Privacy Preservation Model using Machine Learning Techniques

Dr.K.Thirukumar
Associate Professor, Department of Computer Science and Engineering
Dr.Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India
E-mail:thi.kumarr@gmail.com

Dr.A.Rathinavelu
Professor, Department of Computer Science and Engineering
Dr.Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India

Abstract--- An enormous amount of data sharing and information among people all over the world is very much huge. There is a huge demand on personal data which is to be shared to the society to improving their business. But the confidentiality maintenance of an individual's information is a difficult task with the various data releases. When organizations coordinate prior to the data publication, then attacks are possible. Research works like k-anonymity, l-diversity model were proposed but the composition attack avoidance is an issue while publishing the data. Privacy preservation model was developed to resolve various types of disclosure and attacks using Partitioning method and Particle Swarm Optimization (PSO) technique. The proposed model enhances the privacy and data utility better than the existing approaches. Performance has been evaluated based on classification accuracy, disclosure rate, execution time and privacy preservation rate. With the help of PSO model the performance improvement is better when compared to earlier models. Furthermore, this model uses the t-closeness principle which avoids most of the privacy threats.

Keywords--- Privacy Preservation, Partitioning Method, Correlation based Fuzzy model, Swarm Intelligence Particle Swarm Optimization (PSO), t-closeness

I. INTRODUCTION

Large amount of person specific data of an individual is collected by several organizations such as online shopping and insurance agencies etc. The collected data from an individual is verified digitally to identify useful information for research purposes in various areas like medicine and business trend analysis. A micro data set includes privacy information about individuals. Privacy preserving data publishing (PPDP) is an emerging area where individual's privacy will get protected. Data owner may reveal the collected data to public either for data analysis or for research purpose. But the intruders access this data and will be combining it with some external data which is available publicly available to get the person's sensitive information. To break this, the data owner may apply hiding process and release the masked data to maintain the confidentiality of individuals. Even some of the intruders will identify the confidential information about the individual using the masked data which is linked with external data. PPDP is a method to overcome such issues to protect the data from the intruders. Tiancheng Li et al. (2012) proposed slicing method which partition the data horizontally and vertically to achieve 'l –diversity'. Slicing process includes partitioning of attributes, generalization and tuple partitioning. Slicing protects the micro dataset with high data utility and protects membership information. Slicing manages the dimensionality curse by grouping several attributes into single column. Slicing is an efficient algorithm for sliced data with l-diversity needs and it protects the attribute disclosure risks. Organizations share data to public about individuals to improve business process by adhering law and regulation of the government. An adversary may observe private information of an individual with the dissimilar data publications using quasi-identifiers related with the records. Various privacy protection models (e.g., k-anonymity and l-diversity) proposed by researchers are unable to protect the individual's data from the intruders.

Micro dataset consist of several attributes, among these, Disease is considered to be the sensitive attribute. The disease values are swapped among the equivalence class to prevent the identity disclosure. Overlapping among several columns with same attribute prevents the attribute disclosure which also improves the data utility. With this we have proposed a privacy model to protect the privacy of individual's identity and sensitive information from the intruders and to focus on the data utility. Reordering the tuples among the equivalence class takes more time to converge, to reduce this we apply particle swarm optimization (PSO) method for satisfying the diversity constraint in optimized time and improved efficiency. The Particle swarm optimization (PSO) is an optimization technique developed by Kennedy et al.(1995). The working phenomenon is explained in the flow diagram represented in the Figure1.

In PSO, Particles are the optimized solutions, which navigate through the problem space based on the current optimum particles. In the initial step, a group of random particles are initialized and based on the fitness function of the problem it searches in the problem space for finding the optimal solutions by updating generations. In each iteration, every particle is updated based on two "best" values.

