Hindawi Publishing Corporation
Advances in Computer Engineering
Volume 2014, Article ID 396529, 12 pages
http://dx.doi.org/10.1155/2014/396529

Research Article

A Clustering Approach for the l-Diversity Model in Privacy Preserving Data Mining Using Fractional Calculus-Bacterial Foraging Optimization Algorithm

Pawan R. Bhaladhare[1] and Devesh C. Jinwala[2]

[1] *Department of Information Technology, SNJB's College of Engineering, Neminagar, Chandwad, Nashik, Maharashtra 423101, India*
[2] *Department of Computer Engineering, S V National Institute of Technology, Surat, Gujarat 395007, India*

Correspondence should be addressed to Pawan R. Bhaladhare; pawan_bh1@yahoo.com

Received 14 July 2014; Accepted 22 July 2014; Published 16 September 2014

Academic Editor: Lijie Li

In privacy preserving data mining, the l-diversity and k-anonymity models are the most widely used for preserving the sensitive private information of an individual. Out of these two, l-diversity model gives better privacy and lesser information loss as compared to the k-anonymity model. In addition, we observe that numerous clustering algorithms have been proposed in data mining, namely, k-means, PSO, ACO, and BFO. Amongst them, the BFO algorithm is more stable and faster as compared to all others except k-means. However, BFO algorithm suffers from poor convergence behavior as compared to other optimization algorithms. We also observed that the current literature lacks any approaches that apply BFO with l-diversity model to realize privacy preservation in data mining. Motivated by this observation, we propose here an approach that uses fractional calculus (FC) in the chemotaxis step of the BFO algorithm. The FC is used to boost the computational performance of the algorithm. We also evaluate our proposed FC-BFO and BFO algorithms empirically, focusing on information loss and execution time as vital metrics. The experimental evaluation shows that our proposed FC-BFO algorithm derives an optimal cluster as compared to the original BFO algorithm and existing clustering algorithms.

1. Introduction

An immense amount of personal data of an individual is collected by various organizations, namely, e-banking, online shopping, and medical and insurance agencies. Such collected data of an individual can be further analyzed digitally to find out useful information for various purposes such as medical research and market trend analysis. On one hand, data mining techniques are being used to find out the useful information from the collected data. As compared, the collected data might contain sensitive personal information. Therefore, mining such collected data can potentially disclose individuals' personal sensitive information. As a consequence, the personal sensitive information needs to be protected before conducting the data mining. For this reason, preserving the privacy of an individual becomes a prime research issue in the privacy preserving data mining [1].

To achieve the objective of privacy in the privacy preserving data mining, two main approaches have been proposed in the literature, namely, cryptographic approaches [2, 3] and the anonymization based approaches [4–18]. However, our focus here is on anonymization approaches only, owing to the lesser communication and computation cost of the same as compared to their cryptographic counterpart [2, 3].

Among the various anonymization approaches, l-diversity model [13] is one of the anonymization approaches used to preserve the privacy of an individual. However, any anonymization approach suffers from the information loss. The information loss calculates the difference between the original databases with the anonymized database. The information loss increases with the increase in the level of the generalization and/or suppression method [4]. Several metrics of information loss have been proposed in the literature [6, 10–13, 17, 18] to calculate the information loss.

An illustration of information loss has been discussed below. Consider a medical database as shown in Table 1, where the attributes such as age, gender, and zip code are considered as quasi-identifier (QI) whereas the attribute disease is considered as a sensitive attribute (SA). The quasi-identifier is of two types, namely, numeric and categorical. In Table 1, the attributes age and zip code are considered as numeric attribute whereas the attribute disease is considered as categorical attribute.

Table 2 shows a 3-diverse medical patient database of Table 1. Thus, by observing Table 2, a data miner could be able to conclude the sensitive attribute of a person, since all the diseases in equivalence class 1 are related to one type of diseases. Thus, the l-diversity model [13] is not able to protect the sensitive attribute disclosure and also suffers from the information loss.

Motivated by this observation, in this research, we attempt to further explore the issue of privacy preservation and information loss in the l-diversity model using a swarm intelligence optimization algorithm.

In recent years, various swarm intelligence optimization algorithms such as ant colony optimization (ACO) [19], particle swarm optimization (PSO) [20], and bacterial foraging optimization (BFO) [21, 22] have been used for data clustering. Among them, BFO is the recent swarm intelligence optimization algorithm based on the foraging behavior of bacteria. However, the BFO algorithm attends poor convergence and optimization behavior.

In addition, the literature [20, 23, 24] reports that fractional calculus (FC) has been used to boost up the optimization and convergence behavior of the PSO algorithm. We use FC to improve the convergence and optimization behavior of the BFO algorithm for the l-diversity model. However, the same has not been utilized in the l-diversity model for preserving the privacy in data mining to the best of our knowledge.

In this paper, we propose a new data clustering algorithm named as fractional calculus-bacterial foraging optimization (FC-BFO) algorithm. FC-BFO is used for grouping data by using optimization process of bacterial foraging. There are two main objectives of the research. First objective is to enhance optimization and convergence speed of the BFO algorithm by using fractional calculus. Second objective is to protect the sensitive value of an attribute and yield lesser information loss by using generalization and/or suppression. It is tested experimentally on two real datasets from UCI machine learning repository database [25]. The experimental analysis shows that our proposed FC-BFO algorithm achieves lesser information loss and better privacy as compared with state-of-the-art clustering algorithms such as Mondrian [7], greedy k-member [6], systematic clustering [11], recursive l-diversity, and entropy l-diversity [13] in three ways.

(i) First, FC-BFO algorithm uses the swarm intelligence optimization theory and generates an optimal cluster that is faster as compared with Mondrian [7], greedy k-member [6], systematic clustering [11], recursive l-diversity, and entropy l-diversity [13]. In addition, the literature [21, 22] reports that swarm intelligence optimization algorithm gives better result in terms of clustering accuracy and speed as compared to the existing algorithms.

(ii) Second, the BFO algorithm [21, 22] attends poor convergence and optimization behavior. In order to remove this limitation, we used a concept of fractional calculus in the BFO algorithm (as reported in the literature [23, 24]). The fractional order derivative calculus is used to enhance the optimization and convergence speed of the clustering algorithm.

(iii) Third, FC-BFO algorithm generates lesser information loss because of the generation of an optimal cluster via swarm intelligence optimization algorithm as compared with existing clustering algorithms.

Table 1: A medical patient database.

Identifier	Quasi-identifier			Sensitive attribute
Name	Age	Gender	Zip code	Disease
Harry	39	Male	2210	Rash
John	35	Male	2211	Psoriasis
Michal	31	Male	2210	Eczema
Sam	67	Male	2212	Ulcer
Alice	65	Female	2210	Flu
Bob	65	Female	2210	Heart problem

Table 2: A 3-diverse medical patient database.

