Received: July 8, 2018

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Hybrid Clustering Approach and Random Rotation Perturbation (RRP) for Privacy Preserving Data Mining

Sivakumar Kaliappan[1*]

[1]*Department of Mathematics, Sathyabama Institute of Science and Technology, Chennai, India.*
Corresponding author's Email: shivakumar.maths@sathyabama.ac.in

Abstract: The Privacy Preserving Data Mining (PPDM) is known to be the most critical perspective among analysts. As privacy preserving data mining grants, sharing and exchanging of privacy susceptible data for analysis, it has exploited increasingly popular. Since one of the critical aspects of data mining is safeguarding privacy. The diverse technique is embraced for preserving privacy while maintaining the real characteristic of data under consideration. In the proposed work, the high-dimensional data are isolated into various parts by utilizing the k-mean clustering technique and each partition is considered as a cluster. By then the mean estimate of each cluster is processed, after that, the contrast between each cluster member and the mean of the cluster esteem is processed. In the succeeding stage, the clustered information is enhanced by utilizing the Ant Colony Optimization (ACO) algorithm. After that, the optimized clustered particles are perturbed by utilizing the Random Rotation Perturbation (RRP) algorithm which makes the values hard to be recognized. These perturbed values are then stored in the public cloud and the key parameters for randomizing and the clustering is stored in the private cloud. Our approach would contribute in the diminishment of a lot of storage in a private cloud, in case we essentially store the entire sensitive information on private clouds. The experimental results demonstrate that the RRP algorithm has better privacy preserving contrasted with the other existing technique.

Keywords: Data mining, Privacy-preserving, K-mean clustering, Ant colony optimization (ACO), RRP (Random Rotation Perturbation).

1. Introduction

Data mining is the computing process of discovering patterns in extensive dataset including techniques at the intersection of machine learning, statistics, and database systems [1, 2]. The overall objective of the data mining process is to remove information from a data set and transform it into an understandable structure for further use. The privacy-preserving data mining is known to be the most important aspect among researchers [3]. As of late, signal processing in the encrypted area has pulled in light of a legitimate concern for a few scientists. This is especially legitimate in the utilization of cloud computing and appointed computation, where data proprietors need to reveal or give extraordinary data to remote servers for data processing [4]. Information proprietors won't trust these customers or organization suppliers, so a privacy-preserving instrument was utilized in the first data. This sort of study is called PPDM [5]

As privacy preserving data mining grants, sharing and exchanging of privacy susceptible data for analysis, it has developed increasingly famous. Since one of the essential aspects of data mining is safeguarding privacy [6]. The primary objective of privacy-preserving is to shroud the sensitive data before it gets distributed. For instance, a healing center may discharge patient's records to empower the scientists to think about the attributes of various illnesses. The raw data contains any delicate information of people, which are not distributed to ensure individual privacy [7]. However, utilizing some other published attributes and external data,

we can retrieve the individual personalities. There are various strategies, for example, k-means clustering [8], fuzzy base logic [9], neural networks, K-nearest neighbor (KNN) [10] clustering procedures are utilized for preserving the privacy of the data while clustering. A portion of the techniques is the utilization of cryptographic algorithms, noise addition, and data swapping [11].

The prior version of KNN [12] is comparatively moderate, mainly because it has to compute the distances of the inquiry to all data. Notwithstanding, in this technique, the computational complexity is high in big data [13]. Another strategy, Fuzzy base logic [14] is the errand of partitioning a feature space into fuzzy classes. Two strategies are utilized in Neuro-fuzzy classifier [9, 15], such as a feature subset selection and linear discriminate analysis and these techniques are utilized to assess the vital feature subsets and consequently reestablish the attributes of the data distribution in the feature space for training the Neuro-fuzzy classifier [16, 17]. The main drawback of this strategy is utilized by numerous fuzzy rules. In the fuzzy neural classifier algorithm, larger size data sets are utilized and that brought about feeble recognition of data and subsequently resulted in expanded cost and time.

The proposed strategy utilized the hybrid clustering and random rotation perturbation approach for privacy-preserving data mining. Here, initially, the k-mean clustering algorithm is utilized for partitioning the high-dimensional data, each partitioning is considered as a cluster. The ACO algorithm is utilized to select the best-clustered group based on the output of the k-means clustering. The RRP algorithm is used to perturb the clustering data, these perturbed values are difficult to be recognized. The public cloud store the perturbed data and the key parameters for randomizing technique and the clustering techniques are stored in a private cloud. At that point, the data recovery technique is utilized to recover the original data from the perturbed data.

Whatever remains of the segment of the paper is portrayed underneath. In segment 3, the data privacy preservation is depicted, in segment 3.1, the clustering technique is explained. The ACO algorithm for optimization is portrayed in section 3.2. The RRP algorithm is delineated in section 3.3. The test outcome and the conclusion are inspected in section 4 and 5.