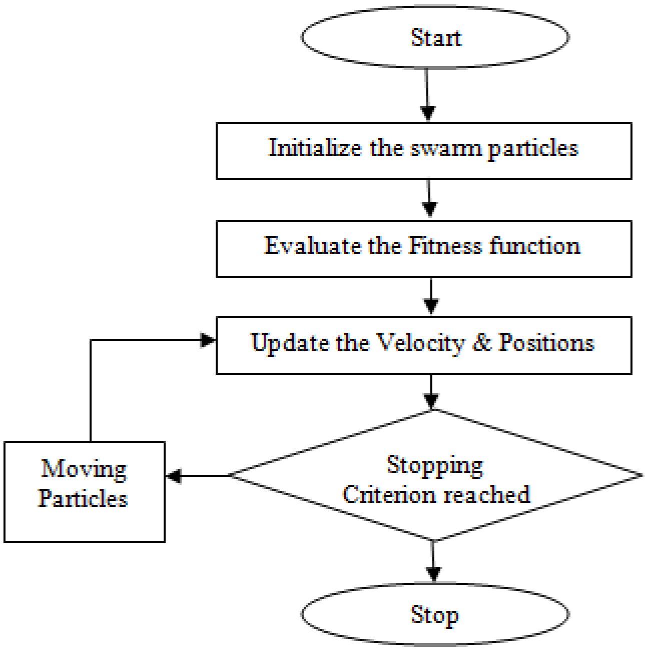
Start
Initialize the swarm particles
Evaluate the Fitness function
Update the Velocity & Positions
Stopping Criterion reached
Moving Particles
Stop

Figure 1. Flow diagram of Particle Swarm Optimization

One is the best solution (fitness) it has achieved so far. It is considered as pbest. Second best value is tracked by the particle swarm optimizer, which is obtained so far by any particle in the population. This is a global best value called as gbest. Another local best value called lbest is a particle which takes part of the population as its topological neighbors. After selecting the two best values i.e pbest and gbest, the particle updates its position and velocity in the search space with following equation (1) and (2).

v[]=v[]+c1*rand()*(pbest[]-present[])+c2*rand()*(gbest[]-present[]) (1)
present[]=present[]+v[] (2)

rand() is a function generates random number between (0,1). Learning factors are C1,C2 where c1 = c2 = 2.
v[] is the particle velocity, present[] is the current particle (solution).
Vmax is the maximum velocity among Particles' velocities on each dimension which is a parameter specified by the user. If the accelerations sum on that dimension exceeds Vmax then the velocity is limited to Vmax. There are 2 major steps involved in the particle swarm optimization process: the construction of fitness function based on problem and the solution representation for the given problem. PSO takes the real numbers as particles which is easy for manipulation.

Our major contributions in this paper are as follows. Correlation based Fuzzy Clustering Algorithm (CFCA) is a Partitioning based privacy protection model which is developed to improve the existing PPDP techniques. The correlation based approach focuses on data privacy by reducing the information loss using fuzzy clustering mechanism. This partitioning method helps to handle the improved privacy with less information loss in an effective manner. The disclosure rate is measured for each tuple in the equivalence class and the minimal resultant value of the disclosure rate is considered to provide the maximum utility rate and the privacy during data publication. During tuple reordering among the buckets, swarm intelligence method PSO was applied to reduce the time taken for obtaining better performance in the anonymized dataset.

The rest of this paper is organized in the following manner. Section II provides a more detailed literature survey of the several privacy preservation models to protect data before publication. Section III describes the proposed model to describe the tradeoff between the privacy and the data utility within the datasets. Section IV provides discussion on results with various parameters. Finally, Section V concludes the paper with a discussion on what needs to be done further.

II. RELATED WORKS

Micro-data contains sensitive information about individuals and publishing the data to the society may reduce the privacy of individual. While distributing the data to public any organization in the world has to adhere the policies and guidelines of the government according to HIPAA law. Sweeney,L (2002) proposed the k-anonymity model which is the basic principle of all the anonymization models. During the data release, intruder cannot able to differentiate the individual from at least k-1 individuals whose value also appear in the release. Ranjit Ganta et al. (2008) propose the solution for composition attacks with auxiliary information. The effectiveness of anonymization model gets increased when various organization releases anonymized data with overlapping populations. Randomization-based models of privacy which limit the breaches of composition attacks with utilization of the arbitrary side information. Multiple sensitive attributes may be anonymized using Multi-dimension bucketization models. SLOMS method was proposed by Jianmin Han et al. (2013). This model hides the sensitive values of multiple attributes by vertically dividing sensitive attributes into tables and buckets which have the characteristics of l-different values in each bucket. MSB-KACA algorithm anonymized the micro-data with multiple sensitive attributes through SLOMS.