Equivalence class	Age	Gender	Zip code	Disease
1	[30–40]	Person	221∗	Rash
	[30–40]	Person	221∗	Psoriasis
	[30–40]	Person	221∗	Eczema
2	[60–70]	Person	221∗	Ulcer
	[60–70]	Person	221∗	Flu
	[60–70]	Person	221∗	Heart Problem

1.1. Our Contribution and Plan of the Paper. In this paper, we propose a fractional calculus based bacterial foraging optimization (FC-BFO) algorithm for the l-diversity model in privacy preserving data mining. The main contributions of this paper are listed as follows.

(1) We improve the convergence and optimization behavior of the bacterial foraging optimization (BFO) algorithm by using a concept of fractional calculus (FC).

(2) We propose an objective function (OB) to identify the fitness of a cluster. A cluster with minimum OB is selected as an optimal cluster among the clusters. In addition, we propose a term privacy factor (PF) to compute the privacy of the anonymized dataset.

(3) We conduct experiments on the various anonymization algorithms to check the efficiency of our proposed algorithm with respect to the execution time and information loss. Our proposed algorithm shows lesser information loss as compared with the existing anonymization algorithms.

The rest of the paper is structured as follows. In Section 2 we survey the anonymization based clustering approaches. In Section 3 we discuss the information loss and the objective function for the l-diversity model. In Section 4 we present our proposed FC-BFO algorithm for the l-diversity model. In Section 5 we present the experimental results. Finally, we conclude in Section 6 with the implications of our results and propose the future research directions.

2. Related Work

In recent times, privacy preservation of an individual has been an active research area in privacy preserving data mining. Several anonymization approaches have been proposed in the literature [4–18] to preserve the privacy and enhance the data utility. The l-diversity model [13] is one of the widely used anonymization based approaches. The l-diversity model maintains l-diverse value in each anonymized group. However, the l-diversity model could not be able to protect the sensitive attribute disclosure and also suffer from the issue of information loss.

Recently, clustering based anonymization approaches such as systematic clustering for the l-diversity model [5], greedy k-member clustering algorithm [6], Mondrian algorithm [7], Loukides and Shao [8], weighted feature c-means clustering algorithm [9], one pass k-mean clustering algorithm [10], and systematic clustering approach for k-anonymity [11] were suggested. However, these anonymization based clustering approaches could not be able to produce an optimal cluster. Therefore, it results in the loss of information via anonymization in the clusters. In addition, this algorithm does not give an optimal solution with the increase in the dimension of the dataset. Thus, finding an optimal solution by using this algorithm becomes NP hard problem.

Among the various clustering algorithms, k-means algorithm [26, 27] is one of the simple and efficient algorithms with lesser computational overhead. However, k-means algorithm gets trapped in the local optimal solution. Thus, in order to find out a global optimal solution in the clustering problem, several swarm intelligence optimization algorithms such as ant colony optimization (ACO) [19], particle swarm optimization (PSO) [20], and bacterial foraging optimization (BFO) [21, 22] were suggested. These algorithms explore an initial population to generate optimal clusters by passing through a number of iterations. Amongst the optimization algorithm, BFO is the latest optimization algorithm applied in the data clustering to generate global optimal solution. However, BFO algorithm attends poor convergence and optimization behavior.

In the last two decades, fractional calculus (FC) has been applied in the area such as biology, physics, signal processing, control, and irreversibility as discussed in the literature [20, 23, 24]. It has been adopted to control the convergence behavior of the PSO algorithm [20]. Therefore, we used FC in our proposed approach to enhance the optimization and convergence behavior of the BFO algorithm.

In this paper, our focus is on the clustering based anonymization approaches for the l-diversity model [13] using swarm intelligence optimization algorithm.

3. Information Loss

Information loss is one of the important issues in an anonymization based approaches. The anonymization approaches via generalization and/or suppression always produce an information loss. Generally, the information loss should be lesser to achieve higher data utility. Two main approaches have been used to calculate the data utility. First approach calculates the amount of data utility remaining in the anonymized data. It includes metrics such as average size of the equivalence classes [12, 13, 17] and discernibility metric [18]. Second approach calculates the amount of data utility loss via data anonymization. It includes metrics such as information loss metrics [6, 10, 11]. In the first approach, if the original data has lower data utility then the anonymized data has lower data utility. Therefore, we adopted the second approach to measure the data utility. In addition, the information loss metrics as discussed in [6, 11] have been used recently. Thus, we utilize the information loss metrics [6, 11] in our work.

Let N denote a set of records in a database with n numeric quasi-identifiers $r_1, r_2, \ldots, r_n$ and c categorical quasi-identifier $d_1, d_2, \ldots, d_c$. Let N be the total number of records partitioned into p clusters. Let $\partial = \{m_1, m_2, \ldots, m_p\}$ be a partitioning of N. Let Mi max and Mi min be maximum and minimum values of the attributes in the cluster. Let NMi max and NMi min be the maximum and minimum value of the attributes in a database. Therefore, the total information loss (IL) of N is the sum of the information loss of each m_i $(i = 1, 2, \ldots, p)$:

$$\mathrm{IL}(N) = \sum_{i=1}^{p} \mathrm{IL}(m_i),$$

$$\mathrm{IL}(m) = |m| \left(\sum_{i=1}^{n} \frac{Mi\max - Mi\min}{NMi\max - NMi\min} + \sum_{j=1}^{c} \frac{H\left(\wedge\left(\cup d_j\right)\right)}{H\left(Td_j\right)} \right), \quad (1)$$

where $|m|$ is the number of records in a cluster. $(\wedge(\cup d_j))$ is a union of the categorical value in a cluster and also specifies the lowest common ancestor for categorical attribute. $H(Td_j)$ is the height of the taxonomy tree T. The taxonomy tree for the quasi-identifier, namely, gender and zip code, are shown in Figures 1 and 2, respectively.

Example 1. The example of finding the information loss for the l-diversity model is discussed as follows. Table 1

shows a medical patient database and Table 2 shows a 3-diverse medical patient database of Table 1 (as discussed in Section 1). Here, we assume the attributes age and zip code as a numerical attribute and gender as a categorical attribute. The attribute disease is taken as a sensitive attribute and the other attributes such as age, gender, and zip code are taken as a quasi-identifier.