2. Related work

Xindong Wu et al. [18] have examined big data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast advancement of networking, data storage, and the data collection capacity, big data were quickly extending in all science and engineering domains, including physical, biological and biomedical sciences. Their article has presented a HACE theorem that portrays the features of the Big Data revolution, and executed a Big Data processing model, from the data mining perspective. D. Adeniyi et al. [19] have introduced an automatic web usage data mining and recommendation system based on current user behavior through his/her clickstream data on the newly developed Really Simple Syndication (RSS) reader website, in order to provide relevant information to the individual without expressly requesting for it. The K-Nearest-Neighbor (KNN) classification strategy has been trained to be utilized online and in real-time to identify clients/visitors clickstream data, matching it to a particular user group and recommend a tailored browsing option that meets the need of the specific user at a particular time. To achieve this, the web user's RSS address file was extracted, cleansed, formatted and grouped into meaningful session and data mart was produced.

A new algorithm based on data perturbation and query restriction (DPQR) has proposed by H. Lou et al. [20] to enhance the privacy-preserving degree by multi-parameters perturbation. In order to enhance the time-efficiency, the calculation to obtain an inverse matrix was improved by isolating the matrix into blocks; in the interim, a further optimization was given to lessen the number of scanning databases by set theory. Both theoretical analyses and experimental results demonstrate that the proposed DPQR algorithm has better performance. A novel hiding-missing-artificial utility (HMAU) algorithm has proposed by C. Lin et al. [21] to conceal sensitive item sets through transaction cancellation. The transaction with the maximal ratio of sensible to nonsensitive one was thus selected to be entirely erased. Three side effects of hiding failures, missing item sets, and artificial item sets are considered to assess whether the transactions are required to be erased for concealing sensitive item sets. Three weights are additionally allotted as the significance to three factors, which can be set by the necessity of clients.

Lei Xu et al.[22] have seen the privacy issues related to data mining from a wider perspective and investigate various methodologies that can ensure to protect sensitive information. Specifically, they have recognized four distinct types of users associated with data mining applications, namely, data provider, data collector, data miner, and decision maker. For

each type of user, they have discussed privacy concerns and the methods that can be adopted to protect sensitive information. Besides exploring the privacy-preserving approaches for each type of user, they also reviewed the game theoretical approaches, which are proposed for analyzing the interactions among different users in a data mining scenario, each of whom has his own valuation on the sensitive information.

3. Proposed method

In this paper, we plan a combined clustering and Random Rotation Perturbation algorithm for improving privacy preservation of various organizational data in the cloud. Initially, the high dimensional data are converted into sensitive attributes and non-sensitive attributes. Here, the sensitive attributes are generally leaked and not properly secured in the cloud. So, in our work, higher dimensional sensitive attributes are partitioned by using the K-means clustering algorithm. After that, the clustered particles are optimized by using the ACO algorithm to select the best-clustered particle from that group. Finally, the data perturbation process is utilized to perturb the optimized clustered particle; it's hard to be recognized by others. The perturbation process is done based on the RRP approach. At last the data retrieval phase is utilized to retrieve the original data from the perturbed data using the key parameters.

3.1 Privacy-preserving K-means clustering

In this section, the proposed model is described briefly and the solution for the K-means clustering, optimization for clustering and the data perturbation approach is described in the privacy preserving data mining.

3.1.1. Our model

Consider the application of a data mining procedure on a $D = \{d_1, d_2, \ldots, d_n\}$ dataset includes different attributes for the same set of entities; wish to cluster their joint data using the K-means clustering algorithm. Initially, consider the common entities are $a_1, a_2, \ldots, a_n$. Here, all the entities a_1 can be viewed as a point $a_{i,1}, a_{i,2}, \ldots a_{i,s}$ in s-dimensional discredited space, where $a_{i,j} \in y_p$ for some large p. In the dataset, let us assume the data set d_1 includes $a_{i,1}, a_{i,2}, \ldots a_{i,s}$ for all i, the data set d_2 includes $a_{i,s1+1}, a_{i,s1+2}, \ldots a_{i,s2}$ for all i, the last data includes $a_{i,sD\text{-}1}+1, a_{i,sD\text{-}2}+2, \ldots a_{i,sD}$ for all i, where $S_D = S$.

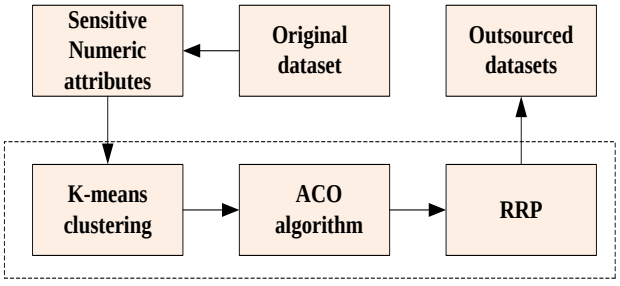

Figure. 1 Model of the proposed method

Generally, the object of the dataset d_1 is represented in the vector format and it is given as,

$$[D_{i1}, D_{i2}, \cdots, D_{im}] \quad (1)$$

The proposed method utilized the k-means clustering algorithm for partitioning the high dimensional dataset. The aim of the K-means clustering algorithm is, initially all the dataset D is converted into a number of k clusters and the cluster centers are randomly selected from the randomly partitioning dataset by minimizing the following cost function.