C. M. Fung et al. (2013) discussed about distributed privacy-preserving data publishing (PPDP) methods for data privacy. PPDP method focuses a large attention of researchers nowadays. Hongtao Li et al. (2013) proposed an e-secrecy solution algorithm called MAIA for mechanism-based attack. Secured Multiparty Computation (SMC) was designed by Zaman et al. (2014) which is a distributed model for privacy preservation instead of centralized model employed for single data owner for data publication. A framework designed which assures the differential privacy standards and guarantees for maximum data usability for classification. This guarantees privacy at record level of the data owner while disclosing the real data in the data set.

Many PPDP models like k-anonymity failed to protect the individual's sensitive data with composition attack. The attack can be avoided when the linking of data sources is not possible with prior to data publication. A (d,α)-linkable - probabilistic model was introduced by Sarowar Sattar et al. (2014) which reduces the composition attack during the coordination among various data releases. An innovative approach called slicing is proposed by Tiancheng Li et al. (2012) which divide the data and attributes horizontal and vertical. Slicing maintains data utility than the generalization & bucketization techniques for membership disclosure protection. Managing high-dimensional data using slicing is very easy. Slicing focuses on both attribute & membership disclosure protection mechanism by applying the l-diversity principle.

Gabriel Ghinita et al. (2011) designed a model for sparse high-dimensional mocro data set. This model is based on K-nearest-neighbor (KNN) model which searches the high-dimensional spaces through locality-sensitive hashing (LSH) method which avoids the curse of dimensionality. Older models like Generalization and Bucketization are designed by Ashwini et al., (2014) for privacy preserving micro data publishing. The generalization method changes the value to more generic leads to information loss in high-dimensional data. Bucketization groups the records which does not protect the membership disclosure. But the Slicing mechanism provides the enhanced data utility and avoids the membership disclosure. A combination of anatomization and enhanced slicing anonymization approach which was proposed by Christopher et al.(2016) which manages the high dimensional data with more sensitive data.

To provide better data utility from Slicing approach a new idea proposed that overlapping of attributes during partitioning approach. Yinghua Lu et al. (2013) propose an enhanced fuzzy c-means algorithm which manages the micro-data set on traditional fuzzy c-means algorithm. Issues faced in traditional fuzzy c-means (FCM) clustering algorithm is to select the initial cluster centers which may take more time for converge.

III. Proposed Privacy Preserving Model

PPDP provides the improved methods and efficient tools for publishing the sensitive information while preserving data privacy. The Correlation based Fuzzy Clustering Algorithm (CFCA) has been developed to handle large number of attributes which contain sensitive attributes in different datasets. Also it focuses to protect various essential disclosure risks like membership disclosure and attribute disclosure. For records reordering this algorithm takes more time for convergence, to avoid this problem, optimization technique called PSO has been introduced. The flow diagram of the CFCA is illustrated in Figure 2.

