Journal of Theoretical and Applied Information Technology

15th October 2023. Vol.101. No 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP LEARNED KERNEL SPECTRAL CLUSTERING AND PEARSON RANK SWAPPING ANONYMIZATION FOR PRIVACY PRESERVED DATA PUBLISHING

A. MALAISAMY [1] AND DR. G. M. KADHAR NAWAZ [2]

[1] Ph.D Research Scholar, Dept. of Computer Applications, S. S. M. College of Engineering,
Komarapalayam, Tamilnadu, India
[2] The Principal, Sona College of Arts and Science, Salem, Tamilnadu, India
E-mail: [1] gamalaisamy@gmail.com, [2] nawazse@yahoo.co.in

ABSTRACT

Privacy-preservation is a challenging issue with the increasing volumes of published data. To overcome such issues in data publishing, different methods of reduced risks related to published data have been developed. However, it has a vital problem to protect the users' sensitive data in a publication. Since the attacker may hack the user data. Therefore, the Deep Learned Kernel Spectral Clustering-based Pearson Rank Proximity Swapping Anonymization (DLKSC-PRPSA) technique is developed for improving the data privacy preservation rate with lesser information loss. The proposed DLKSC-PRPSA technique collects the number of records from the dataset. Then the DLKSC-PRPSA technique trained the input records with several layers namely the input layer, two hidden layers, and the output layer. The numbers of records are given to the input layer of deep neural learning. Then the input is fed into the first hidden layer to group the records into different clusters using a radial basis kernelized spectral clustering technique. The clustered results are transformed into the next hidden layer. In the second hidden layer, the Pearson rank proximity swapping anonymization method is applied for interchanging the value of sensitive attributes to protect the original information. Finally, the anonymized results are obtained at the output layer for further processing. Experimental evaluation is carried out with adult datasets using different metrics such as privacy preservation rate, anonymity level, information loss, and time complexity. The experimental result confirms that the DLKSC-PRPSA technique efficiently increases the privacy preservation rate, anonymity level and minimizes the time complexity as well as information loss of data anonymization than the state-of-the-art methods.

Keywords: *Privacy Preservation, Data Publishing, Deep Learning, Radial Basis Kernelized Spectral Clustering, Pearson Rank Proximity Swapping Anonymization Method.*

1. INTRODUCTION

With the increasing accessibility of public open data, security has become a very important concern for publishing. However, there are a large number of risks in data publishing platforms. The attackers may gain the knowledge to access the published statistical data and discover particular individual's information, which may cause more information loss. To address this problem, a deep learning-based approach is developed for privacy-preserving data publishing in this paper.

A Restricted Sensitive Attributes-based Sequential Anonymization (RSASA) method was introduced in [1] for enhancing the data stream publication with privacy protection. The designed method failed to achieve a higher privacy rate with minimum time complexity. A Sensitive Label Privacy Preservation with Anatomization (SLPPA) method was developed in [2] to preserve the privacy of available data. But the method failed to preserve the privacy of available data with more sensitive attributes.

A data sanitization approach was developed in [3] for protecting the user sensitive information from the attackers. The designed approach did not minimize the time complexity for privacy preservation. A multi-probing Locality-Sensitive Hashing (LSH) technique was developed in [4] to guarantee the protection of recommender systems. But the performance of privacy preservation of the approach was not measured.

A two sanitization approach was introduced in [5] for ensuring latent-data privacy. The method failed to use an efficient anonymization method to further enhance the data privacy. An item-based Collaborative Filtering (ICF) method was introduced in [6] by combining the locality-sensitive hashing technique to enhance secure data publishing. But the method failed to investigate the multiple types of quality data. A novel personalized extended (α, k)-anonymity method was introduced in [7] to provide efficient data privacy. Though the method minimizes the execution time, the performance of information loss remained unsolved.

A Multilevel Privacy-Preserving Data Sharing (MPPDS) approach was developed in [8] for the group of health data shared by different data owners. The performance of the privacy preservation rate was not improved. A Role-Task Conditional Purpose Policy-based protection method was introduced in [9] for improving data privacy. But the method failed to process the multiple records with minimum time. A Secure and Efficient data perturbation Algorithm using Local differential privacy (SEAL) was designed in [10] to provide higher privacy and utility. The algorithm takes more timestamps and latencies to provide the privacy of data records.

From the above discussion, the limitations of the existing methods are overcome by introducing a DLKSC-PRPSA technique. The major contribution of the DLKSC-PRPSA technique is described as follows,

- A privacy-aware structural data publishing technique is developed called as DLKSC-PRPSA to protect the unreleased data privacy effectively as well as guaranteeing the minimum information loss.

- First, the radial basis kernelized spectral clustering is applied to the hidden layer of deep neural learning to group the records into multiple clusters. The radial basis kernel function is applied to measure the similarity between the two records and applying the k-means algorithm to cluster the records. This helps to minimize the time complexity of data anonymization.

- Secondly, the Pearson rank proximity swapping anonymization is applied to preserve the data privacy. For each cluster, the sensitive attribute information is interchanged based on Pearson correlation measure. The anonymized results are obtained at the output layer. This helps to improve the data privacy preservation rate and minimizes the information loss.

1.1 Outline Of The Paper

The paper is organized into five different sections. In Section 2, the proposed methodology DLKSC-PRPSA technique is explained with the help of the architecture diagram. In Section 3, experimental settings are described and the performance results of different techniques are presented in Section 4. Section 5 discusses the related works. Finally, section 6 shows the conclusion of the paper.

2. DEEP LEARNED KERNEL SPECTRAL CLUSTERING-BASED PEARSON RANK PROXIMITY SWAPPING ANONYMIZATION

A Deep Learned Kernel Spectral Clustering-based Pearson Rank Proximity Swapping Anonymization (DLKSC-PRPSA) technique is introduced for protecting the privacy with the increasing volumes of published data. The deep learning-based technique effectively improves data privacy through the two major processes namely clustering and anonymization. The clustering process minimizes the time complexity of the data anonymization. The data anonymization process is considered as the main process in data publishing that aims to hide sensitive information for further data analysis and utilization. The architecture of the DLKSC-PRPSA technique is shown in figure 1.