$$C = \{c_1, c_2, \cdots, c_k\} \quad (2)$$

The cost function is represented as follows,

$$cost(C_i) = \sum_{i=1}^{k} \sum_{d_i \in c_i} (d_i - c_i)^2 \quad (3)$$

In the above equation, the mean of the points in the cluster is represented as c_i and d_i is represented as the dataset. After that considers each information indicates and relegates it to the cluster, which is nearest and updates the cluster center,

$$c_i = \sum_{d_i \in C_j} d_i / |C_i| \quad (4)$$

In the above equation, the cluster point is represented as d_i, the total number of points is represented as C_i and the initial centroid between the cluster points is represented as C_j . Recalculate cluster center by discovering the mean of data points having a place with a similar cluster and the new cluster center is computed as follows,

$$c_k = \frac{1}{|G_k|} \sum_{d_i \in G_k} D_i \quad (5)$$

In the above equation, the cluster G_k contains the set of points d_i that are nearest to the center c_k,

$$G_k = \{d_i / C_i = k\} \quad (6)$$

Then backpedal to and it is computed as follows, the past two stages and rehash the procedure until the group part at no time in the future changes or at a most extreme number of emphases is reached. Finally, the accurate clustering groups are produced by using the cluster center c_k.

Algorithm 1: Privacy-Preserving K-means clustering (PPKM)

Input : $\boldsymbol{D} = \{ d_1, d_2, \dots, d_n \} \in R^X$ ($N \times X$ *input data set*)

Output : $\boldsymbol{D} = \{ c_1, c_2, \dots, c_k \} \in R^X$ (K cluster centers)

Select a random subset C of D as the initial set of cluster centers;

while termination criterion is not met **do**

for ($i = 1$; $i \leq N$; $i = i + 1$) **do**

Assign d_i to the nearest cluster;

Compute cost function which is in eqn. 3

end

Recalculate the cluster centers;

for ($k = 1$; $k \leq K$; $k = k + 1$) **do**

Cluster G_k contains the set of points d_1 that are nearest to the center c_k;

$G_k = \{d_i \mid C_i = k\}$;

Calculate the new cluster center c_k as the mean of the points that belong to c_k;

end

end

3.2 Optimization based on ACO algorithm

In this section, an efficient and well known ACO algorithm is utilized for an efficient and robust optimization process which is a metaheuristic algorithm. In our proposed method ACO algorithm is used for the optimized the clustering data. The ACO algorithm is utilized to solve the hard computational problem in the reasonable amount of time. The foraging behavior of the real ants is the motivating source of the ACO algorithm. The procedure of the ACO algorithm is given as follows,

1) At first, the ants randomly explore the zone surrounding their nest at the time of searching the food.
2) In the wake of finding, the food source, the ant tests the quality and quantity of each food and carries a few its back to the nests.
3) Amid the return trip, a chemical pheromone is generated by the ant and these pheromones are saved on the ground.
4) For the food source, alternate ants are guided by the light of the pheromone quantity. These pheromone depositions are used to assess the shortest path between the food source and their nests. The means of finding the optimized test cases are portrayed as takes after.

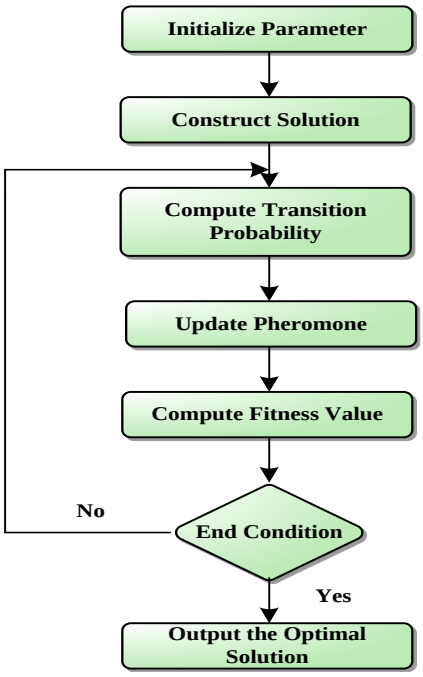

Figure. 2 Flow diagram of the ACO algorithm

In the ACO implementation, initially the candidate solutions $\{d_1, d_2, \dots, d_n\}$ are generated from the initial data set D and it is shown in the clustering scheme (section 3.1.1). Then the cluster center c_k is computed for each individual to know the fitness and this estimation is shown in the section 3.1.1. At first, every ant discovers the clustered particles $C = \{ c_1, c_2, \dots c_n \}$ randomly. Here, each cluster particle has the number of cluster members $c_1 = \{c_{1,1}, c_{1,2}, \dots c_{1,n}\}$ and $c_2 = \{c_{2,1}, c_{2,2}, \dots c_{2,n}\}$ that has built a solution in the iteration itself. The pheromone α_{ij} associated with the edge joining cities i and j, is updated as follows:

$$\alpha_{ij}(t+1) = (1-\rho).\alpha_{ij}(t) + \Delta\alpha_{ij}(t) \tag{7}$$

In the above equation, the evaporation rate is represented as ρ. To form the heuristic rule of the probabilistic transition, the pheromones levels are combined and it is given as follows,

$$p_{ij}^{c_1}(t) = \frac{\alpha_{ij}^{c_1}(t)\beta_{ij}^{c_2}}{\sum_{i,j=1}^{n}\alpha_{ij}^{c_1}(t)\beta_{ij}^{c_2}} \tag{8}$$

In the above equation, to control the importance of the pheromone value and heuristic information, the parameters $c_1 > 0$ and $c_2 > 0$ are utilized. The

amount of pheromone on edge (*i*,*j*) is represented as α_{ij}, the desirability of edge (*i*,*j*) is represented as β_{ij}. The ants use the pheromone concentration to mark their way, between the nest and a source of food. A colony is thus able to choose the shortest way towards a source to exploit without having a global vision of the way.