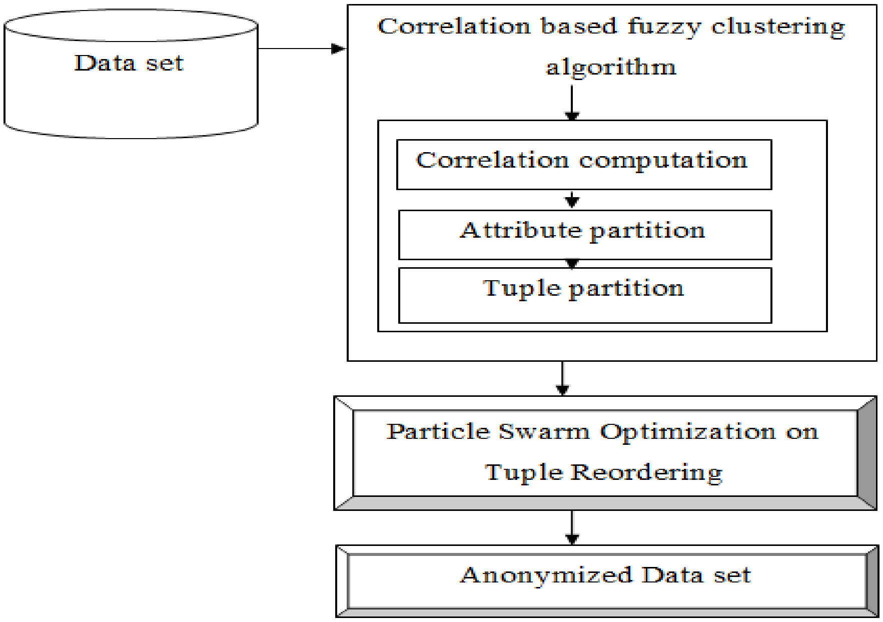
Data set
Correlation based fuzzy clustering algorithm
Correlation computation
Attribute partition
Tuple partition
Particle Swarm Optimization on Tuple Reordering
Anonymized Data set

Figure 2: Flow diagram of Correlation based fuzzy clustering algorithm

Figure 2 illustrates the overall process of the correlation based fuzzy clustering algorithm along with the need of optimization using PSO which is described in Figure 1. The proposed CFCA model consists of three major phases such as selection of attributes based on correlation factor, partitioning attributes and partitioning records in order to preserve the privacy during the data publishing. The selection of attributes from the micro-dataset for combining more attributes into single column the correlation is considered. Secondly by applying fuzzy clustering algorithm the attribute are splitted into several overlapping columns. Finally, by applying systematic clustering algorithm, tuples are partitioned into several buckets by permuting the data across several buckets which enhances the data utility. The detail description of the CFCA model algorithm is provided in following sections.

Let D is the micro-data set which is to be published, contains n number of attributes, the attributes are represented as follows, $A = \{a_{1,}a_{2\ldots\ldots,}a_n\}$. It may contain identification attributes which can be removed before publishing. Quasi identifiers are the attributes which is linked with other publicly available attributes to identify an individual. Sensitive attributes to be protected from the intruders. For selecting the attributes we consider the correlation among attributes in the dataset. Pearson correlation coefficient is widely used technique for evaluating correlations between two attributes. It is a mathematical process to identify the relationship between the attributes.

The formula for Pearson's correlation coefficient (r) is mathematically expressed as ,

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (3)$$

The calculated correlation values ranges between -1 to 1. Table 1 contains the correlation values which are computed based on the Equation (3). X, Y are two different attribute in the micro-data set. Table 1 values are taken as input to the Fuzzy clustering algorithm. Merging of clusters is illustrated using the pair-wise similarity matrix between five clusters as shown in figure. For a threshold of 0.4 , the groups are (1,4),(2),(3.5),(4,1),(5,3). After elimination of duplicate groups and singleton the remaining groups are (1,4) and (3,5) . When the threshold value is 0.2 , the groups are (1,2,4),(2,1,3,4),(3,2,5),(4,1,2),(5,2,3). After elimination of duplicates and sub-groups, the remaining sub groups are (2,1,3,4) and (3,2,5). It can be observed that cluster 2,3 is part of two different groups.

Table 1 Correlation values among the attributes in the Census dataset

	Age	Zipcode	Sex	Occupation	Salary
Age	1.0	0.32	0.16	0.43	0.06
Zipcode	0.32	1.0	0.21	0.32	0.11
Sex	0.16	0.21	1.0	0.17	0.60
Occupation	0.43	0.32	0.17	1.0	0.16
Salary	0.06	0.11	0.60	0.16	1.0

Based on this strategy, the overlapping of attributes among various clusters ensures a fuzzy partitioning of the clusters. Therefore, the attributes which resides in same cluster are partitioned as single column. For example, consider {zipcode, age, sex, occupation}as attributes in the dataset. Based on fuzzy clustering, the attributes are partitioned into three clusters(columns) based on their membership values such as {zipcode, age}, {zipcode, sex}, {age, sex, occupation}.

