymity data from different attacks such as homogeneous attack, background attack. The main goal of CLDPP model is to protect the data. The CLDPP model is designed Clustering based ℓ -Diversity Algorithm to group similar data together with ℓ-diverse sensitive values and then anonymizes each group individually with aiming at improving privacy preservation rate. In CLDPP model, ℓ- diversity concept is used to improve the data utility and to reduce the data loss in anonymity data sets when data point moves any size. Experimental evaluation of CLDPP model is done with the performance metrics such as privacy preservation rate, execution time to preserve privacy, anonymity level, and information loss. Experimental analysis shows that the CLDPP model is able to improve the privacy preservation rate by 16% and also improves the anonymity level by 9 % when compared to the state-of-the-art works.

Keyword: data publishing, privacy, anonymity, attack, clustering, ℓ- diversity, data utility

1. INTRODUCTION

A huge amount of personal data of an individual is collected by various organizations such as e-banking, online shopping medical and insurance agencies etc. Such collected data of an individual can be further examined digitally to discover useful information for a variety of purposes like medical research and market trend analysis. The data mining techniques are being employed to find out the useful information from the collected data.

Currently, privacy becomes a major concern and many research works have been designed for privacy protecting technology. Anonymization techniques present an approach to protect data privacy. In privacy preserving data mining, the l-diversity and k-anonymity models are the mainly employed for preserving the sensitive private information of an individual. Out of these two, l diversity model provides better privacy and lesser information loss while compared to the k-anonymity model.

Recently, most of research works has been developed for data privacy preservation. For example an efficient privacy-preserving K-means clustering algorithm was designed in [1] where the private data of the users, sensitive intermediate values and the final clustering assignments are protected by way of encryption. However, the information loss increases because of the occurrence of the outliers records in the cluster. A novel framework for collaborative fuzzy co-cluster analysis was presented in [2] with the objective of improving the privacy preservation. But, privacy preservation rate is poor.

The clustering approach based on the fractional calculus-bacterial foraging optimization algorithm in the l-diversity model was developed in [3] to reduce the information loss. Though, this approach does not effectively reduce the information loss.

The existing privacy preserving data mining techniques are classified in [4] depends on distortion, association rule, hide association rule, taxonomy, clustering, associative classification, outsourced data mining, distributed and k-anonymity where their notable advantages and disadvantages are highlighted. The various models designed for preserving privacy of social network data to provide privacy preservation with minimum information loss and better utility of released data was analyzed in [5].

In [6], new method was introduced for preserving the privacy of individuals' sensitive information from attribute and detect disclosure attacks which results in improved privacy and reduced information loss. A survey of many attacks techniques used for anonymization-based PPDM & PPDP and their causes on data privacy was described in [7]. A data slicing was developed in [8] to separates the data where efficient algorithm is designed for computing sliced data that accepts l-diversity requirement. The k-anonymity and l-diversity model for privacy preservation [9] was presented in social networks against neighborhood attacks to preserve privacy against neighborhood attacks.

2. RELATED WORKS

Cloud computing has shared resource, privacy and access control over network. It is essential to present the privacy in other vulnerable areas. In [10], Privacy Preservation is attained in Cloud Data Base by Incremental conceptual clustering algorithm in place of incremental k-means Variant of COBWEB is used at Cloud Database Owner to protect the Clients data. A security privacy-preserving data aggregation model implemented in [11] with mixed data aggregation structure. Data integrity is established at cluster head and base station. Several nodes used slicing technology to keep away from the leak of data at the cluster head in inner cluster.

A new technique is designed in [12] to employ the fractional calculus (FC) in chemotaxis step of BFO algorithm. The FC increases the computational results of the algorithm. In [13], collaborative framework is designed for partitioned co-occurrence matrices in fuzzy co-cluster structure estimation where co-occurrence information between the objects and items is stored in many sites. For using distributed data sets lacking the information leaks, a privacy preserving procedure is designed to fuzzy clustering for categorical multivariate data (FCCM).

Lightweight mobile re-authentication protocol is planned in [14] for mobile nodes. The protocol designs the forward secure pair wise key for mobile node when it shifts from one cluster to another. The mobile sensor node is validated by new cluster head and the privacy of origin area is preserved. A practical data publishing framework is planned in [15] for creating masked description of data that protects individual privacy and information effectiveness for cluster analysis. The key issue of masking data for cluster analysis is failed to have the class labels that guide the masking process.

An efficient privacy-preserving demand response (EPPDR) scheme with adaptive key evolution was introduced in [16] for achieving better computation and communication overheads. This EPPDR scheme can adaptively manage the key evolution to equalize the trade-off between the communication efficiency and security level. A privacy-preserving trust-based friend recommendation scheme for online social networks was presented in [17] which enable two strangers establish trust relationships based on the existing 1-hop friendships. The privacy of metadata was protected in [18] by applying meta-data management framework. In [19], novel clustering bee colony algorithm was presented to improve the privacy preservation rate. A comprehensive review on privacy preserving data mining was presented in [20].

Based on the aforementioned methods and techniques presented, in this paper, a Cluster based L-Diversity Privacy Preservation (CLDPP) model is proposed. The main goal of CLDPP model is to improve privacy preservation rate of data publishing and to reduce the information loss in multi dimensional data. The proposed CLDPP model is employed clustering technique for collect similar data together with ℓ-diverse sensitive values and then anonymizes each group separately. The contributions of CLDPP model include the following:

- To improve the privacy preservation of anonymity data from different attacks, Cluster based L-Diversity Privacy Preservation (CLDPP) model is introduced.
- To improve the privacy preservation rate of data publishing, Clustering based ℓ -Diversity Algorithm is employed in CLDPP model.
- To improve the data utility and to reduce the data loss, ℓ-diversity concept is applied in CLDPP model.