In the above-mentioned equation, if the pheromone evaporation ρ is constant, the pheromone concentration given by the exponential form as:

$$\alpha(t) = \alpha_0(t)e^{-\gamma} \tag{9}$$

In the above equation, the initial concentration of pheromone is represented as α_0 and the time is represented as *t*. The values of pheromones are updated by using the Eq. (1), usually from the Eq. (1),

$$\Delta\alpha_{ij}(t) = 1/L \tag{10}$$

After that, the ants add pheromone to all chosen edges, the additional pheromone by each edge (*i*, *j*) is given as follows,

$$\Delta\alpha_{ij}(t) = \sum_{c_1=1}^{n} \alpha_{ij}^{c_1}(t) \tag{11}$$

In the above equation, α_{ij} are the number of cluster members (ants) and the quantity of pheromone laid on each point (*i*, *j*) by ant c_1. The additional pheromone laid on the point (*i* ,*j*) by *c - th* ant at the end of the iteration *t*.

The individual pheromone addition contributed by each ant is given by,

$$\alpha_{ij}^{c_1}(t) = \begin{cases} Q/J_{c_1}\,(i,j)\; c^{th}\; ant\; use\; edge\; (i,j) \in g_b \\ 0 \qquad\qquad otherwise \end{cases} \tag{12}$$

Where g_b denotes global best solution. The ants add pheromone to the edges that they select and that solutions of better quality are rewarded with greater pheromone additions. Therefore, the fitness function of each member of the cluster group *C* is computed as follows,

$$\min J = \frac{1}{N}\sum_{i=1}^{N} \|y_1 = y_2\|^2 \tag{13}$$

The length of data used for cluster estimation is represented as *N*, y_1 and y_2 is a state variable of the original and the estimated cluster. Here, the fitness value is computed for each cluster group and finally, which clustered particle has the minimum fitness value is considered as the best cluster group *K*.

3.3 Random rotation perturbation for privacy preserving

Data perturbation is an imperative system for privacy-preserving the data. To augmenting both information privacy and information utility accomplished is the extraordinary goal for all data perturbation technique. After the k-mean clustering, the clustered result values are optimized by using the ACO algorithm. The random data perturbation technique is connected to the beforehand figured esteems which make the qualities difficult to be perceived. In our paper, a Random Rotation Perturbation (RRP) technique is used for randomizing the previously computed cluster values. The data utility and privacy guarantee are well preserved by the RRP algorithm.

3.3.1. Data perturbation phase

In the RRP method, initially, let as considered the optimized clustered dataset *K* and form the $n \times m$ matrix *M*. Initially, the confidential numerical attributes are selected and then the random rotation matrix is randomly and independently generated by using this confidential numerical attribute. After that, the rotation transformation is applied to get the transformed *data* $P = M \times R$. This distorted dataset is released for clustering analysis. The privacy of individuals is protected by using this distorted dataset and also achieves high accuracy.

Algorithm 2: Data Perturbation Phase

Input: Perturbed data set $\boldsymbol{D}'$
Output: Clustering results *k*and k' of data set $\boldsymbol{D}$ and $\boldsymbol{D}'$respectively.
Step 1: Given the input dataset *D* , its tuple estimate n, and the relating sensitive attribute $[\boldsymbol{F}]_{n\times 1}$
Step 2: Sensitive attribute $[\boldsymbol{F}]_{n\times 1}$is rotated in 180° clockwise course, so the random rotation matrix is generated.
Step 3: The result of $[\boldsymbol{T}]_{n\times 1}$and $[\boldsymbol{F}]_{n\times 1}$ is obtained in step 3. The duplicated esteems will be, $[\mathbf{X}]_{n\times 1} = [\boldsymbol{T}]_{n\times 1} \times [\boldsymbol{F}]_{n\times 1}$.
Step 4: Now the perturbation data is $\boldsymbol{RRP(D)} = [\mathbf{X}]_{n\times 1} = [\boldsymbol{T}]_{n\times 1} \times [\boldsymbol{F}]_{n\times 1}$.
Step 5: The perturbed dataset $D'$is created by supplanting attributes $[\boldsymbol{F}]_{n\times 1}$in the original dataset *D* with $[\boldsymbol{RRP\ (P)}]_{n\times 1}$
Step 6: Apply k- means clustering and ACO algorithm with various estimation of [***k***] on the original dataset [***D***] having the sensitive attribute [***F***].

Step 7: Apply k means clustering and ACO algorithm with various estimation of k on perturbed dataset $\boldsymbol{D}'$ having transformed sensitive attribute $\boldsymbol{RRP(D)}$
Step 8: Create a cluster membership matrix of results from step 6 and step 8 and dissect.