The records are partitioned to several clusters using a different clustering method called systematic clustering which considers the information loss during the data publishing. During every iteration the clusters are formed with group of records which has less information loss is considered as a separate bucket. The amount of information loss is due to generalizing attribute a_i, denoted by IL(y) is defined as,

$$IL(Y) = |Y|\left[\sum_{i=1}^{r}\frac{N_{imax}-N_{imin}}{t_{i\mathrm{Nimax}}-t_{i\mathrm{Nimin}}} + \sum_{j=1}^{s}\frac{H(\Lambda(\cup_{cj}))}{H(\tau_{cj})}\right] \qquad (4)$$

The result of vertical partitioning contains several overlapping attributes into separate column of fuzzy clustering will not reveal any proper information to the intruder. In systematic clustering, tuples are included in buckets to form equivalence class. Initially random records were selected into ‘n’ buckets. If the bucket containing records which are not satisfying the given constraint (less information loss) are dealt separately. By removing those records from the bucket and distributing those non-satisfying tuples which belongs to this bucket are added to other buckets that cause minimum information loss. Because inserting a new record to the existing bucket will not violate the l-diversity constraint.

The CFCA algorithm has been described in the detailed manner by explaining the role of Correlation computation is described in the Algorithm 1. Once the correlation between the chosen attribute is computed using the equation (3), the results with the maximum correlation will be given as input to the Fuzzy Clustering Algorithm (FCA). The clustering with the defined number of ‘c’ is performed using the minimum information loss as described in the Algorithm 2.

Algorithm 1 Correlation based Fuzzy Clustering Algorithm (CFCA)

```
Given Micro dataset (D)
Given k-value and l-value
Output anonymized table
Identify the quasi attributes and sensitive attribute from D
Compute correlation matrix $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ for the chosen attributes
Call: Fuzzy Clustering Algorithm over the matrix which has high correlation value
Construct modified data set by finding overlapping attributes
Apply permutation between vertical partitioning attributes
Construct the horizontal partition of attributes which satisfies l-diversity constraints using clustering
Find the information loss (IL) using
$IL(Y) = |Y|\left[\sum_{i=1}^{r}\frac{N_{imax}-N_{imin}}{t_{iNimax}-t_{iNimin}}+\sum_{j=1}^{s}\frac{H(\Lambda(\cup_{cj}))}{H(\tau_{cj})}\right]$
Reordering the tuple according to the minimum information loss
Call: PSO with maximum privacy and utility as fitness function
Return anonymized table
```

The initial points for the clustering algorithm are selected randomly. The points are clustered based on the distance between other data points and the center.

Algorithm 2 Fuzzy Clustering Algorithm (FCA)

```
Given finite collection of n elements $X=\{x_1,x_2,\ldots x_n\}$
Given c-value, threshold value (t)
Output c number of clusters
Choose the random data point as initial data point
Repeat
Compute the center from random data point to all other points using

$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}$
```

```
Add the data points into the cluster based on minimum distance from center point
Until change between two iterations is no more than threshold value (t)
Returns c number of clusters
```

Algorithm 3 Particle Swarm Optimization (PSO)

```
For each particle
{
   Initialize particle
}
Do until minimum error criteria
{
   For each particle
   {
      Calculate Data fitness value
      If the fitness value is better than pBest
      {
         Set pBest = current fitness value
      }
      If pBest is better than gBest
      {
         Set gBest = pBest
      }
   }
   For each particle
   {
      Calculate particle Velocity
      Use gBest and Velocity to update particle Data
   }
```

Systematic Clustering algorithm helps to overcome the problem of selecting the number of clusters initially. This helps to reduce the information loss which improves the data utility. Tuple reordering process among the equivalence class in every buckets which are not satisfying the l-diversity and minimum information loss constraint is a time consuming process, so to reduce such time we include the optimization technique called PSO, to provide optimal solution in a lesser time. The PSO is described in Algorithm 3.