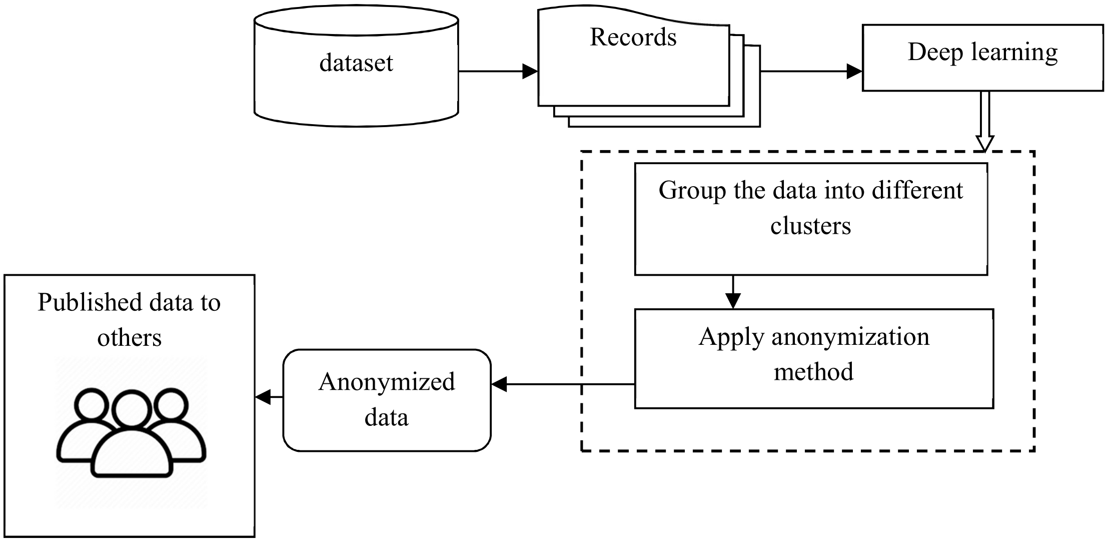

Figure 1 Architecture Of DLKSC-PRPSA Technique

Figure 1 shows the architecture of the proposed DLKSC-PRPSA technique to achieve privacy preservation on published data. As shown in above Figure 1, at first the numbers of records are collected from the dataset. Then, the input records are divided into a number of clusters. The proposed DLKSC-PRPSA technique uses deep learning for both clustering the records and data anonymization for publishing. Thus, the proposed technique enhances the accuracy of privacy preservation with lesser information loss. The structure of the deep learning-based approach is shown in figure 2.

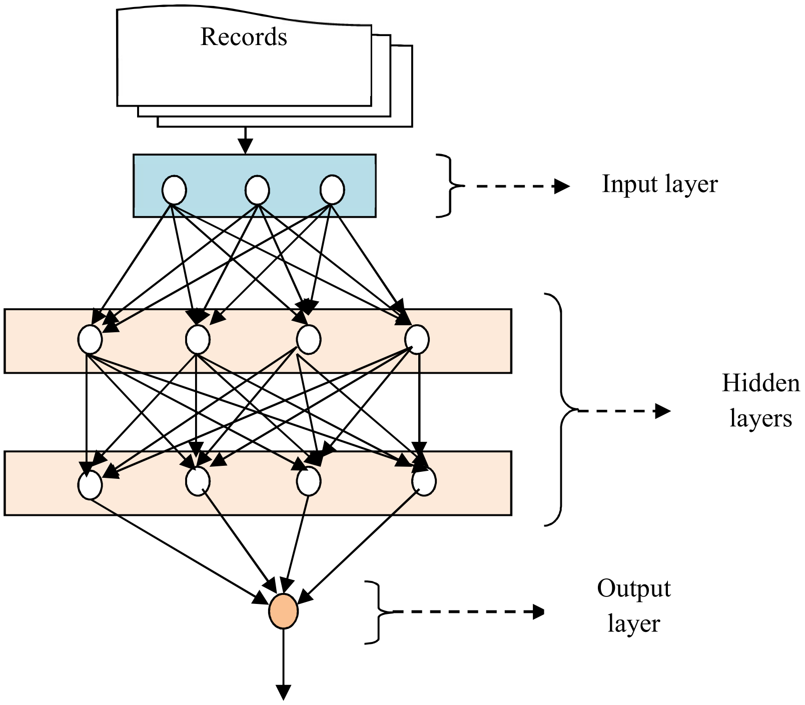

Figure 2 Structure Of The Deep Neural Learning

Figure 2 illustrates the structure of the deep neural learning with the number of records taken from the dataset. The deep neural learning is a type of machine learning methods that uses more than one layer such as one input layer, two hidden layers and one output layer to process the raw input (i.e. records). The word "deep" represents the number of layers in the network by which the input records is transformed from one level to another. Hence the deep learning is called a layer-by-layer method. Each level of the network learns the input records and its transformed into the next layers. The deep neural learning includes the neurons-like the nodes into different layers. The nodes in the one layer are fully connected to the subsequent layers and perform the deep learning of input data hence the name is called deep neural learning. The numbers of records $\mathrm{r}_1, \mathrm{r}_2, \mathrm{r}_3 \ldots . \mathrm{r}_\mathrm{n}$ are given to the input layer at a time't' i.e. $i(t)$. The input layer of the deep learning is expressed as,

$$i(t) = \sum_{i=1}^{n} \mathrm{r}_\mathrm{i} * \mathrm{b}_1 \quad (1)$$

Where, $i(t)$ represents the input at a time't', r_i denotes a number of records, b_1 denotes an adjustable weights between the input and first hidden layer. In the DLKSC-PRPSA technique, two hidden layers are used for preserving the privacy of the data. In the first hidden layers, the clustering is performed to group the records. Here the radial basis kernelized spectral clustering technique is applied in the first hidden layer. The radial basis kernel is applied for calculating the similarity between two records.