Algorithm 3: Data Perturbation Phase

Input: Perturbed data $\boldsymbol{D}'$, sensitive attribute $[\boldsymbol{F}]$.
Intermediate result: Random Rotation perturbation of sensitive attributes $\boldsymbol{RRP(D)}$.
Output: Original clustering data $\boldsymbol{D}$ of the perturbed data $\boldsymbol{D}'$
Step 1: Given the input dataset D', its tuple estimate n, and the relating sensitive attribute $[\boldsymbol{F}]_{n\times 1}$
Step 2: Sensitive attribute $[\boldsymbol{F}]_{n\times 1}$ is rotated in a 180° anti-clockwise direction, so the random rotation matrix $[\boldsymbol{T}']_{n\times 1}$ is generated.
Step 3: The result of $[\boldsymbol{T}']_{n\times 1}$ and $[\boldsymbol{F}]_{n\times 1}$ is obtained in step 3. The duplicated esteems will be, $[\mathbf{X}']_{n\times 1} = [\boldsymbol{T}']_{n\times 1} \times [\boldsymbol{F}]_{n\times 1}$.
Step 4: Now the perturbation data is $\boldsymbol{RRP(D)}' = [\mathbf{X}']_{n\times 1} = [\boldsymbol{T}']_{n\times 1} \times [\boldsymbol{F}]_{n\times 1}$.
Step 5: The perturbed dataset D' is created by supplanting attributes $[\boldsymbol{F}]_{n\times 1}$ in the original dataset D with $[\boldsymbol{RRP(P)}]_{n\times 1}$
Step 6: The original dataset D is created by supplanting attributes $[\boldsymbol{F}]_{n\times 1}$ in a perturbed dataset $\boldsymbol{pD}'$ with $[\boldsymbol{RRP(P)}]'_{n\times 1}$

3.3.2. Data recovering phase

Data recovering is very important in the information technology. In our system, the healthcare database owner gives permission to the three types of users to accept the original database. But they do not get the all original data about the patients, they only get certain data based on the user's requirements and all other data are perturbed. When the data is queried, the request is sent to both the private cloud and the public cloud at the same time. Our system contains two databases; they are private data and the public data. The private data contains the original data and the perturbed data, the public cloud contains the perturbed data only.

Initially, the user sends the request query to the system. This query request checks the private cloud data and takes the perturbed value of that. This perturbed value checks the public cloud data and retrieves all the attributes in that row. The RRP algorithm includes data perturbation phase and the data retrieval phase. Input: Original data *D*, its size *n*,and delicate characteristic [*F*].

4. Results and discussions

The performance and the evaluation results of the proposed method are depicted in this section and also the proposed method is compared with the existing privacy-preserving techniques. Here, the proposed method utilized the RRP algorithm for perturbing the high dimensional data by utilizing the K-means clustering and ACO algorithm. The proposed RRP algorithm is more interpretable and quickly manages the huge number of data. To perturb the data, the privacy-preserving data mining approach incorporates data partitioning, optimization, and random rotation perturbation. Different types of clustering approaches are compared with the proposed K-means clustering method to evaluate the performance of the proposed approach.

4.1 Dataset description

In this section, the real-time healthcare data sets are utilized to perform an effective analysis for RRP that are commonly used in result analysis for data mining. Here, the healthcare dataset contains number attributes and more information (age, sex, height, weight, and diseases) about the patients. This information about the patients is classified as sensitive attributes and non-sensitive attributes. Here, the non-sensitive attributes are publicly readable and the sensitive attributes are preserved privately. Here, the perturbation approach applies to the sensitive attributes (age, name, height, and weight).

4.2 Performance evaluation

In this section, the performance of the proposed method is evaluated by comparing the proposed clustering and RRP method with the existing privacy-preserving approaches. Here, the proposed K-means clustering approach is compared with the KNN (K-Nearest Neighbour) and the fuzzy C-means classifier. The proposed ACO optimization approach is compared with the existing PSO (Particle Swarm Optimization) and Genetic algorithm. The presented RRP approach is compared with the RDT (Reversible Data Transform) and PDE (Privacy Difference Expansion) algorithm. The below mentioned table shows the three optimization approaches with ACO, PSO [25] and Genetic Algorithm [26], in order to perform an optimization accuracy analysis of datasets after going through RRP perturbation, to knowledge reservation thereby analyzing the impact of the algorithm. The experimental results from the table show that the optimization accuracy for using the different number

of QI attributes in the test data with RRP perturbation all came very close to the classification accuracy of the original data sets. This means the data sets with RRP protection do not lose its original knowledge because of the perturbation, the objective of knowledge reservation proving that RRP can indeed achieve.

4.2.1. Accuracy of clustering results

To compute the accuracy of our proposed method, the same K-mean clustering algorithm is applied to the original data and the perturbed data. Here, clustering approach is compared with the existing KNN [23] and the Fuzzy C-mean clustering [24] approaches. The accuracy of the clustering data is computed as follows,

$$Acc = \frac{1}{N}\sum_{i=1}^{k}(|Cluster_i(D)| - |Cluster_i(D')|) \quad (14)$$

Where,

N- Number of original data sets

K- Number of clustered groups

$|Cluster_i(D)|$ - Number of data in the original data sets.