The cluster 1 consists of the first three records and the cluster 2 consists of other three records as shown in Table 2. The information loss of Table 2 is calculated by relating to the original Table 1. Thus, by using Table 1, we calculate the minimum and the maximum value from the cluster 1. Similarly, we calculate the minimum and the maximum value from the complete Table 1. The minimum and the maximum value for the cluster 1 are 31 and 39, respectively. In the same way, the minimum and the maximum values for the cluster 2 are 65 and 67, respectively. The value used for the categorical attribute gender and zip code is based on the height of the taxonomy tree as shown in Figures 1 and 2.

Thus, the total information loss (IL) of the l-diverse Table 2 by using information loss metrics as discussed in [6, 11] is arrived at as follows:

$$\begin{aligned} &\mathrm{IL}(N) \\ &= |3| * \left(\frac{39-31}{67-31} + 1 + 1\right) + |3| * \left(\frac{67-65}{67-31} + 1 + 1\right) \\ &= 12.84. \end{aligned} \tag{2}$$

3.1. An Objective Function and Privacy Factor for the l-Diversity Model. In this section, we propose the objective function (OB) and privacy factor (PF) of the anonymized dataset for the l-diversity model. The OB computes the fitness and PF computes the privacy of the anonymized dataset. The OB finds an optimal cluster which results in the lesser information loss and higher privacy. Thus, we incorporated both the objectives, namely, information loss and privacy, in the design of OB. The objective function (OB) is denoted by

$$\mathrm{OB} = \alpha * \mathrm{IL}(N) + \beta * (1 - \mathrm{PF}), \tag{3}$$

where OB is an objective function, IL(N) is an information loss of the anonymized dataset, PF is the privacy factor introduced in our work to improve the privacy of the dataset, and α and β are the weighting factor (which are generally given as 0.5 and 0.5, resp.); PF contains two parameters: first is the total number of tuples changed during the anonymization process and second is the number of tuples in the dataset.

The PF of the dataset can be computed based on the tuples changed during an anonymization process. Therefore, the privacy factor (PF) of the anonymized database is

$$\mathrm{PF} = \frac{1}{p} \sum_{i=1}^{P} \frac{T_i'}{T_i}, \tag{4}$$

where T_i' is the total number of cells with anonymized value in each tuples; T_i is the total number of cells in an equivalence class or cluster of a database; P indicates the total number of clusters. If the value of T_i' is higher, then the PF of data will be higher. This is because the number of changed cells value T' is higher. However, our objective in this study is to minimize the information loss.

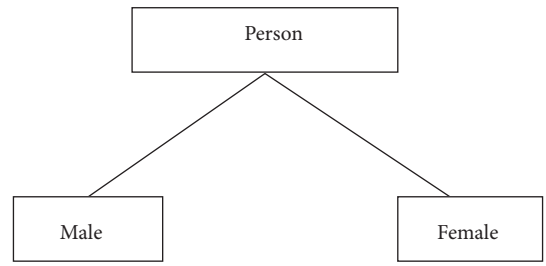

Figure 1: Taxonomy tree of gender.

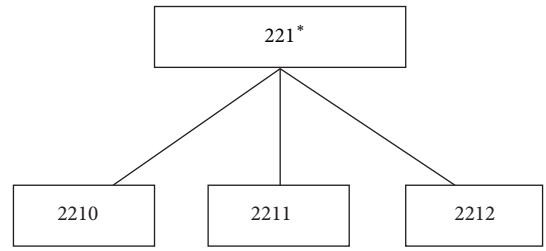

Figure 2: Taxonomy tree of zip code.

The privacy factor (PF) of Table 2 is computed as follows:

$$\mathrm{PF} = \frac{1}{2}\left(\frac{9}{12} + \frac{9}{12}\right) = 0.75. \tag{5}$$

The objective function (OB) is computed as follows:

$$\mathrm{OB} = (0.5) * (12.84) + (0.5) * (1 - 0.75) = 6.54. \tag{6}$$

It has been observed from Table 3 that, with the increase in anonymized cell value, the PF value increases. However, the OB value decreases with the increase in the anonymized cell value. The PF value varies between 0 and 1. Case 1 represents that no cell in the database has been anonymized. Thus, it represents the original database before the data anonymization. In addition to this, the data utility and data privacy have been represented on the scale from 1 to 5, in which, the smaller value on the scale represents the smaller data utility and data privacy. In Case 1, the data utility and data privacy are 5 and 1, respectively. This indicates that the data utility is higher in Case 1 as compared to all other cases. However, the data privacy is lesser as compared to all other cases. This observation indicates the tradeoff of data utility with data privacy. Thus, we conclude that the higher the cell anonymization (use of generalization and/or suppression in a cell value of the database), the higher the PF and the lesser the OB.

4. FC-BFO Algorithm for the l-Diversity Model

In this section, we present a new clustering algorithm named as FC-BFO algorithm for the l-diversity model in privacy

Table 3: The tradeoff between data utility and data privacy.

Case number	Anonymized cell value	PF	OB	Data utility (on scale from 1 to 5)	Data privacy (on scale from 1 to 5)
1	0	0	6.905	5	1
2	3	0.25	6.78	4	2
3	6	0.5	6.655	3	3
4	9	0.75	6.53	2	4
5	12	1	6.405	1	5

preserving data mining. The BFO algorithm [21, 22] consists of three major steps such as chemotaxis, reproduction, and elimination-dispersal. These steps are used to generate optimal clusters. However, the BFO algorithm attains poor convergence and optimization behavior. Thus, we applied FC in the chemotaxis step of the BFO algorithm.

In this work, we used two main steps such as initial solution-encoding and grouping and BFO algorithm.

4.1. Initial Solution-Encoding and Grouping. The initial solution-encoding and grouping step consists of four main substeps. It includes solution-encoding and grouping, distance matrix computation, grouping with the constraint of k and l parameter, and objective function (OB) computation.

The solution-encoding and grouping step finds the central record randomly from a set of records. After finding the central record, the distance matrix computation step is applied. The distance matrix is calculated based on the distance between the randomly selected central records with respect to other records in a database. Generally, two types of attributes are present in the database such as numerical attributes and categorical attributes.

First, we derive the distance matrix for the numerical attributes. Let r_1 and r_2 be a numerical value of an attribute. Then, the normalized distance between two values r_1 and r_2 is given as follows:

$$\text{dist}_n(r_1, r_2) = \frac{|r_1 - r_2|}{|R|}, \qquad |R| = r\max - r\min, \tag{7}$$

where r max and r min are the maximum and the minimum values of an attribute.