$$k(r_i, r_j) = \exp\left(-0.5 * \frac{t_{ij}}{\sigma^2}\right) \, (2)$$

Where, $k(r_i, r_j)$ represents the radial basis kernel function measure the similarity between the pair of records (r_i, r_j) in Euclidean space, σ denotes a deviation and t_{ij} denotes a distance between the pair of records which is measured as follows,

$$t_{ij} = \left\|r_i - r_j\right\|^2 \, (3)$$

From (3), t_{ij} represents a Euclidean distance between the two records r_i and r_j. Based on the distance measure, the weight matrix v_{ij} is constructed. Then the unnormalized Laplacian matrix is constructed with the weight matrix,

$$NL_{ij} = g_{ij} - v_{ij} \quad (4)$$

Where, NL_{ij} denotes a unnormalized Laplacian matrix, g_{ij} denotes a diagonal matrix, v_{ij} denotes a weight matrix. Followed by, the diagonal matrix is constructed with the degrees $\mathrm{d}_1, \mathrm{d}_2, \mathrm{d}_3, \ldots . \mathrm{d}_\mathrm{n}$ on the diagonal as given below,

$$g_{ij} = \begin{bmatrix} \mathrm{d}_1 & & & \\ & \mathrm{d}_2 & & \\ & & \ddots & \\ & & & \mathrm{d}_\mathrm{n} \end{bmatrix} \quad (5)$$

Where, g_{ij}represent a diagonal matrix with a size of $4X\,4$. Then the normalized Laplacian matrix is constructed as given below,

$$\mathrm{L}_{\mathrm{ij}}\,(n) = \frac{1}{{g_{ij}}^{1/2}} \, k \, {g_{ij}}^{-1/2} \quad (6)$$

Where, $\mathrm{L}_{\mathrm{ij}}\,(n)$ denotes a normalized Laplacian matrix, $g_{ij}$is a diagonal matrix, k represents a kernel function to measure the similarity. Therefore, the normalized Laplacian matrix is constructed with the Eigen vectors 'v' and eigen values 'n'. Let us consider the matrix 'M_{ij}' whose columns are the eigenvectors equal to the 'n' smallest Eigen values. Therefore, the new matrix U_{ij} is constructed as given below,

$$U_{ij} = \frac{\mathrm{M}_{\mathrm{ij}}}{\sum \mathrm{M}_{\mathrm{ij}}} \, (7)$$

Where U_{ij} denotes a new matrix, rows of the matrix 'M_{ij}' as a collection of 'n' data and it grouped into 'k' number of clusters by applying the k-means algorithm. By applying the k-means algorithm, the 'k' number of clusters and centroid is initialized. Then the records are grouped by using the following equation,

$$Y = \arg\min \sum_{i=1}^{n} \sum_{j=1}^{n} \left\|r_i - c_j\right\|^2 \, (8)$$

Where Y represents the output of clustering, $arg\ min$ denotes argument of the minimum function to find the minimum distance between the records (r_i) and cluster centroid (c_j). $\left\|r_i - c_j\right\|^2$ is the squared distance between the records and cluster centroid. Similarly, the record which is closer to the centroid is grouped into cluster 'j' if and only if a row of the matrix is assigned to cluster j. In this way, all the records are grouped into the particular cluster in order to minimize the time and information loss of anonymization.

Data anonymization

The clustering results are transferred into the next hidden layer for data anonymization to protect the data. Data anonymization is applied for preserving the confidential or sensitive user information. This is also called as data masking. The proposed DLKSC-PRPSA technique uses the Pearson rank proximity swapping anonymization method for creating the anonymized data to preserve privacy. Pearson rank proximity swapping is a process of interchanging the neighborhood values of attributes across the records. Here the proximity refers to the neighborhood attribute values. In order to find the neighborhood, the correlation between the attribute values is measured. Based on the correlation, the attribute values are ranked and perform the swapping. The swapping based data anonymization provides the accurate results of data protection and minimizes information loss. The process of the Pearson rank proximity swapping anonymization method is shown in figure 3.

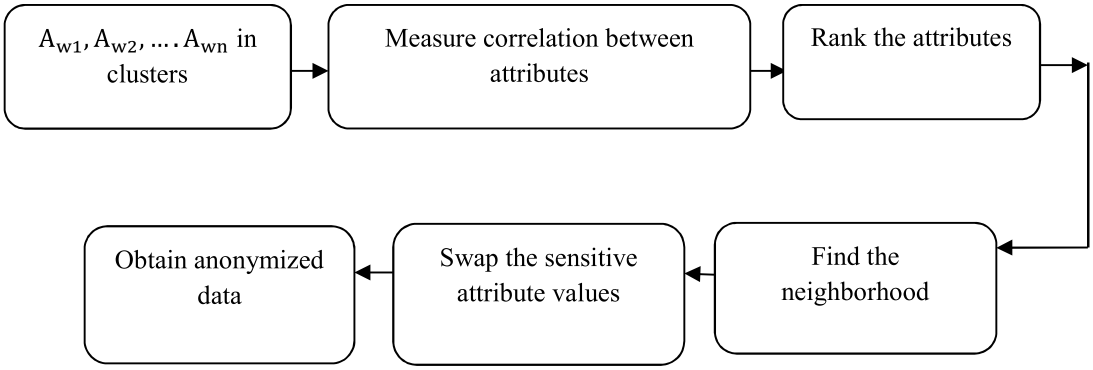
$A_{w1}, A_{w2}, \dots . A_{wn}$ in clusters
Measure correlation between attributes
Rank the attributes
Find the neighborhood
Swap the sensitive attribute values
Obtain anonymized data