The rest of the paper organized as follows. In Section 2, a summary of different privacy preservation methods are explained. In Section 3, the proposed CLDPP model is described with the help of neat architecture diagram. In Section 4, simulation environment is presented with detailed analysis of results explained in Section 5. In Section 6, the concluding remarks are included.

3. CLUSTER BASED L-DIVERSITY PRIVACY PRESERVATION MODEL

The Cluster based L-Diversity Privacy Preservation (CLDPP) model is designed in this work with objective of improving the privacy preservation rate of data publishing. The proposed CLDPP model is used ℓ-diversity concept where it assumes that every group of indistinguishable records contains at least ℓ distinct sensitive attributes values. In CLDPP model, ℓ-diversity concept is applied with the aiming at improving the data utility and reducing the information loss of data. Besides, CLDPP model is proposed Clustering based ℓ -Diversity Algorithm to improve the privacy preservation of anonymity data from different attacks. The architecture diagram of Cluster based L-Diversity Privacy Preservation model is shown in below Figure 1.

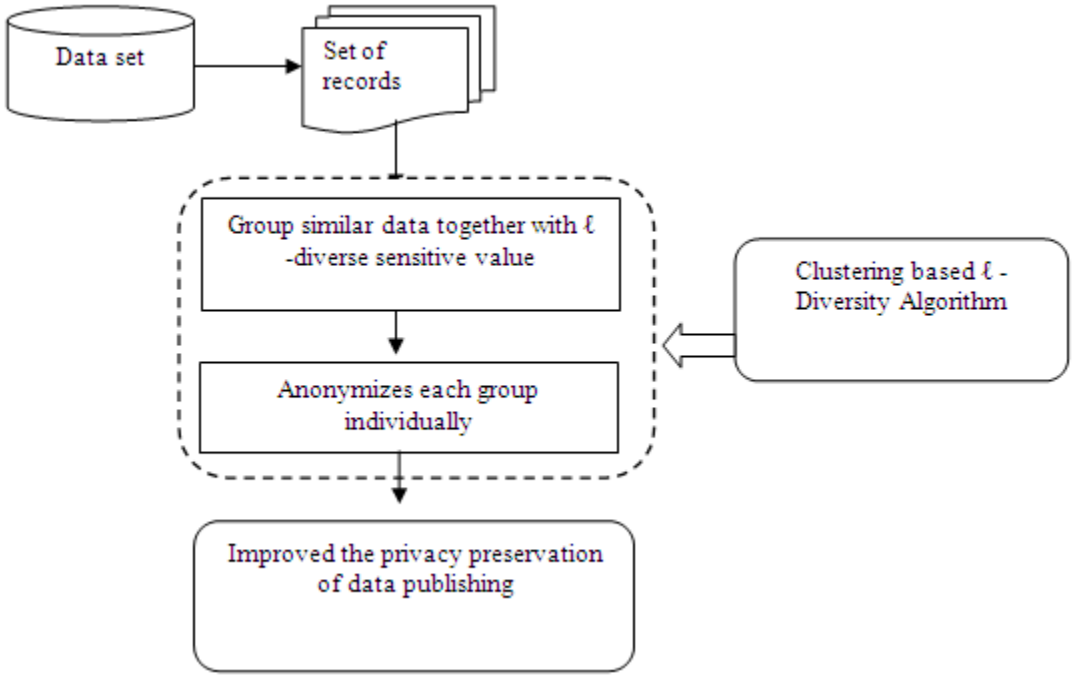

Figure 1 Architecture Diagram of Cluster Based L-Diversity Privacy Preservation Model

As shown in Figure 1, CLDPP model is used Clustering based ℓ-Diversity Algorithm with objective of improving the privacy preservation rate of data publishing. Clustering based ℓ -Diversity Algorithm in CLDPP model initially clusters similar data together with ℓ-diverse sensitive value and then Anonymizes each group separately which in turn improves the privacy preservation of anonymity data from different attacks. Therefore, CLDPP model is improves privacy preservation rate of data publishing in a significant manner.

3.1. ℓ-Diversity Concept

CLDPP model is used ℓ-diversity concept, it supposes that every group of identical records contains at least ℓ distinct sensitive attributes values. With the help of ℓ-diversity concept, CLDPP model is significantly improves the data utility and also reduces the information loss of data. In CLDPP model, ℓ-diversity is a type of group based anonymization which is used to preserve privacy in data sets by means of reducing the granularity of a data representation. This reduction is a trade off which result in some defeat of efficiency of data management or mining algorithms in order to achieve privacy. The ℓ-diversity concept is an expansion of the k-anonymity model that decreases the granularity of data representation. Therefore, CLDPP model is improves the data utility and reduces the information loss of data in an effective manner.

Due to the lack of the sensitive attribute value constraints in k-anonymity model, data is easily attacked by homogeneous attacks and background knowledge attacks. As a result, the l-diversity method is constructed to further k-anonymity as well maintaining the diversity of sensitive fields. The drawback of k-anonymization due to the background knowledge attack can be removed in CLDPP model by diversifying the values of sensitive attribute within a block. The ℓ -diversity model is a very useful model for preventing attribute disclosure. In CLDPP model, an every similarity group is said to have ℓ-diversity while it contains at least ℓ well signified values for the sensitive attribute. Besides, an anonymized table is said to have ℓ-diversity while every similarity group of the table has ℓ-diversity.

In ℓ-diversity concept, the entropy of a similarity group G is described by using the following mathematical formula,

$$\text{Entropy}(G) = -\Sigma s \epsilon S\, p(G,s) \log p(G,s) \ldots\ldots (1)$$

Where S indicates the field of the sensitive attribute and $p(G,s)$ characterize the fraction of records in G that include a sensitive value s. With the help of entropy of a similarity group G description, the sensitive attributes in CLDPP model is fine diverse through each group therefore it does not consider the semantically closeness of these values.