$|Cluster_i(D')|$- Number of data in the perturbed data sets.

Fig. 3 shows the accuracy of the clustering results and it is compared with the existing KNN and Fuzzy C-means clustering algorithm. Here, compared with the K-means clustering approach, the KNN is comparatively slow, mainly because it has to compute the distances of the query to all data. And also the computational complexity is also high on the KNN clustering approach. Another method, Fuzzy C-means clustering is the well-known approach and its partitioning the high dimensional data sets into a number of fuzzy classes. It requires many fuzzy rules to cluster the big data, so the accuracy is very less. This is the main drawback of the fuzzy C-means clustering. Compared with these two existing methods, the proposed K-means clustering approach has the best accuracy (98.50%), because it takes less time to cluster the high dimensional data sets.

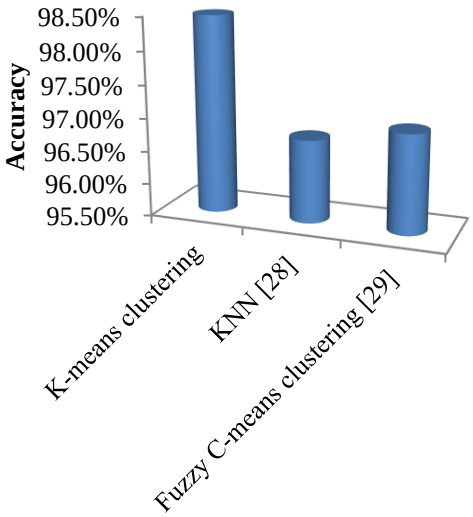

Figure. 3 Accuracy of clustering results

4.2.2. Accuracy of optimization results

To compute the accuracy of the optimization phase, the ACO algorithm used in the proposed method compared with the existing well-known optimization algorithms namely PSO and Genetic algorithm. In higher dimensional datasets operation, the PSO [25] and the genetic algorithm [26] play a significant role to optimize the data. However, in high dimensional space, the PSO is easy to fall into a local optimum. When the PSO solves the high dimensional and complex problems, the computational complexity of PSO is high. In the genetic algorithm, the time consumed to optimize the high dimensional data is very high, because it requires a number of parameters to solve the optimization problem. Due to this main drawback of these two existing algorithms, the accuracy is very less. Compared with these two optimization approaches, the ACO is most widely used probabilistic approach for solving the computational complexity in the high dimensional data. So the accuracy of our proposed method is high when compared with the existing approaches.

4.2.3. Accuracy of perturbation results

To compute the accuracy of the perturbation results, the proposed random rotation perturbation algorithm is compared with the existing perturbation

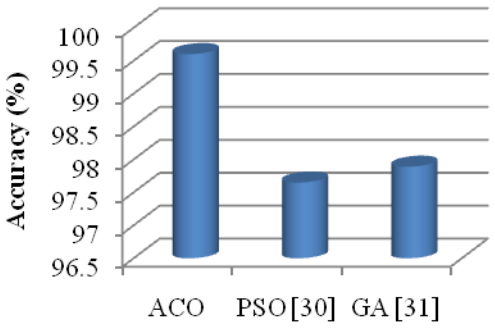

Figure. 4 Accuracy of optimization results

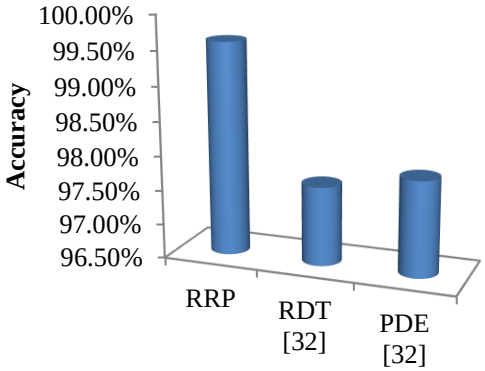

Figure. 5 Accuracy of perturbation results

algorithms namely PDE and RDT [27]. Comparing to our system RDT uses an adjustable weighting mechanism. In PDE algorithm, the parameters does not have any certain relationship with the knowledge reservation and in the result there is an difficulty to setting the proper parameters. For perturbing the high dimensional data sets, the computational complexity is very high by using the PDE and RDT. In the perturbation phase, the good rotation perturbation is done by using the RRP approach. Compared with the existing approaches, the proposed RRP approach has better accuracy.

4.3 Comparison analysis

In this section, to compute the performance of the proposed method, the data privacy ratio, iteration for perturbing and retrieving the data and the execution time for perturbing and retrieving the data is computed.