Figure 3 Block Diagram Of Pearson Rank Proximity Swapping Anonymization

Figure 3 illustrates a block diagram of the Pearson rank proximity swapping anonymization. Let us consider the number of attribute values $A_{w1}, A_{w2}, \dots . A_{wn}$ in the 'j' number of clusters $c_1, c_2, \dots c_j$. The Pearson correlation between the attribute values are calculated as follows,

$$P_r = \frac{\sum A_{w1}A_{w2} - \frac{(\sum A_{w1})(\sum A_{w2})}{n}}{\sqrt{\left(\sum A_{w1}{}^2 - \frac{(\sum A_{w1})^2}{n}\right)\left(\sum A_{w2}{}^2 - \frac{(\sum A_{w2})^2}{n}\right)}} \; \dots\dots (9)$$

Where, P_r represents the Pearson correlation coefficient, n denotes the number of attributes, A_{w1}, A_{w2} are the two attribute values, $\sum A_{w1}A_{w2}$ refers to the sum of the product of paired score, $\sum A_{w1}$ is the sum of A_{w1}score, $\sum A_{w2}$ is the sum of A_{w2} score, $\sum A_{w1}{}^2$ is the sum of the squared score of A_{w1} and $\sum A_{w2}{}^2$ is the sum of the squared score of A_{w2}. The coefficient provides the correlation value between −1 and +1. Based on the correlation, sensitive attributes columns are first ordered i.e. ranked. After ranking, the highly correlated columns are considered as neighborhood. Then the swapping is carried out between the two neighborhood attribute values to preserve the privacy of the data.

Data swapping is a methodology is more effectively used for creating the anonymized data that also used to minimizing the information loss since it avoids the data deletion. Finally, the swapped results are obtained at the hidden layer,

$$h(t) = b_1 * r_i\ (t) + b_{\mathrm{h}} * h(t-1) \quad (10)$$

Where, $h(t)$denotes a hidden layer output at a time 't'. Here, '$h(t-1)$' represents a output of previous hidden layer and 'b_{h}' denotes a weights of the hidden layer, b_1 denotes a weight between input and hidden layer, $r_i\ (t)$represents the input records. In the output layer, the anonymized results are obtained. Let us consider the two sensitive attribute a and b are swapped independently as follows,

$$s(a,b) = y\ (t)\ =\ (b,a) \quad (11)$$

Where $y\,(t)$ denotes an output layer, $s(a,b)$ denotes a swapping of two attribute values in the cluster. This anonymized data only known to the controller. As a result, the sensitive attributes values within the cluster are protected and hence it improves the privacy preservation rate. The step by step process of proposed DLKSC-PRPSA technique is described as follows,

```
Input: Dataset, Number of records $r_1, r_2, r_3 \dots, r_n$
Output: Obtain higher privacy preservation rate
Begin
Given the input dataset into input layer $i(t)$
\\ first hidden layer
    Measure the radial basis kernel between two records $k\,(r_1, r_2)$
    Construct unnormalized Laplacian matrix $NL_{ij}$ with the diagonal matrix $g_{ij}$ and weight matrix $v_{ij}$
    Find the first 'k' eigenvectors
    Construct normalized Laplacian matrix $\mathrm{L_{ij}}\,(n)$
    Define '$k$' number of clusters
    for each cluster
        Initialize the mean value $\mu_j$
        Group data $r_i$ into clusters $j$ with $\arg\min \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|r_i - c_j\right\|^2$
    end for
\\ Second hidden layer
For each attribute value in clustering results
    Measure the correlation $P_r$
    Rank the attribute columns
    Find the neighborhood
    Swap the two attributes $s(a,b) = y\,(t)\ i.e\ (b,a)$
    Generate anonymized data at the output layer
end for
End
```

Algorithm 1 Deep Learned Kernel Spectral Clustering-based Pearson Rank Proximity Swapping Anonymization

Algorithm 1 describes the step by step process of deep learning-based privacy preservation for data publishing. The number of records in the dataset is considered as input and it was given to the input layer. Then the inputs are fed into the first hidden layer. In the hidden layer, the clustering is performed to group similar data for minimizing the anonymization time for privacy preservation. The clustering results are transformed into the second hidden layer. In the second hidden layer, the data anonymization is carried out to protect the sensitive attribute values by applying the Pearson correlative rank proximity swapping anonymization method. The two neighboring columns get interchanged and the intruder difficult to know the specific values of the variable. As a result, the data privacy preservation rate gets improved with lesser information loss.

3. EXPERIMENTAL SETTINGS

An experimental evaluation of the proposed DLKSC-PRPSA technique and existing methods RSA-SA approach [1] and SLPPA [2] are implemented in the Java language. The Adult Data Set (https://archive.ics.uci.edu/ml/datasets/Adult is used for conducting the experiments with 14 attributes. These attributes values are preserved for data publishing by swapping the column's values. The dataset includes 32561 instances. The objective of the dataset is to predict the person's income exceeds 50k per year. The attributes characteristics are an integer and categorical and dataset characteristics are multivariate. Totally ten different runs are carried out with the number of records (i.e. instances) taken in the range from 500 to 5000. The experimental evaluation is conducted with different performance metrics such as privacy preservation rate, anonymity level, time complexity, and Information loss.

4. RESULTS AND DISCUSSION

The experimental results of the proposed DLKSC-PRPSA technique and existing methods RSA-SA approach [1] and SLPPA [2] are discussed with certain parameters such as privacy preservation rate, anonymity level, information loss and time complexity based on the number of records taken from the adult dataset. The performances of the proposed technique against the existing methods are discussed using graphical representation.

4.1 Performance Results Of Privacy Preservation Rate

The privacy preservation rate is defined as the ratio of a number of records preserved from the others i.e. attackers to the total number of records taken as input. The formula for calculating the privacy preservation rate is given below,

$$R_{pp} = \left(\frac{N_P}{n}\right) * 100 \quad (12)$$

Where R_{pp} represents privacy preservation rate, n denotes a total number of records, N_P is the number of records preserved from the others. The privacy preservation rate is measured in the unit of percentage (%).