Let assume a data set $D = \{rec_1, rec_2, \ldots, rec_n\}$ which stores private information about a set of individuals with attributes like $A_1, A_2 \ldots, A_m$ where each record signifies an individual. CLDPP model consider that data set D is a subset of some larger population Ω in which all tuple $rec_i \in D$ signifies an individual from the population. For instance, CLDPP model assume a data set D is a medical dataset and then Ω could be the population of the Japanes, Russian, American, and Indian. Let A symbolize the set of all attributes $\{A_1, A_1, \ldots, A_m\}$ and $rec[A_i]$ indicate the value of attribute A_i for record ' rec'.

In CLDPP model, the attributes in data set D is divided into two types like sensitive attribute and non-sensitive attribute for improving the privacy preservation of data publishing. In CLDPP model, sensitive attribute is an attribute whose value for any particular individual must be kept secret from people who have no direct access to the original data with aiming at improving the data privacy-preserving of data publishing. We consider that the data publisher in CLDPP model knows which attributes are sensitive. All attributes that are not sensitive in data set D are termed as non-sensitive attributes in CLDPP model. An example of an attributes in medical data set is illustrated in below Table 1.

Table 1 Original Data Table of Medical Data Set

Serial number	Non-Sensitive attribute			Sensitive attribute
	Register code	Age	Nationality	Condition
1	1011	25	Japanes	Heart disease
2	1012	28	Russian	cancer
3	1013	29	American	Heart disease
4	1014	30	Indian	cancer

Table 1 shows the sensitive attributes and non-sensitive attributes in medical data set. In Table 1, the relationship between individuals and medical condition should be kept secret, therefore we should not known which particular patients have cancer, but it is allowable to known the information that cancer patients exist in the hospital. For performing the above process effectively in data publishing, CLDPP model is employed ℓ-diversity concept and clustering technique with aiming at improving the privacy preservation rate. In CLDPP model, ℓ-diversity concept is used for enhancing the data utility and reducing the information loss. Then, CLDPP model is used Clustering based ℓ-Diversity Algorithm for improving the performance of privacy preservation of data publishing in an effective manner. With the help of Clustering based ℓ-Diversity Algorithm, the attributes values which are not visible to the people who have no direct access to the original database is kept to be secret which results in improved privacy preservation rate in CLDPP model. The output of Clustering based ℓ-Diversity Algorithm is anonymous table of medical data set which shown in below Table 2.

Table 2 Anonymized Table of Medical Data Set

Serial number	Non-Sensitive attribute			Sensitive attribute
	Register code	Age	Nationality	Condition
1	10**	>30	***	Heart disease
2	10**	>30	***	cancer
3	10**	>30	***	Heart disease
4	10**	<30	***	cancer

The algorithm process of Clustering based ℓ-Diversity Algorithm is explained in following subsection.

3.2 Clustering based ℓ-Diversity Algorithm

CLDPP model is designed Clustering based ℓ-Diversity Algorithm to improve the privacy preservation of anonymity data from different attacks and to improve the privacy preservation of data publishing. The main objective of ℓ Clustering based ℓ-Diversity Algorithm is to compute an anonymized table D^*. The advantages of using Clustering based ℓ-Diversity Algorithm is CLDPP model is as follows. The output of Clustering based ℓ-Diversity Algorithm includes all the attributes

of data set D except the sensitive attribute. The Clustering based ℓ-Diversity Algorithm is used in CLDPP model comprises of a generalized record for every record in D and each similarity group includes at least l different sensitive attribute values. The Clustering based ℓ-Diversity Algorithm preserves as large amount of information of D as possible.

Given a dataset D, Clustering based ℓ-Diversity Algorithm first divides the records into different buckets depends on sensitive attribute values. So that every bucket comprises the same sensitive attribute value and then sort these buckets based on their size. Then, Clustering based ℓ-Diversity Algorithm in CLDPP model group records into a number of clusters in which every group comprises at least l different sensitive attribute values with the reduced information loss. In Clustering based ℓ-Diversity Algorithm, there are three different ways used to select record from buckets like selecting record randomly, selecting record on ascending order or descending order based on their bucket size. In CLDPP model, Clustering based ℓ-Diversity Algorithm is used descending order for selecting the record from buckets which described in following algorithmic steps.

// Clustering based ℓ-Diversity Algorithm
Input: a dataset D and a diverse anonymity threshold value l.
Output: the anonymized dataset D^*
Begin
Step 1: create a collection of buckets for different sensitive attribute values and the perform sorting based on their sizes resulting in $B = b_1, b_{2....} b_n$
Step 2: Return if the number of buckets is < l
Step 3: Let result =∅
Step 4: While (the number of non-empty buckets is ≥ l)
Step 4:1 Randomly select a record rec_i from the maximal non-empty bucket b and create it as a cluster $c = \{rec_i\}$
Step 4:2 $b = b - \{rec_i\}$
Step 4:3 While $|c| < l$
Step 4:3.1 select a record rec_j from the smaller non-empty bucket so that Information Loss$(c \cup \{rec_j\})$ is minimal
Step 4:3.2 $b = b - \{rec_j\}$
Step 4:3.3 $c = c \cup \{rec_j\}$
Step 4:4 result = result ∪ c
Step 5: While (the number of non-empty buckets is > 0)
Step 5:1 Randomly select a record rec_k from the non-empty bucket b
Step 5:2 $b = b - \{rec_k\}$
Step 5:3 select a cluster c so that Information Loss$(c \cup \{rec_k\})$ is minimal
Step 5:4 $c = c \cup \{rec_k\}$
Step 6: generate anonymous similarity group by using local recoding techniques on each Cluster
Step 7: Return a anonymized dataset D^*