Second, we derive the distance matrix for the categorical attributes (as shown in (8)). Let d be a value of a categorical attribute. Consider

$$\text{dist}_c(d) = \frac{H\left(\wedge\left(\cup d_j\right)\right)}{H\left(Td_j\right)}, \tag{8}$$

where $H(\wedge(\cup d_j))$ is the union of the categorical attribute whose value rooted at the lowest level in a taxonomy tree. $H(Td_j)$ is the height of the taxonomy tree for the categorical attributes. Figures 1 and 2 show the taxonomy tree (as discussed in Section 3). Thus, the formula for calculating the distance for the numerical and the categorical attributes is shown as follows:

$$\text{dist}(p, N) = \text{dist}_n(r_1, r_2) + \text{dist}_c(d), \tag{9}$$

where p is a randomly selected central record of a cluster and N is the number of records in a database. After calculating the distance matrix, every record is grouped into relevant cluster based on minimum distance.

In the third step, we check the k and l parameters in every cluster using grouping with the constraint of l-diverse step. If the solution in the l-diverse table is satisfied with the k and l parameter constraint, then the fitness function is computed (as discussed in Section 3.1) by using the fourth step called objective function (OB). The OB finds out the information ioss (IL) in the l-diverse table. The above steps are repeated for all the initial population till an initial population is changed. Once the initial population is changed, the BFO algorithm is applied to obtain the best cluster with the minimum objective function (OB).

4.2. Bacterial Foraging Optimization (BFO) Algorithm. The changed initial population (as discussed in Section 4.1) is applied on BFO algorithm [21, 22] to get an optimal cluster. The BFO algorithm is a stochastic global search optimization algorithm that uses foraging behavior of bacteria. It consists of three steps such as chemotaxis, reproduction, and elimination-dispersal step which have been used to find an optimal cluster.

Generally, the groups of bacteria try to find a food using a foraging behavior. A bacterium moves in two alternate modes such as tumbling and swimming in its lifetime. Such alternate mode is called chemotactic step. The chemotactic step helps a bacterium to move in random direction to search for food. The bacterium received sufficient amount of food during the chemotactic step. After the chemotactic step, the bacterium goes into the reproduction step. The reproduction step changes the population of the bacteria. In the elimination-dispersal step, the weak bacterium is eliminated or killed from the bacteria population. This keeps the population of the bacteria constant.

The BFO algorithm is explained as follows.

Let the total population of a bacteria be represented as

$$P(l, m, n) = \{(l, m, n) \mid k = 1, 2, \ldots, S\}, \tag{10}$$

where l, m, and n represent a chemotaxis step, reproduction step, and an elimination-dispersal step, respectively. Let S be the number of centroid set in the population. Let k be the position of the bacterium related to the attribute value of cluster centroid.

The chemotaxis step consists of two modes such as swimming and tumbling. Bacteria in the swimming step move in the predefined direction. As compared, it moves

in different directions in the tumbling step. Let $\theta^i(l,m,n)$ represent ith bacterium at lth chemotaxis, mth reproductive, and nth elimination-dispersal step. Consdier

$$\theta^i(l+1,m,n) = \theta^i(l,m,n) + B(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}}, \quad (11)$$

where Δ is a random direction of a vector and its elements lie in the range of $[-1,1]$. $B(i)$ is the chemotactic step size which is taken in the random direction for the tumble step. The generation of new neighbor in BFO algorithm is purely designed for numerical values. However, in the anonymization process of the l-diversity model, we need to consider both attributes such as numerical and categorical attributes. Therefore, with the consideration of these points, the neighbor solution generation procedure is adopted with respect to our requirement. In (11), $\theta^i(l,m,n)$ is the current centroid set.

In reproduction step, the better solutions are identified and they are kept in the same location. Thus, the healthy bacteria are identified based on higher fitness function (as discussed in Section 3.1). The healthy bacteria are divided into two bacteria, and then they are placed in the same location. This keeps the population size of the bacteria constant. In the elimination-dispersal step, the noneffective solutions are removed from the problem space. The new solutions are added into the problem space. The new solutions are generated based on the idea of the random addition.

4.3. Fractional Calculus. Fractional calculus (FC) [20, 23, 24] is one of the branches of applied mathematics. It plays a significant role in boosting the performance of the numerous algorithms such as modeling, curve fitting, filtering, pattern matching, and edge detection.

The FC solves integral and derivative equation. The fractional order differential and integral equation have been solved by using Laplace transforms. Three steps have been used to solve the integral and derivative equation. The first step finds the Laplace transform of the equation. The second step solves the transform of unknown function, and the third step finds the inverse Laplace to obtain the desired solution.

The Grunwald-Letnikov is an alternative technique that utilized the concept of fractional differential. Please refer to [23] for more details. The Grunwald-Letnikov report that the fractional order derivative needs an infinite number of terms for an integer order derivative and possess inherent memory capacity. For this reason, we add the fractional order derivative to the chemotaxis step of the BFO algorithm.

Thus, we incorporated FC in the chemotaxis step of the BFO algorithm. Equation (11) shows the chemotaxis step, which is rearranged and shown as follows:

$$\theta^i(l+1,m,n) - \theta^i(l,m,n) = B(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}}. \quad (12)$$

In Equation (10), the left side represents the discrete version of the derivative of order $\hbar = 1$. Assuming $T = 1$ as discussed in the literature [20] leads to

$$D^{\hbar}\left[\theta^i(l+1,m,n)\right] = B(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}}. \quad (13)$$

The FC perspective directs to a smoother variation and a longer memory effect. To study the behavior, we carried a set of simulations on the values of $\hbar$ which varies from 0 to 1 with the increment of $\Delta\hbar = 0.1$. Thus, considering the first $r = 4$ terms of the differential derivative as discussed in [20] yields

$$\begin{aligned}
&\theta^i(l+1,m,n) - \hbar\left[\theta^i(l,m,n)\right] - \frac{1}{2}\hbar\left[\theta^i(l-1,m,n)\right] \\
&\quad - \frac{1}{6}(1-\hbar)\left[\theta^i(l-2,m,n)\right] \\
&\quad - \frac{1}{24}\hbar(1-\hbar)(2-\hbar)\left[\theta^i(l-3,m,n)\right] \\
&= B(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}}.
\end{aligned} \quad (14)$$

Thus, the changed initial population for the BFO algorithm is represented as follows:

$$\begin{aligned}
&\theta^{\hbar}(l+1,m,n) \\
&\quad = \hbar\left[\theta^i(l,m,n)\right] + \frac{1}{2}\hbar\left[\theta^i(l-1,m,n)\right] \\
&\quad + \frac{1}{6}\hbar(1-\hbar)\left[\theta^i(l-2,m,n)\right] \\
&\quad + \frac{1}{24}\hbar(1-\hbar)(2-\hbar)\left[\theta^i(l-3,m,n)\right] \\
&\quad + B(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}}.
\end{aligned} \quad (15)$$

In (15), the term $B(i)(\Delta(i)/\sqrt{\Delta^T(i)\Delta(i)})$ is used to generate a new position. The same procedure is followed to generate a new position for the numerical value. The solution for the categorical attribute is generated based on the hierarchical tree. The value of a current position is identified and a random value -1 or 1 is generated. If the value is in the negative direction, then the child node is taken as new position. As compared, if the value is in the positive direction, the parent node is taken as new position.