Mathematical calculation:

- **Existing RSA-SA:** Let us consider the total number of records is 500 and the records preserved from others are 400. The overall privacy preservation rate is estimated as follows,

$$R_{pp} = \left(\frac{400}{500}\right) * 100 = 80\%$$

- **Existing SLPPA:** Let us consider the total number of records is 500 and the records preserved from others are 410. The overall privacy preservation rate is estimated as follows,

$$R_{pp} = \left(\frac{410}{500}\right) * 100 = 82\%$$

- **Proposed DLKSC-PRPSA:** Let us consider the total number of records is 500 and the records preserved from others are 435. The overall privacy preservation rate is estimated as follows,

$$R_{pp} = \left(\frac{435}{500}\right) * 100 = 87\%$$

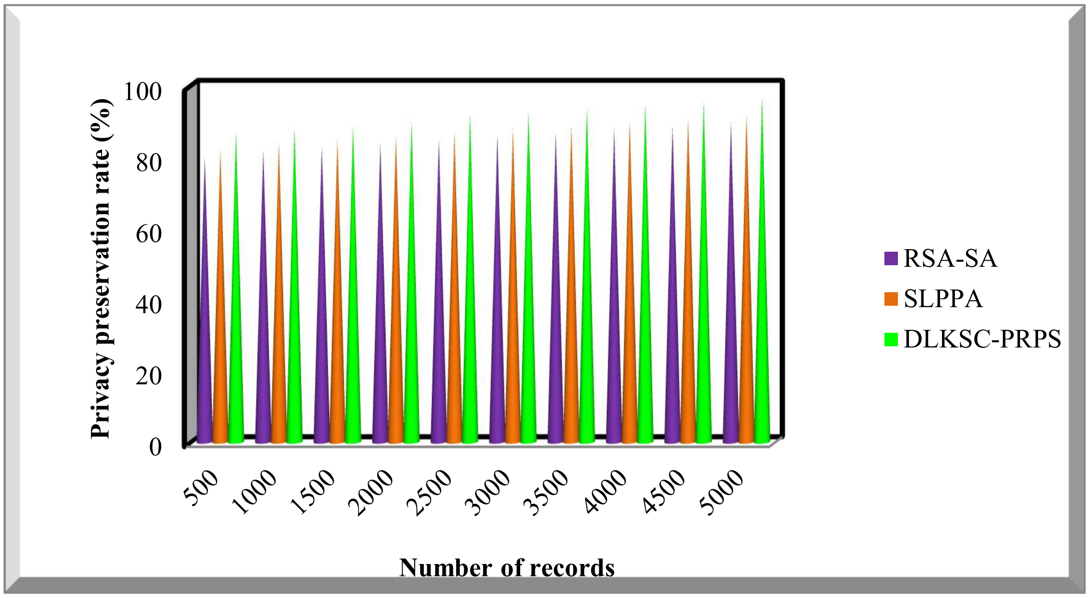

Figure 4 Privacy Preservation Rate Versus Number of Records

From the above figure, the performance results of data privacy preservation rate with the number of records taken in the range of 500-5000. The graphical results show that the privacy preservation rate of three methods namely the DLKSC-PRPSA technique and existing methods RSA-SA approach [1] and SLPPA [2] are obtained with three different colors namely green, violet and orange respectively. From the results, the data privacy preservation rate of DLKSC-PRPSA technique is higher than the other two existing methods. This significant improvement is achieved by applying the swapping anonymization method. The swapping is an anonymization method that effectively interchanging the attribute value in the records using the Pearson ranks proximity method for preserving the data privacy within the cluster. Based on the Pearson correlation coefficient value, the attribute values are interchanged and protect the original information from the others. This, in turn, achieves a higher privacy preservation rate.

.

With experimentation of '500' records, '435' records were preserved and the percentage is

87% whereas, 80% and 82% privacy preservation rate are obtained using the DLKSC-PRPSA, RSA-SA approach [1] and SLPPA [2]. Ten various results of the privacy preservation rate are obtained with respect to a number of records. The average comparison results evidently prove that the privacy preservation rate is found to be improved by 8% as compared to the RSA-SA approach [1] and 5% as compared to SLPPA [2].

4.2 Performance Results Of Anonymity Level

Anonymity level is measured as the ratio of the size of the record that maintained the anonymity to the size of the record. Mathematically the anonymity level is computed using the following equation,

$$At_l = \left(\frac{A\ size}{r\ size}\right) * 100 \quad (13)$$

Where At_l denotes an anonymity level, r_{size} represents a size of the records, $A_{\ size}$ represents the size of the record that maintained the anonymity. The anonymity level is measured in the unit of percentage (%).

Mathematical calculation:

- **Existing RSA-SA:** Let us consider the total size of the record is $100MB$ and the size of the record that has maintained anonymity is $76MB$. Then, the anonymity level is calculated as follows,

$$At_l = \left(\frac{76MB}{100MB}\right) * 100 = 76\%$$

- **Existing SLPPA:** Let us consider the total size of the record is $100MB$ and the size of the record that has maintained anonymity is $79MB$. Then, the anonymity level is calculated as follows,

$$At_l = \left(\frac{79MB}{100MB}\right) * 100 = 79\%$$

- **Proposed DLKSC-PRPS:** Let us consider the total size of the record is $100MB$ and the size of the record that has maintained anonymity is $82MB$. Then, the anonymity level is calculated as follows,