Figure 2 Clustering based ℓ-Diversity Algorithm

As shown in Figure 2, CLDPP model is used above algorithmic step for improving the privacy preservation of data pushing. In CLDPP model, Clustering based ℓ-Diversity Algorithm initially create a collection of buckets and then performs sorting based on bucket sizes. Next, Clustering based ℓ-Diversity Algorithm arbitrarily select a record from the bucket and then construct it as a cluster. After that, this algorithm selects a record from the next bucket. This procedure is repeated until $|c| < l$. While $|c|$ reaches l, ℓ Clustering based ℓ-Diversity Algorithm select a record from the current bucket and reiterate the clustering process. Then, Clustering based ℓ-Diversity Algorithm is iterates over these remaining records and adds each record into a cluster which in turn preserves the data from the different attacks like homogeneous attack, background attack in an effective manner. Finally, Clustering based ℓ-Diversity Algorithm is generate anonymous similarity group by using local recoding techniques on each cluster which results in improved privacy preservation rate of data publishing in a significant manner.

3.3 Information Loss

While achieving l-diverse anonymity in proposed CLDPP model, the information loss is a very significant problem. To create a similarity group in Clustering based ℓ-Diversity Algorithm, it is obvious to cluster the records which have the

nearest distance together to have the least information loss. In a medical data set, there are two types of data like numeric and categorical data. Generally, the categorical data can be discreted to numeric data, but not all categorical data can do so. For that reason, we require two distance functions to compute numeric data and categorical data respectively.

Let us consider a cluster $c = \{r_1, r_2, \ldots, r_n\}$ in which the quasi-identifiers comprise of numeric attributes such as $NA_1, NA_2, \ldots NA_m$ and categorical attributes such as $CA_1, CA_2, \ldots CA_m$. Let assume T_{CA_i} be the taxonomy tree described for the domain of categorical attribute CA_i. Let MIN_{NA_i} and MAX_{NA_i} be the min and max values in c with respect to attribute NA_i, and let $\cup_{CA_i}$ be the union set of values in cluster c with respect to attribute CA_i. Then, the amount of information loss happened by generalizing cluster c is denoted as Information Loss(c) which is mathematically formulated as,

$$\text{Information Loss}(c) = |c|.\left(\sum_{i=1,\ldots,m} w_i \frac{(MAX_{NA_i} - MIN_{NA_i})}{|NA_i|} + \sum_{j=1,\ldots,n} w_j \frac{H(\wedge(\cup_{CA_j}))}{H(T_{CA_j})}\right) \ldots (2)$$

From (2), $|c|$ is the number of records in cluster c, $|NA_i|$ symbolizes the size of numeric domain NA_i. And w_i is indicates the weight of attribute NA_i, $\wedge(\cup_{CA_j})$ represent the sub-tree rooted at the lowest common ancestor of every value in $\cup_{CA_j}$ and H(T) is the height of taxonomy tree T.

Let E be the set of all similarity groups in the anonymized dataset D^*. Then the amount of total information loss of D^* is mathematically formulated as follows,

$$\text{Total} - \text{Information Loss}(D^*) = \sum_{e \in E} \text{Information Loss}(c) \ldots\ldots (3)$$

The proposed CLDPP model is used ℓ-diversity concept, therefore the information loss of data publishing is reduced in an effective manner.

4. EXPERIMENTAL SETTINGS

The proposed Cluster based L-Diversity Privacy Preservation (CLDPP) model is implemented using Java language. The CLDPP model is used the Adult dataset from the University of California Irvine data repository. The Adult dataset consists of information about the individuals such as age, level of education and current employment type. The dataset used for CLDPP model to improve the privacy preservation of data publishing includes forty nine thousand records and a binomial label representing a salary of less or greater than fifty thousand US dollars. The dataset take account of fourteen attributes consists of seven polynomials, one binomial and six continuous attributes which is shown in below Table 3.

Table 3 Adult dataset with description

Attributes	Description
Employment class	Self employed, state government, local government
Education level	High school, college, bachelor, masters
Relationship	Husband, own child, married
Race	White, black, asian
Marital status	Married, never married, divorced
Occupation	Scales, armed forces, technical support, admin clerical
Country	United states, Germany, Canada
Salary	>50K, < 50K
Gender	Male, Female
Age	29, 31, 35
Hours worked per week	25, 120,110
Education number	9, 13, 18

The efficiency of the CLDPP model is evaluated with the metric such as privacy preservation rate, execution time to preserve privacy, anonymity level, and information loss. The performance of proposed CLDPP model is compared against with the existing two methods namely, privacy-preserving K-means clustering algorithm [1] and collaborative fuzzy co-cluster analysis framework [2] respectively.

5. DISCUSSION

In this section, the result analysis of CLDPP model is estimated. The performance of CLDPP model is compared against with exiting two methods namely, privacy-preserving K-means clustering algorithm [1] and collaborative fuzzy co-cluster analysis framework [2]. The performance of CLDPP model is evaluated along with the following metrics.

5.1 Measurement of Privacy Preservation Rate

In CLDPP model, the privacy preservation rate is defined the amount of data that are must be kept secret from people who have no direct access to the original to the total number of available data. The privacy preservation rate is measured in terms of percentage (%) and formulated as,

$$\text{privacy preservation rate} = \frac{\text{amount of data that are must be kept secret from people}}{\text{total number of available data}} \ldots (4)$$

When the privacy preservation rate is higher, the method is said to be more efficient.