IV. Experimental Analysis

The proposed algorithm is developed in Java language. Adult Dataset has been downloaded from UCI Machine learning repository [17] for testing the efficiency of the proposed model. It is the benchmark dataset for Privacy preservation model. The experiments are carried out on a system configured with Intel core i5 processor, 3 GB RAM and 500 HDD. The performance of the proposed algorithm is evaluated based on Classification accuracy and Privacy preservation rate. Adult Dataset contains 48842 records in total, after removing the noisy data, it contains 45,222 valid records. Also it contains 15 Categorical and numeric attributes. To evaluate our algorithm, we have considered the Occupation as the Sensitive attribute and the Quasi identifiers are "Education", "Sex" ,"Age", "Marital-Status", "Race" and "Work class".

Classification accuracy in adult dataset

Classification accuracy is computed in adult dataset using the CFCA method with Decision Tree C4.5 (J48) is illustrated in following figure.

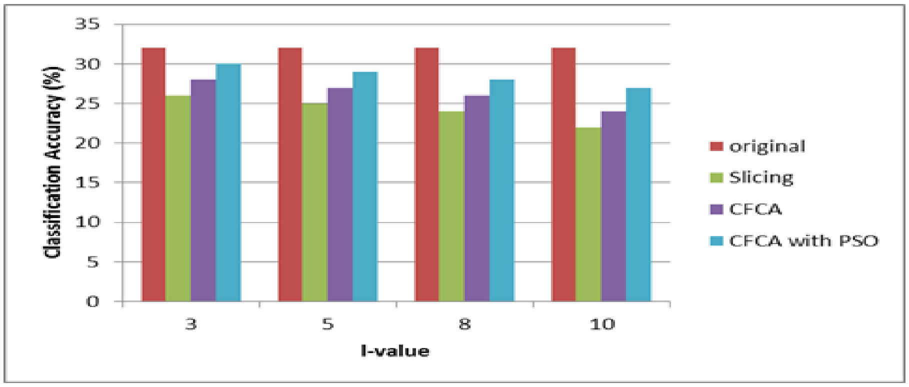

Figure 3 Classification accuracy using Decision Tree C4.5 (J48)

The Figure3 represents the classification accuracy of adult dataset using Decision Tree C4.5 (J48) classifier. For our experiments, we vary the privacy thresholds from 3 to 10 and the classification accuracy has been measured, from this graph, it is clearly illustrated that CFCA with PSO yields a good accuracy value while compared to other existing models. Without anonymization, we considered the baseline accuracy for the original dataset. From the results it is closer to the baseline accuracy in the proposed model. With varying l-values, for occupation attribute as the classifier, the accuracy has been improved because of more unique values in each bucket.

We have recorded the results of the CFCA approach with different set of quasi identifier as the target attribute for the classifier. The analysis of the proposed approach shows acceptable accuracy has been obtained with varying privacy factor. Also it is observed that when increasing the l-value the accuracy of the target attribute is slowly decreased. But the accuracy is better for sensitive attribute as the target attribute. Another significant experiment is done by comparing the J48 with Naïve Bayes classifier by considering Occupation as sensitive attribute. Accuracy results are better in Naïve Bayes Classification. The inference from the experiments is accuracy is improved as 30.5% from the existing privacy preservation models.

Analysis of membership disclosure in adult dataset

In CFCA, while permuting the records among the buckets, fake tuples will get generated. Fake tuple generation is an advantage in our model, where we will be hiding the original data in the crowd. From Figure 4, the l-values are varied from 2 to 14 in the buckets for our experiment. The inference made out in this experiment is that with t the more number of fake tuples is good enough to hide the original records among the buckets. The proposed CFCA model results shows better performance than the existing models. This is because; correlation among attributes is taken as a measure for generating column in the anonymized data set. This factor improves both data utility and the privacy measure with maximum accuracy. From the experiments, the percentage of fake records generated is more than that of original records. This fake record among buckets provides the membership protection from the intruders. CFCA with PSO model produces more to protect against membership disclosure with a lesser execution time. Figure 5 represents the privacy preservation rate among various models.