$$At_l = \left(\frac{82MB}{100MB}\right) * 100 = 82\%$$

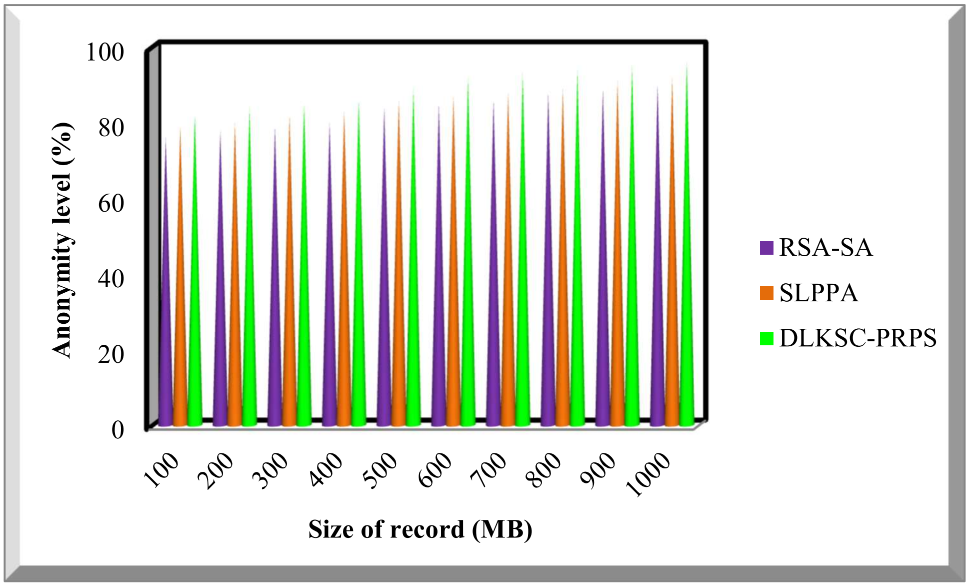

Figure 5 Anonymity Level Versus The Size of Records

Figure 5 given above illustrates the performance of the anonymity level under various sizes of records which ranged from 100 MB to 1000 MB. The graphical results clearly show that the Anonymity level of the proposed DLKSC-PRPS technique is found to be higher than the other two existing anonymization methods. The deep learning-based clustering and anonymization method is applied in the DLKSC-PRPS technique to maintain the privacy level. The clustering technique minimizes the information loss and preserves anonymized information of data publishing. The number of attribute values in each group maintains the anonymized information for protecting the original data from the attacker. Totally ten different results of anonymity level are obtained for three different methods. The results of the proposed technique are compared to the existing methods. The average of comparison results evidently proves that the anonymity level of the DLKSC-PRPS technique is said to be improved by 7% and 5% when compared to the existing RSA-SA approach [1] and SLPPA [2] respectively.

4.3 Performance Results Of Time Complexity

The time complexity is defined as an amount of time taken by the algorithm to anonymize the given record for privacy-preserving data publishing. The formula for calculating time complexity is given below,

$$C_{time} = Number\ of\ records * T\ (\ anonymize\ one\ r) \quad (14)$$

Where C_{time} represents the time complexity, 'T' denotes a time taken to anonymized one record r. The time complexity is measured in terms of a millisecond (ms).

Mathematical calculation:

- **Existing RSA-SA:** Let us consider the number of records is 500 and the time for anonymizing one record is $0.045ms$. Therefore the overall time complexity is measured as follows,

$$C_{time} = 500 * 0.045ms$$
$$= 22.5ms \sim 23ms$$

- **Existing SLPPA:** Let us consider the number of records is 500 and the time for anonymizing one record is $0.042ms$. Therefore the overall time complexity is measured as follows,

$$C_{time} = 500 * 0.042ms = 21ms$$

- **Proposed DLKSC-PRPS:** Let us consider the number of records is 500 and the time for anonymizing one record is $0.036ms$. Therefore the overall time complexity is measured as follows,

$$C_{time} = 500 * 0.036ms = 18ms$$

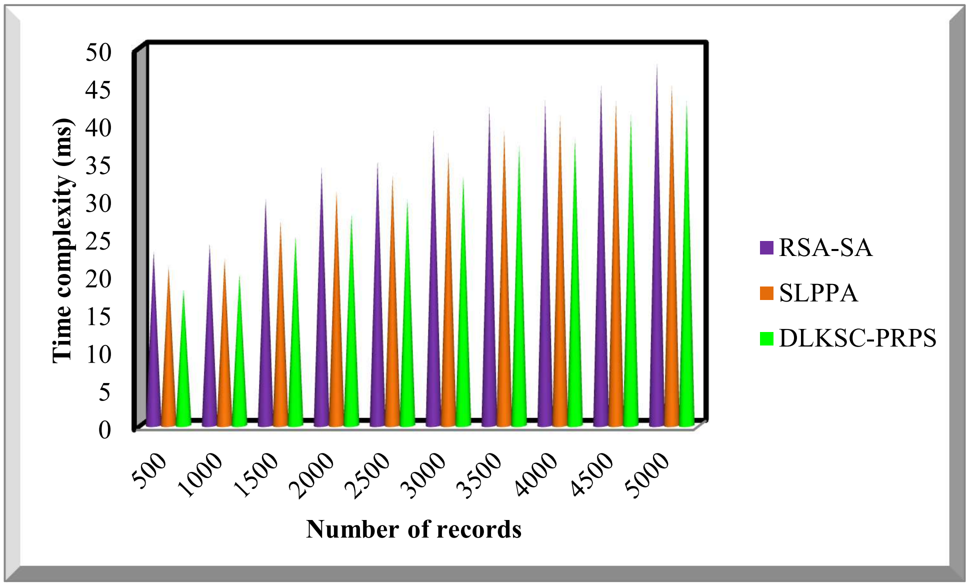

Figure 6 Time Complexity Versus The Number of Records

Figure 6 depicts the comparison of time complexity versus a number of records. From the graphical results, the numbers of records are given as input to the horizontal axis, the performance results of time complexity are obtained at the vertical axis. The comparison of the time complexity is minimized using the DLKSC-PRPS technique. This is because of applying the radial basis kernelized spectral clustering at the hidden layer. The deep learning approach effectively performs the clustering of input records into different groups. The input layer receives the dataset for data anonymization. While processing the whole dataset, the technique takes more time to anonymize the data. Therefore, the DLKSC-PRPS technique initially performs the clustering process in the hidden layers. The clustering algorithm groups similar data. Followed by, the rank swapping anonymization method is applied to protect the data. The comparison result of three methods evidently proves that the DLKSC-PRPS technique minimizes the time complexity by 15% and 8% as compared to two state-of-the-art methods.