Table 4 Tabulation for Privacy Preservation Rate

Record size (MB)	Privacy Preservation Rate (%)		
	CLDPP model	privacy-preserving K-means clustering algorithm	collaborative fuzzy co-cluster analysis framework
150	82.44	63.25	74.56
300	84.56	65.47	76.54
450	86.23	67.52	78.21
600	88.14	69.42	80.44
750	90.58	71.47	82.41
900	92.78	73.58	84.36
1050	94.58	75.56	86.97

To determine the performance of the privacy preservation rate of CLDPP model, comparison is made with two other existing methods, privacy-preserving K-means clustering algorithm [1] and collaborative fuzzy co-cluster analysis framework [2]. In Table 4, the number of record size is varied in range of 150 to1050. From the table value, it is illustrative that the privacy preservation rate of data publishing using the proposed CLDPP model is improved when compared to other existing methods [1], [2].

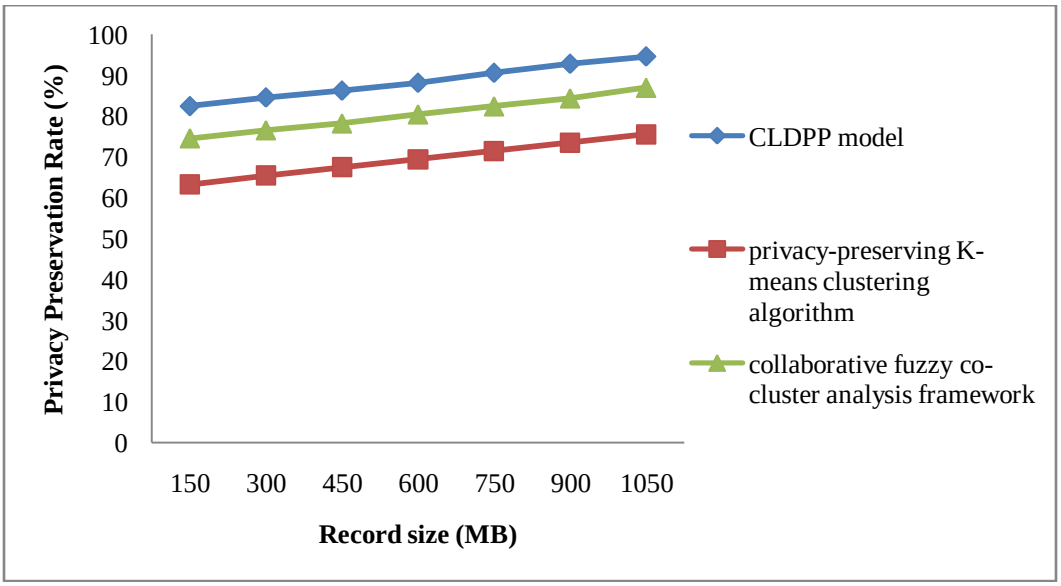

Figure 3 Measurement of Privacy Preservation Rate

Figure 3 shows the privacy preservation rate of data publishing versus different number of record size in the range of 150 to 1050. As shown in figure, privacy preservation rate of data publishing using CLDPP model provides better performance when compared to two other methods namely privacy-preserving K-means clustering algorithm [1] and collaborative fuzzy co-cluster analysis framework [2]. Besides, while increasing the number of record size, the data privacy preservation rate is gets also increased. But comparatively the privacy preservation rate using proposed CLDPP model is higher.

This is because of the application of Clustering based ℓ-Diversity Algorithm in CLDPP model. With the support of Clustering based ℓ-Diversity Algorithm, the attributes values which are not visible to the people who have no direct access to the original database is kept to be secret efficiently which results in improved privacy preservation rate in an effective manner. As a result, CLDPP model is improves the privacy preservation rate of data publishing by 22% as compared to privacy-preserving K-means clustering algorithm [1] and 9% as compared collaborative fuzzy co-cluster analysis framework [2] respectively.

5.2 Measurement of Execution Time to Preserve Privacy

In CLDPP model, to measure the impact of execution time to preserve privacy, the size of the records and time taken to add a significant amount of noise factor is considered. Therefore, the execution time to preserve privacy is defined as the time taken to process the entire privacy preservation scheme based on the addition of the random noise (i.e. the product of the size of the records and the total time taken to add the noise). The execution time to preserve privacy is mathematically formulated as given below,

$$\text{Execution Time} = \text{Record size} * \text{Time}\ (\varepsilon_j) \quad (5)$$

From (5), the execution time to preserve privacy is based on the time taken to add the noise '(ε_j)' with respect to the record size obtained from Adult dataset. The execution time to preserve privacy is measured in terms of milliseconds (ms). When the execution time to preserve privacy is lower, the method is said to be more efficient.

Table 5 Tabulation for Execution Time to Preserve Privacy

Record size (KB)	Execution time to preserve privacy (ms)		
	CLDPP model	collaborative fuzzy co-cluster analysis framework	privacy-preserving K-means clustering algorithm
200	9	14	16
400	13	18	20
600	17	22	24
800	21	26	28
1000	25	30	32
1200	29	34	36
1400	33	38	40

The execution time to preserve privacy using CLDPP model is elaborated in Table 5. We consider the framework with different number of record size in the range of 150 to 1050 is taken for experimental purpose using Java language. From the table value, it is illustrative that the execution time to preserve privacy using CLDPP model is reduced when compared to the other existing methods [1], [2].