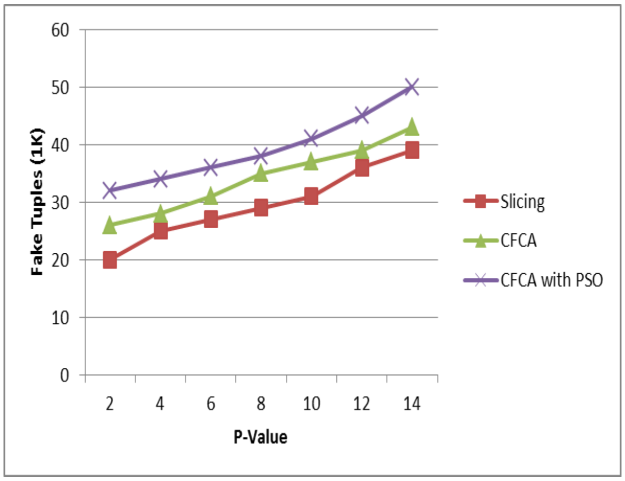

Figure 4: Fake Records Generation

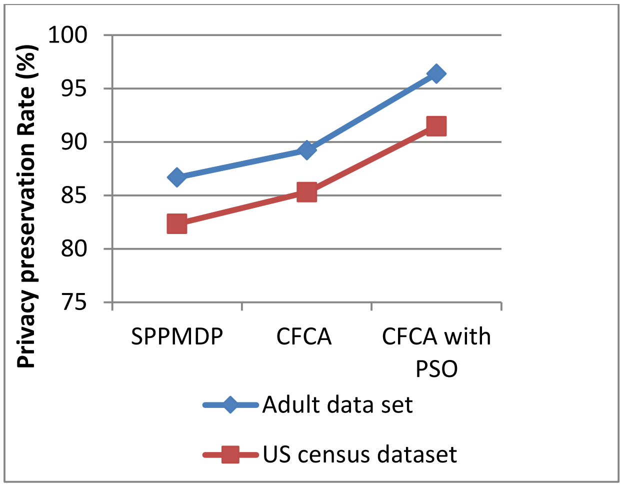

Figure 5: Privacy Preservation Rate

Execution time analysis

From Figure 6, the proposed model CFCA with PSO produces the anonymized dataset in a minimum execution time while compared to the existing models CFCA & SLAMSA methods. PSO algorithm plays a major role in the reduction of execution time during the tuple reordering phase. In US census dataset, the execution time is reduced by 11% and 29% using CFCA-PSO methods compared to existing approaches.

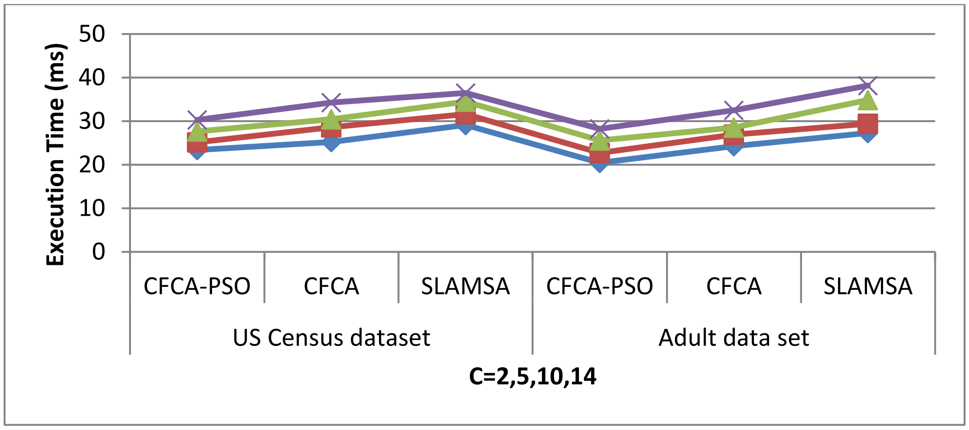

Figure 6: Execution Time Measure

V. CONCLUSION

CFCA - Privacy-preserving data publishing is a model for privacy preservation with enhanced data utility which prevents the attribute disclosure and membership disclosure during the data publication. It is an efficient method to distribute anonymous data and ensure privacy against identity disclosure rate of an individual. A Correlation Approach based on Fuzzy Clustering Algorithm (CFCA) is developed for ensuring the privacy based on attribute composition method. The fuzzy clustering approach is used for attribute composition to achieve the data privacy and reducing the information loss. CFCA is enhanced with PSO approach to improve the efficiency of the algorithm. From the experimental results shows that it has been improved 20% higher than existing approaches.