4.4 Performance Results Of Information Loss

Information loss of anonymization is measured as the ratio of the difference between the total number of records taken as input and the number of records received with higher privacy to the total number of records. The information loss is mathematically computed as follows.

$$Loss_{inf} = \frac{(n - n_{RP})}{Number\ of\ records} * 100 \quad (15)$$

Where $Loss_{inf}$ represents an Information loss, n denotes a total number of records, n_{RP} denotes a number of records received with higher privacy. The information loss is measured in terms of percentage (%).

Mathematical calculation:

- **Existing RSA-SA:** Let us taken as a total number of records is 500 and the number of records received with higher privacy is 400. Then the Information loss is mathematically calculated as follows,

$$Loss_{inf} = \left(\frac{500-400}{500}\right) * 100 = 20\%$$

- **Existing SLPPA:** Let us taken as a total number of records is 500 and the number of records received with higher privacy is 410. Then the Information loss is mathematically calculated as follows,

$$Loss_{inf} = \left(\frac{500-410}{500}\right) * 100 = 18\%$$

- **Proposed DLKSC-PRPS:** Let us taken as a total number of records is 500 and the number of records received with higher privacy is 435. Then the Information loss is mathematically calculated as follows,

$$Loss_{inf} = \left(\frac{500-4}{500}\right) * 100 = 13\%$$

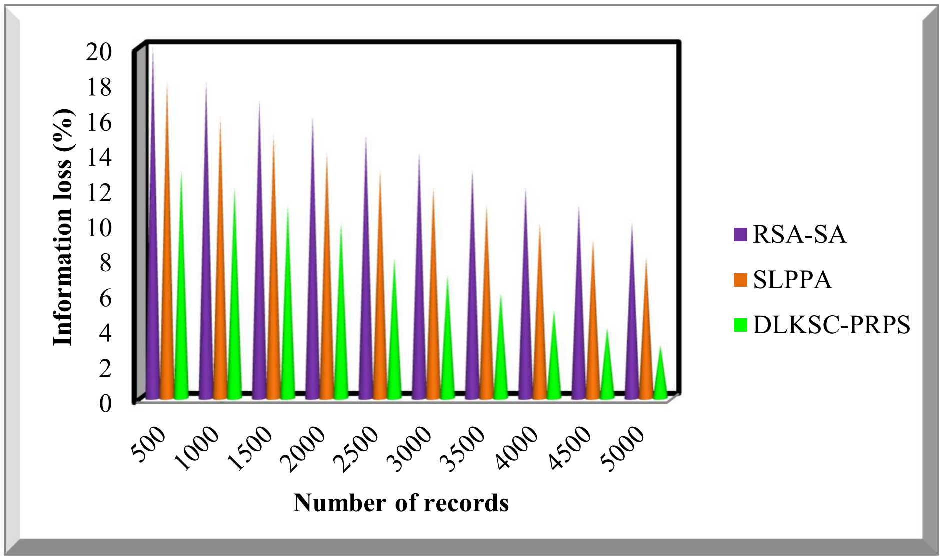

Figure 7 Information Loss Versus Number of Records

The experimental results of information loss versus a number of records are shown in figure 7. The figure clearly depicts the information loss is found to be minimized using the DLKSC-PRPS technique. The clustering process before the data anonymization minimizes the information loss. In addition, the rank swapping anonymization measures the Pearson correlation between the attribute values. Rank swapping provides accurate anonymization results about the trade-off between information loss and data protection. It also useful for protecting the multiple sensitive attributes values, therefore, it effectively anonymized the original information and minimizes the information loss. The mathematical evaluation shows that the performance of information loss using DLKSC-PRPS is minimized. The comparison results show that the information loss in data anonymization is reduced using the DLKSC-PRPS technique by 48% compared to [1] and 40% compared to [2].

The above-discussed results of the various performance metrics clearly show that the DLKSC-PRPS technique effectively improves the privacy preservation of data publishing with minimum time and minimum information loss.

5. RELATED WORKS

Privacy is a significant concern in the publication of datasets that includes sensitive information. The several methods have been developed for preventing privacy and offering useful information to legal users.

A conditional probability distribution and machine learning-based data privacy-preserving approach were developed in [11]. But the approach failed to use the anonymization method for achieving a higher privacy preservation rate. A Privacy-Preserving Tabular Data Publishing (PPTDP) with the data anonymization approach was introduced in [12] to provide accurate protection. However, the performance of information loss was not reduced.

A privacy-preserving content-based publish/subscribe (PCP) method was developed in [13] to guarantee the differential privacy and security. But the method failed to use the clustering-based privacy-preservation to minimize the processing time. A k-anonymity technique was introduced in [14] for ensuring the privacy of big data and balancing the data utility. The technique failed to calculate its complexity of data anonymization.

A customizable and continuous privacy-preserving social media data publishing method was introduced in [15] for efficiently protecting the user data. The method failed to consider more records with continuous values. A novel anonymization method was introduced in [16] for improving the privacy and anonymous data utility during the e-health data publishing. The designed method failed to use more sensitive attributes for data anonymization.

A new on-line spatial-temporal k-anonymity model was developed in [17] for protecting the data from the attacks. The method failed to resist the attacks using large-scale data. A privacy-aware structural data publishing approach was introduced in [18] for protecting the data from the attacks. But the data privacy protection with multiple attributes was not performed.

A graph-based multifold method was introduced in [19] to perform data anonymization with attributes of various types. The clustering-based fuzzy algorithm was applied for protecting the data and minimizing the time complexity but the anonymity level was not improved. A Merging method was developed in [20] using l-diversity privacy requirements to preserve data privacy before the publication. Though the method reduces the information loss, the time complexity was not minimized.