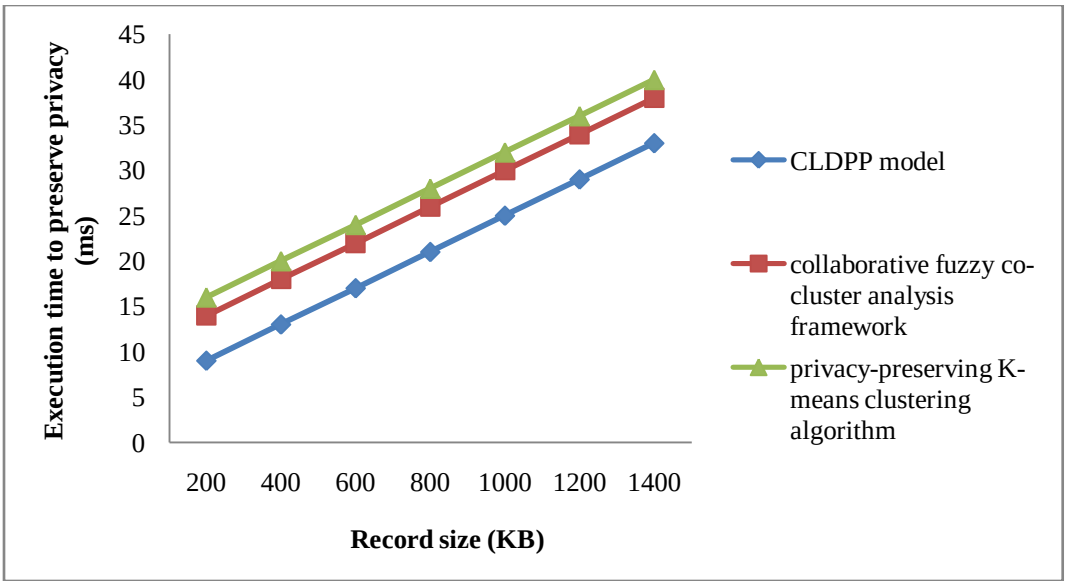

Figure 4 Measurement of Execution Time to Preserve Privacy

Figure 4 shows the execution time to preserve privacy using CLDPP model and exiting privacy-preserving K-means clustering algorithm [1] and collaborative fuzzy co-cluster analysis framework [2]. As shown in figure, the execution time to preserve privacy using CLDPP model provides the better performance when compared to other existing methods [1], [2]. Besides, while increasing the number of record size, the execution time to preserve privacy is gets also increased. But comparatively the execution time to preserve privacy using proposed CLDPP model is reduced. This is because of the application of Clustering based ℓ-Diversity Algorithm in CLDPP model. In CLDPP model, Clustering based ℓ-Diversity Algorithm is divides the records into different buckets based on sensitive attribute values and then sort these buckets based on their size. As a result, CLDPP model is reduces the execution time to preserve privacy by 29% as compared to privacy-preserving K-means clustering algorithm [1] and 40% as compared collaborative fuzzy co-cluster analysis framework [2] respectively.

5.3 Measurement of anonymity level

In CLDPP model, Anonymity refers to the level that in a way allows the owner of the database without disclosing the content of the tuple to the third party or the person those who are not intended to (i.e. data provider and database owner). Therefore, the anonymity level is the ratio of the size of the record to the size of the record that has maintained anonymity and is mathematically formulated as follows,

$$\text{anonymity level} = \frac{\text{Size}_r}{\text{Size}_a} * 100 \qquad (6)$$

From (6), the anonymity level is measured based on the size of the record to the size of the record that has maintained anonymity. The anonymity level is measured in terms of percentage (%). When the anonymity level is higher, more efficient the method is said to be.

Table 6 Tabulation for Anonymity Level

Record size (MB)	Anonymity level (%)		
	CLDPP model	collaborative fuzzy co-cluster analysis framework	privacy-preserving K-means clustering algorithm
150	80.21	75.65	70.85
300	83.65	78.62	73.63
450	86.61	81.25	76.59

600	89.57	84.63	79.45
750	92.13	87.32	82.14
900	95.26	90.45	85.54
1050	98.45	93.25	88.14

The anonymity level using CLDPP model is elaborated in Table 6. We consider the framework with different number of record size in the range of 150 to 1050 is taken for experimental purpose using Java language. From the table value, it is illustrative that the anonymity level using CLDPP model is higher when compared to the other existing methods [1], [2].