Pseudocodes 1 presents the FC-BFO algorithm. In FC-BFO algorithm, all the bacteria use three iteration loops such as chemotaxis N_c, reproduction N_r, and elimination-dispersal N_e and generate various clusters $C = \{c_1, c_2, \ldots, c_p\}$ in the search space. Thus, the total iteration in the optimization process will be $N_{\text{it}} = N_c \times N_r \times N_e$. If the value of N_{it} is larger, the optimization process is better and the computational complexity is higher. If the value of N_{it} is smaller, the optimization process is lesser and the algorithm could not be able to get the global optimum solution. On the other hand, if the value of N_s is larger, the bacteria will

Input: Dataset D Output: Clusters $C = \{c_1, c_2, \ldots, c_p\}$
Parameters:
l chemotaxis step; m reproduction step; n elimination-dispersal step; P dimension of the search space; S total number of bacteria in the population; N_c number of chemotactic steps; N_r number of reproduction steps; N_e number of elimination-dispersal steps; N_s swim step; P_e probability of elimination-dispersal; $B(i)$ step size during tumble.
FC-BFO Algorithm:
(1) Initialize the parameters $P, S, N_c, N_r, N_e, N_s, P_e, B(i)$ where $i = 1, 2, \ldots, S$ and θ^i
(2) Elimination-dispersal loop $n = n + 1$: N_e
(3) Reproduction loop $m = m + 1$: N_r
(4) Chemotaxis loop $l = l + 1$: N_c
(5) Apply a chemotaxis step for the ith bacterium (where, $i = 1, 2, \ldots, S$)

$$\theta^i(l+1, m, n) = \theta^i(l, m, n) + B(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}}$$

(6) Calculate fitness function $\mathrm{OB}(i, l, m, n)$
(7) Store the value as $\mathrm{OB}_{\mathrm{old}} = \mathrm{OB}(i, l, m, n)$ to find a better fitness function
(8) Tumble: Generate a random vector $\Delta(i)$ with each element of a numerical attribute
(9) Generate a random direction of categorical attribute
(10) Move: Make a move in the direction of the tumble for the bacterium i

$$\theta^{\hbar}(l+1, m, n) = \hbar\left[\theta^i(l, m, n)\right] + \frac{1}{2}\hbar\left[\theta^i(l-1, m, n)\right] + \frac{1}{6}\hbar(1-\hbar)\left[\theta^i(l-2, m, n)\right] + \frac{1}{24}\hbar(1-\hbar)(2-\hbar)\left[\theta^i(l-3, m, n)\right] + B(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}}$$

(11) Compute fitness function $\mathrm{OB}(i, l+1, m, n)$ with $\theta^i(l+1, m, n)$
(12) Swim:
(13) Initialize the swim counter $t = 0$
(14) While $t < N_s$ do
(15) $t = t + 1$
(16) if $\mathrm{OB}(i, l+1, m, n) < \mathrm{OB}_{\mathrm{old}}$ then
(17) $\theta^i(l+1, m, n) = \theta^i(l+1, m, n) + B(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}}$
(18) $\mathrm{OB}_{\mathrm{old}} = \mathrm{OB}(i, l+1, m, n)$
(19) else $t = N_s$
(20) End if
(21) End while
(22) If $l < N_c$, then go to step 4.
(23) Reproduction (D, OB) of S bacteria with higher OB.
(24) If $m < N_r$, then go to step 3. Start again with the Chemotaxis step.
(25) Elimination-dispersal (D, P_e): eliminate the bacterium that has highest fitness value and disperse it to a random location
(26) For $i = 1$ to P: C
(27) Anonymize (D)
(28) End for
(29) End

Pseudocode 1: The pseudocode of FC-BFO algorithm.

move in different direction and the computational complexity increases. Similarly, if the value of the elimination-dispersal loop N_e is larger, the random exhaustive search ability gets degraded. The step size $B(i)$ used in the iteration process should not be too small or too large. If the step size is too small, the convergence speed is slow and algorithm gets trapped in a local optimum solution. If the step size is too large, the algorithm will miss the position of the local optima solution. Therefore, the value should be chosen properly to get out from the local optima to the global optimum solution.

Pseudocodes 2 presents an anonymization function for the FC-BFO algorithm. It uses the bottom-up generalization approach to anonymize the original data. An equivalence class has been constructed from the optimal cluster by considering the constraints of size of k and l. The anonymization function has been used to generate the anonymized data.

```
Function: Anonymize (D, D′, C, E, T, T′, n, k, l)
// D is the original dataset;
// D′ is the anonymized dataset;
// C is the number of optimal cluster generated using FC-BFO algorithm;
// E is the number of an equivalence class;
// T is the number of tuples in the dataset D;
// T′ is the number of tuples in the dataset D′;
// n is the number of records in the dataset D;
// k is the constraint in the k-anonymity;
// l is the constraint in the l-diversity;
Begin
 (1) Create an equivalence classes E with the constraint size of k and l
     from D
 (2) Number of tuples T in each E < number of records n in D
 (3) If the size of equivalence class E is < the constraint k and l then
 (4) Generalize all the tuples T which resulting T′ tuples in an equivalence
     class E
 (5) Return the anonymized result in D′ with C
End
```

Pseudocode 2: The pseudocode of anonymization function for the FC-BFO algorithm.

5. Experimental Evaluation

In this section, we will present the effectiveness of our proposed FC-BFO approach with respect to the parameters such as information loss and execution time. The main objective of the experiment is to achieve higher data utility with lesser information loss and execution time. We also compare our proposed approach with five most notable approaches, namely, Mondrian algorithm [7], greedy k-member algorithm [6], systematic clustering approach [11], recursive l-diversity, and entropy l-diversity [13]. The experiment is implemented in Java with JDK 1.6 in a system configured with Intel core i5 processor, 4 GB RAM, and 500 GB hard disk.

5.1. Experiment Setup. We used two benchmark datasets such as ADULT database and the university database from UCI machine learning repository [25] for the experimentation. The ADULT database contains 32561 records and 15 attributes. Out of them, we retain only attribute, namely, age, race, marital status, sex, fnlwt, and occupation. The attribute age and fnlwt are numeric attributes whereas race, marital status, sex, and occupation are the categorical attributes. The attribute occupation is taken as a sensitive attribute in the database.