4.3.1. Data privacy ratio and iteration results

The data privacy ratio and the iteration results are computed in this phase. Here, the proposed RRP method is compared with the existing RDT and PDE approach. The data privacy ratio computation is utilized to determine, how well the technique is utilized to precisely preserve the privacy data. The number of rounds taken to complete the operation is known as the iteration results. The existing RDT and PDE approaches involve a number of mathematical and computational methods to perturb the sensitive data. So, here, the data privacy ratio is low and the iteration for complete the operation is high. Compared with these existing methods, the proposed RRP method has high privacy ratio and less number of iterations because the RRP method is a simple perturbation approach

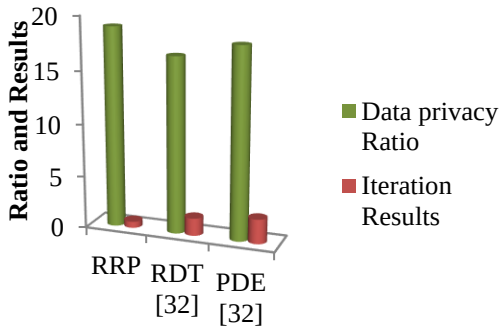

Figure. 6 Data privacy ratio and Iteration results

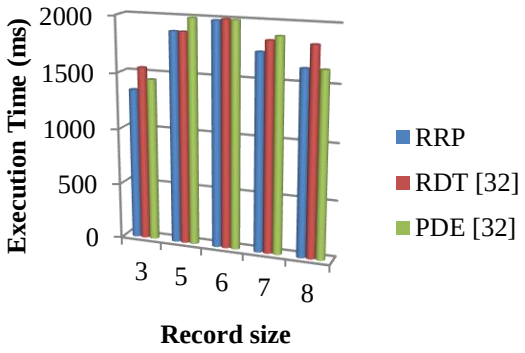

Figure. 7 Execution time of data perturbation phase

4.3.2. Execution time of data perturbation phase

In this section, the execution time of the data perturbation phase and the data retrieval phase is computed. The proposed RRP method is compared with the existing RDT and PDE approach to compute the execution time. The RRP method gives better execution time over the other two approaches. In record size 3, the execution time in the data perturbation phase is 1926 (ms) and 1556 (ms) in the RDT and PDE, but the proposed RRP approach has 1426 (ms). In record size 5, the execution time in the data perturbation phase is 1965 (ms) and 1885 (ms) in the RDT and PDE, but the proposed RRP approach has 1765 (ms). In record size 6, the execution time in the data perturbation phase is 1754 (ms) and 1774 (ms) in the RDT and PDE, but the proposed RRP approach has 1654 (ms). In record size 7, the execution time in the data perturbation phase is 1865 (ms) and 1876 (ms) in the RDT and PDE, but the proposed RRP approach has 1766 (ms). In record size 8, the execution time in the data perturbation phase is 1898 (ms) and 1897 (ms) in the RDT and PDE, but the proposed RRP approach has 1876 (ms).

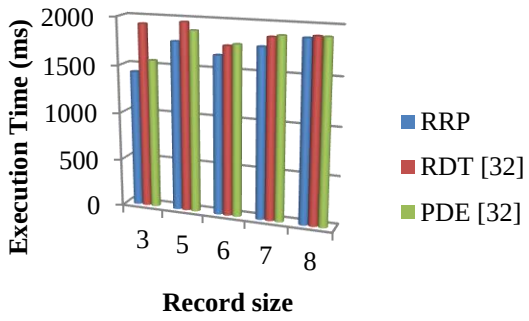

Figure. 8 Execution time of data retrieval phase

4.3.3. Execution time of data retrieval phase

In this section, the execution time of the data retrieval phase is computed. The proposed RRP method is compared with the existing RDT and PDE approach to computing the execution time of the retrieval phase. In record size 3, the execution time in the data perturbation phase is 1545 (ms) and 1445 (ms) in the RDT and PDE, but the proposed RRP approach has 1345 (ms). In record size 5, the execution time in the data perturbation phase is 1875 (ms) and 1985 (ms) in the RDT and PDE, but the proposed RRP approach has 1874 (ms). In record size 6, the execution time in the data perturbation phase is 1999 (ms) and 1996 (ms) in the RDT and PDE, but the proposed RRP approach has 1986 (ms). In record size 7, the execution time in the data perturbation phase is 1845 (ms) and 1885 (ms) in the RDT and PDE, but the proposed RRP approach has 1745 (ms). In record size 8, the execution time in the data perturbation phase is 1834 (ms) and 1633 (ms) in the RDT and PDE, but the proposed RRP approach has 1630 (ms).

5. Conclusion

In order to protect the privacy information in the original data set, this work implemented the concept of data rotation transformation in the image processing domain and developed the RRP algorithm, which can perturb and restore the data. And also the proposed method utilized the combined K-means clustering and the ACO algorithm for the cluster formation. This clustered data is perturbed by using the RRP algorithm. The perturbation result is very hard to recognize. After that, the recovering phase is utilized to recover the original data set from the perturbed dataset. The experimental results show that the proposed method accurately protects the privacy data when compared with the existing method. In future, to accomplish the more elevated level of privacy-preserving, the work can additionally explore other data hiding techniques. Furthermore, choosing suitable QI attributes are a work in progress.

References

[1] C. Aggarwal, *Data Mining*, Vol. 1, Springer, Cham, Switzerland, 2015.

[2] L. Dunning and R. Kresman, "Privacy-Preserving Data Sharing With Anonymous ID Assignment", *IEEE Transactions on Information Forensics and Security*, Vol. 8, No. 2, pp. 402-413, 2013.