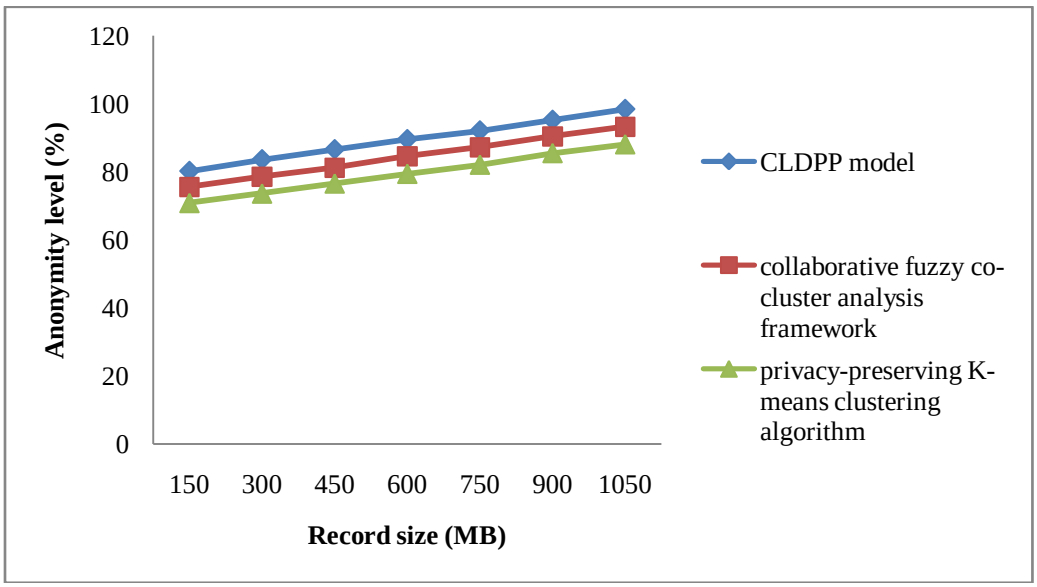

Figure 5 Measurement of Anonymity Level

Figure 5 shows the Anonymity Level of data publishing versus different number of record size in the range of 150 to 1050. As shown in figure, Anonymity Level of data publishing using CLDPP model provides better performance when compared to two other methods namely privacy-preserving K-means clustering algorithm [1] and collaborative fuzzy co-cluster analysis framework [2]. Besides, while increasing the number of record size, the Anonymity Level is gets also increased. But comparatively the Anonymity Level using proposed CLDPP model is higher. This is because of the application of Clustering based ℓ-Diversity Algorithm in CLDPP model. With the support of Clustering based ℓ-Diversity Algorithm, CLDPP model is preserves as much information of data publishing. As a result, CLDPP model is improves the Anonymity Level of data publishing by 6% as compared to privacy-preserving K-means clustering algorithm [1] and 11% as compared collaborative fuzzy co-cluster analysis framework [2] respectively.

5.4 Measurement of Information Loss

In CLDPP model, information loss measures the information loss of an anonymous data table. The concept of information loss or data distortion often is used to reflect the data quality in privacy-preserving publishing. Information loss usually decreases the quality of the data and affects data utility. When information loss is low, the method is said to be more efficient.

Table 7 Tabulation for Information Loss

Methods	Information Loss (%)
CLDPP model	14
collaborative fuzzy co-cluster analysis framework	18
privacy-preserving K-means clustering algorithm	20

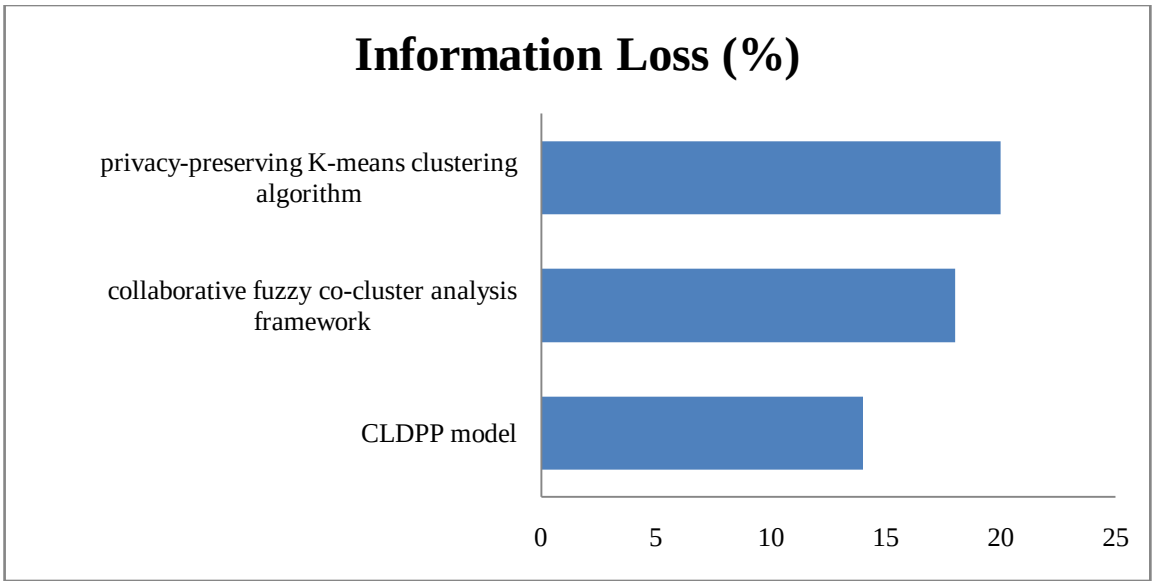

Figure 6 Measurement of Information Loss

Table 7 and Figure 6 demonstrate the Information Loss of CLDPP model versus different number of record size in the range of 150 to 1050. As shown in figure, CLDPP model of using CLDPP model provides better performance when compared to two other methods namely privacy-preserving K-means clustering algorithm [1] and collaborative fuzzy co-cluster analysis framework [2]. This is because of the application of ℓ-Diversity concept in CLDPP model. With the support of ℓ-Diversity concept, CLDPP model is significantly reduce the amount of total information loss. As a result, CLDPP model is reduce the information loss of data publishing by 29% as compared to privacy-preserving K-means clustering algorithm [1] and 13% as compared collaborative fuzzy co-cluster analysis framework [2] respectively.

CONCLUSION

In this work, an effective novel framework is designed called Cluster based L-Diversity Privacy Preservation (CLDPP) model to improve the performance privacy preservation of anonymity data from different attacks such as homogeneous attack, background attack. In proposed CLDPP model, Clustering based ℓ -Diversity Algorithm is developed to group similar data together with ℓ-diverse sensitive values and then anonymizes each group independently which resulting in improved privacy preservation rate of data publishing. Clustering based ℓ -Diversity Algorithm is applied in CLDPP model initially splits the records into number of different buckets depends on sensitive attribute values and then sort these buckets based on their bucket size. Then, Clustering based ℓ-Diversity Algorithm in CLDPP model group records into a number of different clusters where every cluster comprises at least l different sensitive attribute values with the reduced information loss of data. Afterward, Clustering based ℓ-Diversity Algorithm is iterates over these remaining records and adds each record into a cluster which in turn preserves the data from the different attacks. Finally Clustering based ℓ-Diversity Algorithm is generate anonymous similarity group with the help of local recoding techniques on each cluster which in turn improves the privacy preservation rate of data publishing. The performance of CLDPP model is tested with Adult dataset and compared against with existing methods. With the experiments conducted for model, it is observed that the privacy preservation rate of data publishing provides more accurate results as compared to state-of-the-art works. The experimental results show that CLDPP model is provides better performance with an improvement of privacy preservation rate by 16% and also improves the anonymity level by 9 % when compared to state-of-the-art works.