The university database contains 285 records and 17 attributes. For the experimentation purpose, we retain only the attribute, namely, state, location, number of students, number of applicants, university name, and expenses as quasi-identifiers. Whereas, the attribute university-name is taken as a sensitive attribute.

5.2. Parameter Selection. The appropriate selection of parameter is one of the important factors in the optimization algorithms. We used the following parameter values in our experimentation as discussed in the literature [22] to achieve good convergence: the chemotaxis step size $B(i) = 0.1$; swim step $N_s = 4$; number of chemotactic step $N_c = 100$; number of reproduction step $N_r = 4$; number of elimination-dispersal step $N_e = 2$; probability of elimination-dispersal $P_e = 0.25$.

5.3. Methodology of Evaluation. We compare the efficiency of our proposed FC-BFO clustering algorithm with two parameters, namely, iteration and objective function. Figures 3, 4, and 5 show the objective function (OB) obtained for various quasi-identifiers (QI) size with a fixed value of k, $l = 5$. Initially, we run the experiment for the various iterations, namely, 5, 10, 15, and 20.

From Figures 3 to 5, we can find that when the size of the QI increases, the OB function also increases. Thus, we conclude that the lesser the QI size, the lesser will be the information loss and objective function (OB).

We run our proposed FC-BFO algorithm by varying the size of quasi-identifier attributes. In this work, the quasi-identifier attributes of sizes 2, 3, 4, 5, and 6 have been used. Here, we kept the value of k and l equal to 5. The total information loss and execution time are computed during each run of the experiment on two different datasets [25]. Figure 6 shows that our proposed FC-BFO and BFO approach for the l-diversity model achieve lesser information loss as compared with a state-of-the-art clustering approaches, namely, Mondrian algorithm [7], greedy k-member approach [6], systematic clustering approach [11], recursive l-diversity [13], and entropy l-diversity [13] on the adult dataset [25].

The reason for yielding lesser information loss of our proposed FC-BFO and BFO algorithm (as shown in Figures 6 and 7) as compared with state-of-the-art approaches [6, 7, 11, 13] is as follows.

(i) Our proposed FC-BFO algorithm for the l-diversity model shows improvement over proposed BFO algorithm because of the following reasons. First, BFO

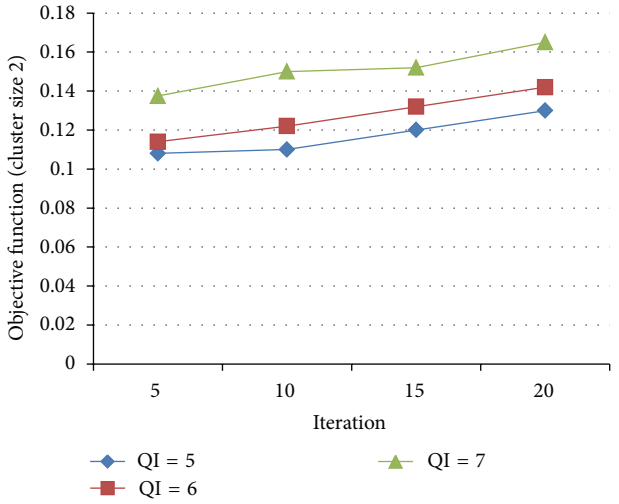

Figure 3: Iteration versus objective function for cluster size 2 with varied quasi-identifier (QI).

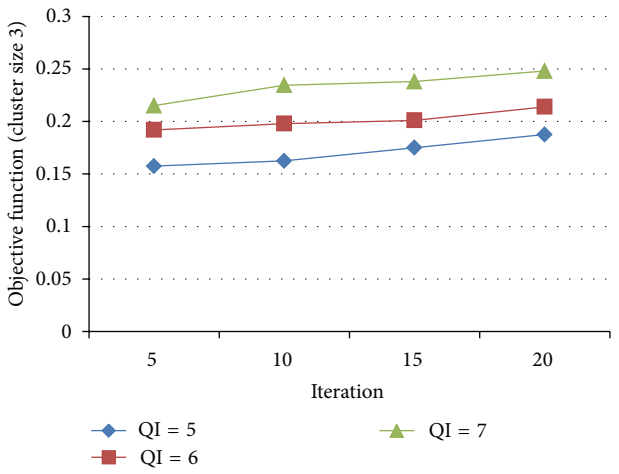

Figure 4: Iteration versus objective function for cluster size 3 with varied quasi-identifier (QI).

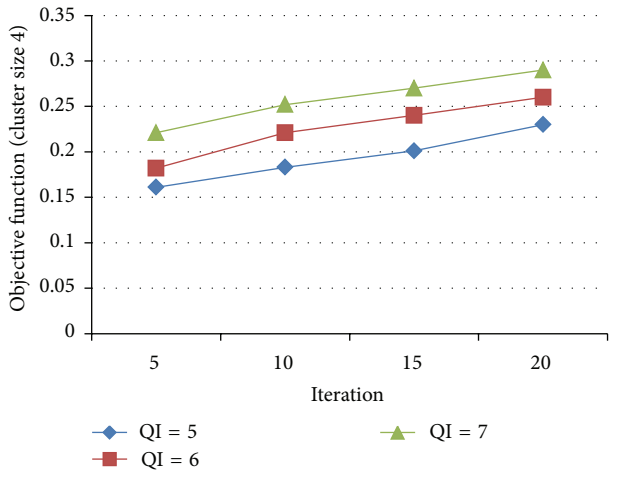

Figure 5: Iteration versus objective function for cluster size 4 with varied quasi-identifier (QI).

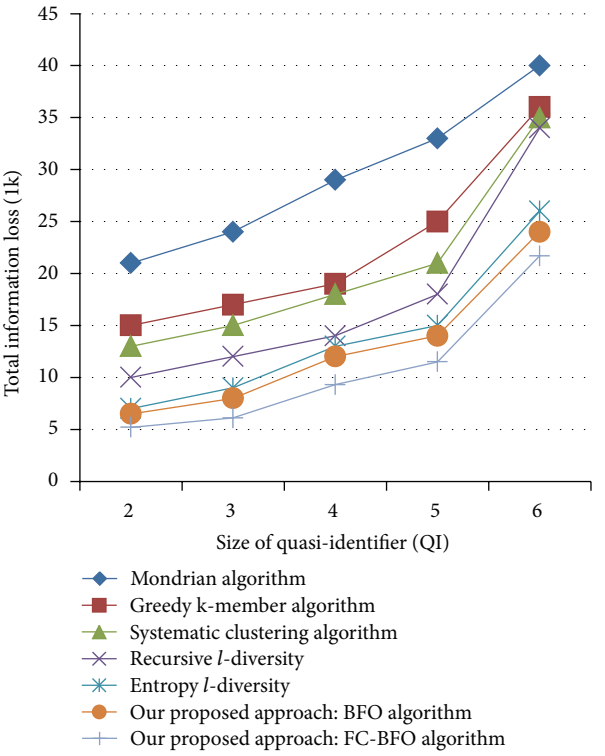

Figure 6: Total information loss for the adult dataset.