6. CONCLUSION

An efficient deep learning technique called DLKSC-PRPS is introduced for enhancing the privacy preservation of data publishing. In DLKSC-PRPS technique, mainly focused on grouping the records into the several clusters for data anonymization using the radial basis kernelized spectral clustering technique. After grouping the records, the data anonymization is carried out by applying the rank swapping based on the Pearson correlation measure. Based on the swapping of sensitive attribute values, the original information is preserved and securely publishing the data. As a result, the deep learning approach efficiently performs the data anonymization with minimum time. The performance of the DLKSC-PRPS technique is tested with parameters such as privacy preservation rate, anonymity level, information loss and time complexity using the adult dataset. The experimental results demonstrate that the DLKSC-PRPS technique provides better performance in terms of privacy preservation rate, anonymity level and minimizes the information loss as well as time complexity when compared to state-of-the-art works.

References

[1] Saad A. Abdelhameed, Sherin M. Moussa, Mohamed E. Khalifa, "Restricted Sensitive Attributes-based Sequential Anonymization (RSA-SA) approach for privacy-preserving data stream publishing", Knowledge-Based Systems, Elsevier, Volume 164, 2019, Pages 1-20

[2] Lin Yao, Zhenyu Chen, Xin Wang, Dong Liu, Guowei Wu, "Sensitive Label Privacy Preservation with Anatomization for Data Publishing", IEEE Transactions on Dependable and Secure Computing, 2019, Pages 1-14

[3] Zhipeng Cai, Zaobo He, Xin Guan, Yingshu Li, "Collective Data-Sanitization for Preventing Sensitive Information Inference Attacks in Social Networks", IEEE Transactions on Dependable and Secure Computing, Volume 15, Issue 4, 2018, Pages 577 – 590

[4] Xuening Chen, Hanwen Liu, Dan Yang, "Improved LSH for privacy-aware and robust recommender system with sparse data in edge environment", EURASIP Journal on Wireless Communications and Networking, Springer, 2019, Pages 1-11

[5] Zaobo He, Zhipeng Cai, Jiguo Yu, "Latent-Data Privacy-Preserving With Customized Data Utility for Social Network Data", IEEE Transactions on Vehicular Technology, Volume 67, Issue 1, 2018, Pages 665 – 673

[6] Chao Yan, Xinchun Cui, Lianyong Qi, Xiaolong Xu, Xuyun Zhang, "Privacy-Aware Data Publishing and Integration for Collaborative Service Recommendation", IEEE Access, Volume 6, 2018, Pages 43021 – 43028

[7] Xiangwen Liu, Qingqing Xie, Liangmin Wang, "Personalized extended (α, k)-anonymity model for privacy-preserving data publishing", Concurrency and Computation: Practice and Experience, 2016, Volume 29, Issue 6, Pages 1-18

[8] Jong Wook Kim, Kennedy Edemacu, Beakcheol Jang, "MPPDS: Multilevel Privacy-Preserving Data Sharing in a Collaborative eHealth System", IEEE Access, Volume 7, 2019, Pages 109910 – 109923

[9] Rana Elgendy, Amr Morad, Hicham G. Elmongui, Ayman Khalafallah, Mohamed S. Abougabal, "Role-task conditional-purpose policy model for privacy-preserving data publishing" Alexandria Engineering Journal, Elsevier, Volume 56, 2017, Pages 459–468

[10] M.A.P. Chamikara, P. Bertok, D. Liu, S. Camtepe, I. Khalil, "An Efficient and Scalable Privacy-Preserving Algorithm for Big Data and Data Streams", Computers & Security, Elsevier, Volume 87, 2019, Pages 1-52

[11] Chaobin Liu, Shixi Chen, Shuigeng Zhou, Jihong Guan, Yao Ma, "A novel privacy-preserving method for data publication", Information Sciences, Elsevier, Volume 501, 2019, Pages 421–435

[12] Saad A. Abdelhameed, Sherin M. Moussa, Mohamed E. Khalifa, "Privacy-preserving tabular data publishing: A comprehensive evaluation from the web to cloud", Computers & Security, Elsevier, Volume 72, 2018, Pages 74-95

[13] Qixu Wang, Dajiang Chen, Ning Zhang, Zhe Ding, Zhiguang Qin, "PCP: A Privacy-Preserving Content-Based Publish-Subscribe Scheme With Differential Privacy in Fog Computing", IEEE Access, Volume 5, 2017, Pages 17962 – 17974

[14] Zakariae El Ouazzani and Hanan El Bakkali, "A new technique ensuring privacy in big data: K-anonymity without the prior value of the threshold k", Procedia Computer Science, Elsevier, Volume 127, 2018, Pages 52–59

[15] Dingqi Yang, Bingqing Qu, Philippe Cudré-Mauroux, "Privacy-Preserving Social Media Data Publishing for Personalized Ranking-Based Recommendation", IEEE Transactions on Knowledge and Data Engineering, Volume 31, Issue 3, 2019, Pages 507 – 520

[16]Abdul Majeed, "Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data", Journal of King Saud University - Computer and Information Sciences, Volume 31, Issue 4, 2019, Pages 426-435

[17] Haitao Zhang, Chenxue Wu, Zewei Chen, Zhao Liu, Yunhong Zhu, "A novel on-line spatial-temporal k-anonymity method for location privacy protection from sequence rules-based inference attacks", PLoS ONE, Volume 12, Issue 8, Pages 1-32

[18] Xuangou Wu, Panlong Yang, Shaojie Tang, Xiao Zheng, Xiaolin Wang, "Privacy-aware data publishing against sparse estimation attack", Journal of Network and Computer

Applications, Elsevier, Volume 109, 2018, Pages 78-88

[19] Li-E. Wang and Xianxian Li, “A graph-based multifold model for anonymizing data with attributes of multiple types”, computers & security, Elsevier, Volume 72, 2018, Pages 122–135

[20] A S M Touhidul Hasan, Qingshan Jiang, Hui Chen and Shengrui Wang, “A New Approach to Privacy-Preserving Multiple Independent Data Publishing”, Applied Science, Volume 8, Issue 5, 2018, Pages 1-22