REFERENCES

[1]. Zekeriya Erkin, Thijs Veugen, Tomas Toft and Reginald L Lagendijk, "Privacy-preserving distributed clustering", Springer, EURASIP Journal on Information Security, Volume 2013, Issue 4, November 2013, Pages 1-15.

[2]. Katsuhiro Honda, Toshiya Oda, Daiji Tanaka, and Akira Notsu, "A Collaborative Framework for Privacy Preserving Fuzzy Co-Clustering of Vertically Distributed Cooccurrence Matrices", Hindawi Publishing Corporation, Advances in Fuzzy Systems, Volume 2015, March 2015, Pages 1-9.

[3]. Pawan R. Bhaladhare and Devesh C. Jinwala, "A Clustering Approach for the -Diversity Model in Privacy Preserving Data Mining Using Fractional Calculus-Bacterial Foraging Optimization Algorithm", Advances in Computer Engineering, Volume 2014 (2014), Article ID 396529, 12 pages

[4]. Yousra Abdul Alsahib S. Aldeen, Mazleena Salleh and Mohammad Abdur Razzaque, "A comprehensive review on privacy preserving data mining", Springer, 2015
[5]. Amardeep Singh, Divya Bansal, Sanjeev Sofat, "Privacy Preserving Techniques in Social Networks Data Publishing-A Review", International Journal of Computer Applications (0975 – 8887) Volume 87, Issue.15, February 2014
[6]. R.Mahesh and Meyyappan, "New Method for Preserving Privacy in Data Publishing Against Attribute and Identity Disclosure Risk", International Journal on Cryptography and Information Security (IJCIS), Vol.3, No. 2, June 2013
[7]. Abou-el-ela Abdou Hussien, Nermin Hamza, Hesham A. Hefny, "Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing", Journal of Information Security, 2013, 4, Pages 101-112
[8]. Alphonsa Vedangi, V.Anandam, "Data Slicing Technique to Privacy Preserving and Data Publishing", IJRET: International Journal of Research in Engineering and Technology, Volume: 02, Issue: 10, Oct-2013
[9]. Bin Zhou, Jian Pei, "Pei, J.: The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks, Knowledge and Information Systems, Research gate, 28(1), Pages 47-77
[10]. Nazneen Mulani, Ambika Pawar, Preeti Mulay, Ajay Dani., "Variant of COBWEB Clustering for Privacy Preservation in Cloud DB Querying", Procedia Computer Science, Elsevier, Volume 50, 2015, Pages 363 – 368
[11]. Changlun Zhang, Chao Li, and Jian Zhang, "A Secure Privacy-Preserving Data Aggregation Model in Wearable Wireless Sensor Networks", Journal of Electrical and Computer Engineering, Hindawi Publishing Corporation, Volume 2015, September 2015, Pages 1-9
[12]. Pawan R. Bhaladhare and Devesh C. Jinwala, "A Clustering Approach for the *l*-Diversity Model in Privacy Preserving Data Mining Using Fractional Calculus-Bacterial Foraging Optimization Algorithm", Hindawi Publishing Corporation, Advances in Computer Engineering, Volume 2014, September 2015, Pages 1-12
[13]. Katsuhiro Honda, Toshiya Oda, Daiji Tanaka, and Akira Notsu, "A Collaborative Framework for Privacy Preserving Fuzzy Co-Clustering of Vertically Distributed Cooccurrence Matrices", Hindawi Publishing Corporation, Advances in Fuzzy Systems, Volume 2015, March 2015, Pages 1-8
[14]. Shunrong Jiang, Jiapeng Zhang, JingJun Miao, and Conghua Zhou, "A Privacy-Preserving Reauthentication Scheme for Mobile Wireless Sensor Networks", Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, Volume 2013, April 2013, Pages 1-8
[15]. Benjamin C.M. Fung, Ke Wang, Lingyu Wang, Patrick C.K. Hung, "Privacy-preserving data publishing for cluster analysis", Data and Knowledge Engineering, Elsevier, Volume 68, 2009, Pages 552–575
[16]. Hongwei Li, Xiaodong Lin, Haomiao Yang, Xiaohui Liang, Rongxing Lu, and Xuemin (Sherman) Shen, "EPPDR: An Efficient Privacy-Preserving Demand Response Scheme with Adaptive Key Evolution in Smart Grid", IEEE Transactions on Parallel and Distributed Systems, Year: 2014, Volume: 25, Issue: 8, Pages: 2053 – 2064
[17]. Linke Guo, Chi Zhang, and Yuguang Fang, "A Trust-based Privacy-Preserving Friend Recommendation Scheme for Online Social Networks", IEEE Transactions on Dependable and Secure Computing, Volume: 12, Issue: 4, Pages 413 - 427
[18]. Yves-Alexandre de Montjoye, Erez Shmueli1, Samuel S. Wang, Alex Sandy Pentland, "openPDS: Protecting the Privacy of Metadata through SafeAnswers", PLoS ONE, Volume 9, Issue 7, July 2014, Pages 1-9.
[19]. Jinchao Ji, Wei Pang, Yanlin Zheng, Zhe Wang, Zhiqiang Ma, "A Novel Artificial Bee Colony Based Clustering Algorithm for Categorical Data", PLoS ONE, Volume 10, Issue 5, May 2015, Pages 1-17.
[20]. Yousra Abdul Alsahib S. Aldeen1, Mazleena Salleh and Mohammad Abdur Razzaque, "A comprehensive review on privacy preserving data mining", SpringerPlus, Volume 4, November 2015, Pages 1-36.