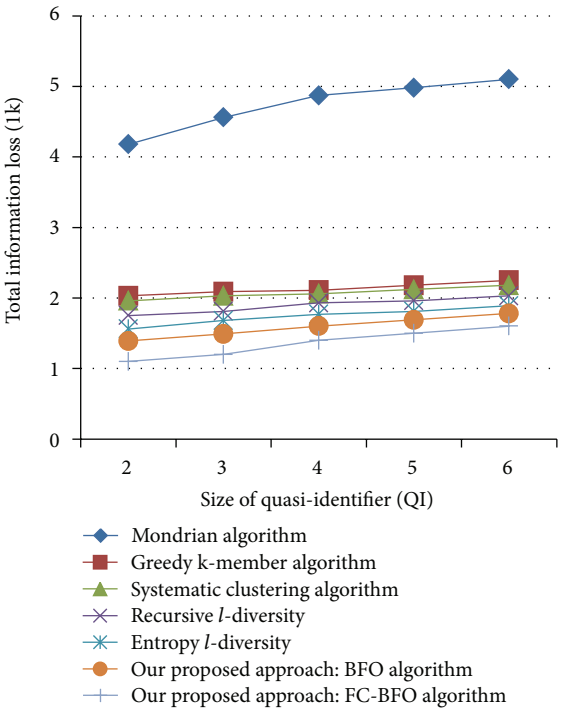

Figure 7: Total information loss for the university dataset.

algorithm has poor optimization capability and convergence behavior. Second, we use the fractional calculus in chemotaxis step of the BFO algorithm to improve the optimization and convergence behavior of the BFO algorithm, since the fractional calculus uses the concept of integral and derivative to find the solution. Third, we use a Laplace transform for the fractional order differential equation. As a result, the convergence speed of the BFO algorithm increases in creating the clusters as compared with existing approaches [6, 7, 11, 13].

(ii) Nonetheless, the information loss of each of the algorithms increases with respect to the increase in the quasi-identifier for a fixed k and l value. There are two main reasons behind this. The first reason is the higher the k value, the higher the information loss. The second reason is the higher the value of l, the better the privacy. Therefore, in real life, we need to adopt an optimal value of k for lesser information loss and higher value of l for better privacy.

(iii) The Mondrian algorithm [7] makes use of the concept of global recoding. Our proposed approach employs the concept of local recoding. However, literature [16] reports that local recoding is better compared to the global multidimensional recoding. Therefore, Mondrian algorithm generates higher information loss comparing to our proposed approach.

(iv) The Greedy k-member algorithm [6] is slow in building the cluster and also sensitive to the outliers records. Therefore, the information loss increases due to the presence of the outliers records in the cluster. On the other hand, our proposed FC-BFO approach is faster in searching strategy and makes use of nature-inspired algorithm [21, 22] for building the clusters.

Our observations related to the comparison of our proposed FC-BFO and BFO approaches with the existing approaches [6, 7, 11, 13] with respect to the execution time (as shown in Figures 8 and 9) are listed as follows.

(i) The Greedy k-member algorithm [6] takes higher time in selecting the records from the input dataset. Therefore, the execution time of greedy k-member algorithm [6] is higher as compared with the existing clustering algorithm [7, 11, 13].

(ii) The Mondrian algorithm [7] procures lesser time in selecting and faster time in partitioning the records. Therefore, the execution time of the Mondrian algorithm [7] is lesser as compared with the other existing algorithms [6, 13, 23].

(iii) The entropy l-diversity model uses the pruning step for the searching. Therefore, it takes lesser execution time when compared with the algorithms [6, 7, 11].

(iv) The BFO algorithm takes lesser searching and building time to build the clusters of size k and l. Therefore, our proposed FC-BFO and BFO algorithm show lesser execution time when compared with the existing algorithms [6, 11, 13].

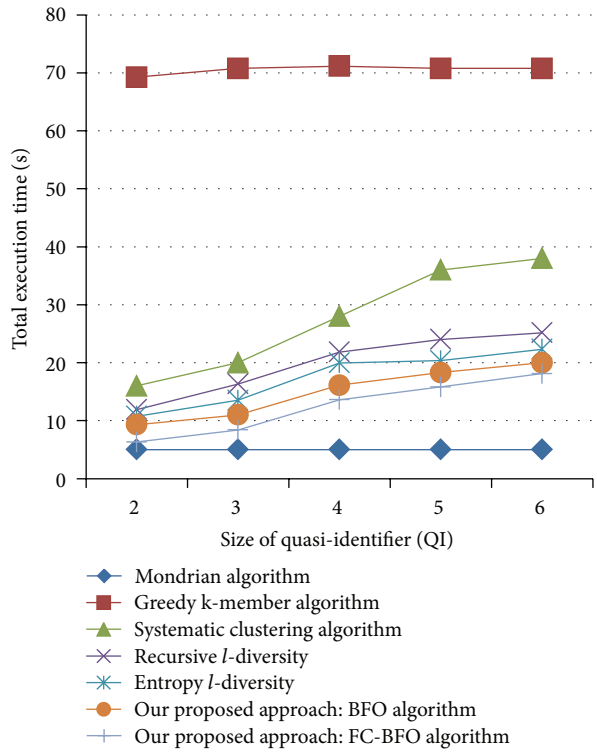

Figure 8: Total execution time for the adult dataset.