ACKNOWLEDGMENT

We would like to express our heartfelt thanks to the anonymous reviewers for their helpful suggestions on the organization of the paper. We would like to express our sincere thanks to the Management and Secretary, Dr.Mahalingam College of Engineering and Technology, Pollachi for their support and facilities provided to complete this work.

REFERENCES

[1] Hemanta Kumar Bhuyan, Narendra Kumar Kamilab Department “Privacy preserving sub-feature selection in distributed data mining”,Applied Soft Computing, Elsevier, Volume 36, Pages 552–569, November 2015

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkita-subramaniam, “l-diversity: Privacy beyond k-anonymity”, 22nd International Conference on Data Engineering (ICDE'06), Pages 1-24,April 2006

[3] Amar Paul Singh and. DhanshriParihar, “A Review of Privacy Preserving Data Publishing Technique”, International Journal of Emerging Research in Management and Technology, Volume 2, Issue 6, Pages 32-38, June 2013

[4] Dervis Karaboga and Celal Ozturkient, “A novel clustering approach: Artificial Bee Colony (ABC) algorithm”, Applied Soft Computing, Elsevier, Volume 11, Issue 1, Pages 652–657, January 2011

[5] Gabriel Ghinita, PanosKalnis and Yufei Tao, “Anonymous Publication of Sensitive Transactional Data”, IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 2, Pages 161 – 174, February 2011

[6] Hongtao Li, Jianfeng Ma, and Shuai Fu, “Analyzing mechanism-based attacks in privacy-preserving data publishing”, Optik, Elsevier, Volume 124, Pages 6939– 6945, 2013

[7] Huang Xuezhen, Liu Jiqiang, Han Zhen, Yang Jun, “A New Anonymity Model for Privacy-Preserving Data Publishing”,China Communications, Volume 11, Issue 9, Pages 47-59, September 2014

[8] Jianmin Han, FangweiLuo, Jianfeng Lu and HaoPeng, “SLOMS: A Privacy Preserving Data Publishing Method for Multiple Sensitive Attributes Microdata”, Journal of Software, Volume 8, Issue 12, Pages 3096-3104, 2013

[9] Katarzyna Pasierb, Tomasz Kajdanowicz, Przemysław Kazienko, “Privacy-Preserving Data Mining, Sharing and Publishing”, Journal of Medical Informatics and Technologies, Volume 18, Pages 69-76, 2011

[10] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, “Closeness: A New Privacy Measure for Data Publishing”, IEEE Transactions on knowledge and data engineering, Volume 22, Issue 7, Pages 943-956, 2010

[11] Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing", IEEE transactions on knowledge and data engineering, Volume 24, Issue 3, Pages 561 – 574, 2012

[12] V. Shyamala Susan and T. Christopher, “Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes”, SpringerPlus, Volume 5, Issue 964, Pages 1-21, 2016

[13] Yinghua Lu, Tinghuai Ma, Changhong Yin, XiaoyuXie, Wei Tian and Shui Ming Zhon, “Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data”, International Journal of Database Theory and Application, Volume 6, Issue 6, Pages 1-18, 2013

[14] Yang Xu, Tinghuai Ma, Meili Tang and Wei Tian, “A Survey of Privacy Preserving Data Publishing using Generalization and Suppression”, Applied Mathematics & Information Sciences, Volume 8, Issue 3, Pages 1103-1116, 2014

[15] K.Thirukumar, and A. Rathinavelu. "A correlation approach for preserving privacy in health care data using fuzzy clustering algorithm.", Journal of Medical Imaging and Health Informatics 6.3 ,816-821, 2016

[16] http://www.swarmintelligence.org/tutorials.php

[17] https://archive.ics.uci.edu/ml/datasets.php