[3] R. Chen, B. Fung, N. Mohammed, B. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression", *Information Sciences*, Vol. 231, No. 1, pp. 83-97, 2013.

[4] S. Scardapane, R. Altilio, V. Ciccarelli, A. Uncini, and M. Panella, *Multidisciplinary Approaches to Neural Computing*, Vol. 69, Springer, Cham, 2017.

[5] F. Giannotti, L. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases", *IEEE Systems Journal*, Vol. 7, No. 3, pp. 385-395, 2013.

[6] J. Yang, J. Li, and Y. Niu, "A hybrid solution for privacy-preserving medical data sharing in the cloud environment", *Future Generation Computer Systems*, Vol. 43-44, No. 1, pp. 74-86, 2015.

[7] J. Li, Z. Liu, X. Chen, F. Xhafa, X. Tan, and D. Wong, "L-EncDB: A lightweight framework for privacy-preserving data queries in cloud computing", *Knowledge-Based Systems*, Vol. 79, No.5, pp. 18-26, 2015.

[8] F. Yu, S. Fienberg, A. Slavković, and C. Uhler, "Scalable privacy-preserving data sharing methodology for genome-wide association studies", *Journal of Biomedical Informatics*, Vol. 50, No.8, pp. 133-141, 2014.

[9] G. Afzali and S. Mohammadi, "Privacy-preserving big data mining: association rule hiding using fuzzy logic approach", *IET Information Security*, Vol.12, No.1, pp. 15-24, 2017.

[10] J. Zhang and J. Yang, "Linear reconstruction measure steered nearest neighbor classification framework", *Pattern Recognition*, Vol. 47, No. 4, pp. 1709-1720, 2014.

[11] Z. Gheid and Y. Challal, "Novel Efficient and Privacy-Preserving Protocols for Sensor-Based Human Activity Recognition", In: *Proc. of the IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted*

Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, pp. 301- 308, 2016.

[12] A. Azar and A. Hassanien, "Dimensionality reduction of medical big data using neural-fuzzy classifier", *Soft Computing*, Vol. 19, No. 4, pp. 1115-1127, 2014.

[13] G. Li, "A New Bayesian-Based Method for Privacy-Preserving Data Mining", In: *proce. of the International Conference on Intelligent and Interactive Systems and Applications*, pp. 171-177, 2017.

[14] A. Marrella, A. Monreale, B. Kloepper and M. Krueger, "Privacy-Preserving Outsourcing of Pattern Mining of Event-Log Data - A Use-Case from Process Industry", In: *Proc. of the IEEE International Conference on Cloud Computing Technology and Science*, pp.545-551, 2016.

[15] Z. Fu, F. Huang, K. Ren, J. Weng, and C. Wang, "Privacy-Preserving Smart Semantic Search Based on Conceptual Graphs Over Encrypted Outsourced Data", *IEEE Transactions on Information Forensics and Security*, Vol. 12, No. 8, pp. 1874-1884, 2017.

[16] Q. Jiang, M. Khan, X. Lu, J. Ma, and D. He, "A privacy-preserving three-factor authentication protocol for e-Health clouds", *The Journal of Supercomputing*, Vol. 72, No. 10, pp. 3826-3849, 2016.

[17] B. Colaco and S. Khan, "Privacy-preserving data mining for social networks", In: *Proc. of the International Conference on Advances in Communication and Computing Technologie*s, pp.1-4, 2014.

[18] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 1, pp. 97-107, 2014.

[19] D. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", *Applied Computing and Informatics*, Vol. 12, No. 1, pp. 90-108, 2017.

[20] H. Lou, Y. Ma, F. Zhang, M. Liu, and W. Shen, "Data mining for privacy-preserving association rules based on improved MASK algorithm", In: *Proc. of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design*, pp. 265-270, 2014.

[21] C. Lin, T. Hong, and H. Hsu, "Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining", *The Scientific World Journal* , Vol. 2014, No. 1, pp. 1-12, 2017.

[22] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information Security in Big Data: Privacy and Data Mining", *IEEE Access*, Vol. 2, No. 1, pp. 1149-1176, 2014.

[23] H. Rong, H. Wang, J. Liu, and M. Xian, "Privacy-Preserving k-Nearest Neighbor Computation in Multiple Cloud Environments", *IEEE Access*, Vol. 4, No. 1, pp. 9589-9603, 2016.

[24] P. Kumar, K. Varma, and A. Sureka, "Fuzzy-based clustering algorithm for privacy-preserving data mining", *International Journal of Business Information Systems*, Vol. 7, No. 1, p. 27, 2011.

[25] K. Saranya and K. Premalatha, "Privacy-Preserving Data Clustering using hybrid Particle Swarm Optimization Algorithm", *Asian Journal of Research in Social Sciences and Humanities*, Vol. 7, No. 3, p. 601, 2017.

[26] R. Purohit and D. Bhargava, "An illustration of a secured way of data mining using privacy-preserving data mining", *Journal of Statistics and Management Systems*, Vol. 20, No. 4, pp. 637-645, 2017.

[27] C. Lin, "A reversible data transform algorithm using integer transform for privacy-preserving data mining", *Journal of Systems and Software*, Vol. 117, No. 7, pp. 104-112, 2016.