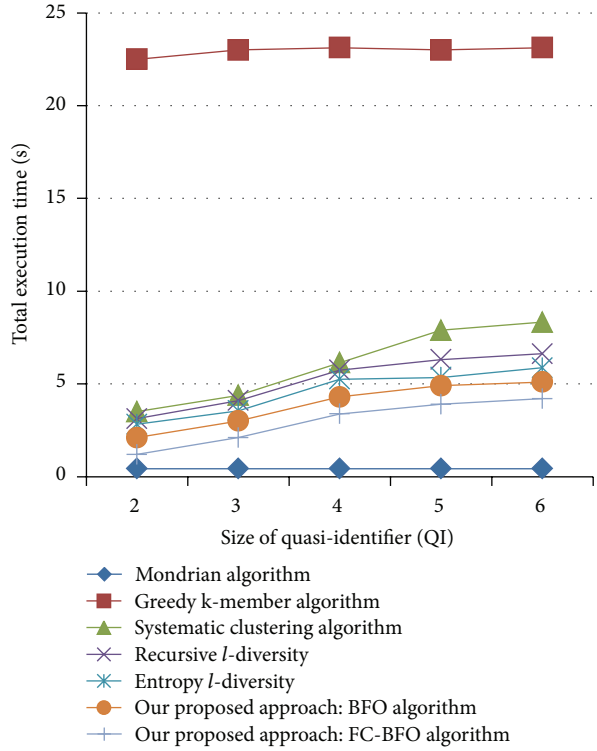

Figure 9: Total execution time for the university dataset.

(v) The execution time of our proposed FC-BFO algorithm is lesser when compared with the BFO algorithm. The reasons for yielding the lesser execution time are as follows. The first reason is that we adopted the concept of FC in the BFO. Secondly, the FC is used to boost the performance of the algorithm. Finally, FC controls the convergence and optimization behavior of the algorithm.

Our proposed FC-BFO approach attains higher execution time and lesser information loss when compared with a Mondrian algorithm [7]. The primary research issue of privacy preserving data mining is to achieve lesser information loss, since we are fulfilling the primary research issue of minimizing the information loss. Therefore, we can conclude that our proposed FC-BFO algorithm for the l-diversity model generates lesser information loss when compared with the existing clustering approaches.

6. Conclusion

In this paper, we present a clustering approach based on the fractional calculus-bacterial foraging optimization algorithm in the l-diversity model to minimize the information loss. The proposed approach uses the concept of fractional calculus in the chemotaxis step of the bacterial foraging optimization algorithm. The experimental results report that our proposed clustering approach show lesser information loss as compared with Mondrian algorithm, greedy k-member algorithm, systematic clustering algorithm, recursive l-diversity, and entropy l-diversity.

In privacy preserving data mining, information loss is the prime research issue addressed in the existing literature. However, various methods have been proposed in the literature for minimizing the information loss such as t-closeness [14], (α, k) anonymity [15], and anatomy [12]. Therefore, our further work is to apply the fractional calculus based bacterial foraging optimization algorithm to t-closeness [14], (α, k) anonymity [15], and anatomy [12] methods.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for reviewing their paper and for the suggestions to improve the quality of their paper.

References

[1] Y. Lindell and B. Pinkas, "Privacy preserving data mining," *Journal of Cryptology*, vol. 15, no. 3, pp. 177–206, 2002.

[2] M. Upmanyu, A. M. Namboodiri, K. Srinathan, and C. V. Jawahar, "Efficient privacy preserving K-means clustering," in *Intelligent and Security Informatics*, vol. 6122 of *Lecture Notes in Computer Science*, pp. 154–166, 2010.

[3] G. Jagannathan, K. Pillaipakkamnatt, R. N. Wright, and D. Umano, "Communication-efficient privacy preserving clustering," *Transactions on Data Privacy*, vol. 3, no. 1, pp. 1–25, 2010.

[4] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[5] E. Kabir, H. Wang, E. Bertino, and Y. Chi, "Systematic clustering method for l-diversity model," in *Proceeding of 21st Australasian Conference on Database Technologies*, vol. 104, pp. 93–102, Australian Computer Society, Darlinghurst, Australia, 2010.

[6] J. W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," in *Proceedings of the 12th International Conference on Database Systems for Advanced Applications*, pp. 188–200, Springer, 2007.

[7] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, p. 25, April 2006.

[8] G. Loukides and J. Shao, "Capturing data usefulness and privacy protection in K-anonymisation," in *Proceedings of the ACM symposium on Applied computing (SAC '07)*, pp. 370–374, March 2007.

[9] C.-C. Chiu and C.-Y. Tsai, "A k-anonymity clustering method for effective data privacy preservation," in *Proceeding of the 3rd International Conference on Advanced Data Mining and Application*, vol. 4632, pp. 89–99, 2007.

[10] J.-L. Lin and M.-C. Wei, "An efficient clustering method for k-anonymization," in *Proceedings of the International Workshop on Privacy and Anonymity in Information Society (PAIS '08)*, pp. 46–50, March 2008.

[11] M. E. Kabir, H. Wang, and E. Bertino, "Efficient systematic clustering method for k-anonymization," *Acta Informatica*, vol. 48, no. 1, pp. 51–66, 2011.

[12] X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preservation," in *Proceedings of the 32nd International Conference on Very Large Data Bases*, pp. 139–150, 2006.

[13] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, p. 24, Atlanta, Ga, USA, April 2006.

[14] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: privacy beyond k-anonymity and l-diversity," in *Proceedings of the 23rd International Conference on Data Engineering*, pp. 106–115, IEEE, Istanbul, Turkey, April 2007.

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

[16] J. Goldberger and T. Tassa, "Efficient anonymizations with enhanced utility," *Transactions on Data Privacy*, vol. 3, no. 2, pp. 149–175, 2010.

[17] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain k-anonymity," in *Proceeding of the ACM SIGMOD International Conference on Management of Data (SIGMOD '05)*, pp. 49–60, June 2005.

[18] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proceedings of the 21st International Conference on Data Engineering (ICDE '05)*, pp. 217–228, April 2005.

[19] M. Dorigo and L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem,"

IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, pp. 53–66, 1997.

[20] E. J. S. Pires, J. A. T. MacHado, P. B. de Moura Oliveira, J. B. Cunha, and L. Mendes, "Particle swarm optimization with fractional-order velocity," *Nonlinear Dynamics*, vol. 61, no. 1-2, pp. 295–301, 2010.

[21] S. Das, A. Biswas, S. Dasgupta, and A. Abraham, "Bacterial foraging optimization algorithm: theoretical foundations, analysis and applications," *Foundation of Computational Intelligence*, vol. 3, pp. 23–55, 2009.

[22] M. Wan, L. Li, J. Xiao, C. Wang, and Y. Yang, "Data clustering using bacterial foraging optimization," *Journal of Intelligent Information Systems*, vol. 38, no. 2, pp. 321–341, 2012.

[23] M. S. Couceiro, R. P. Rocha, N. M. F. Ferreira, and J. A. T. Machado, "Introducing the fractional-order Darwinian PSO," *Signal, Image and Video Processing*, vol. 6, no. 3, pp. 343–350, 2012.

[24] J. A. T. Machado, M. F. Silva, R. S. Barbosa et al., "Some applications of fractional calculus in engineering," *Mathematical Problems in Engineering*, vol. 2010, Article ID 639801, 34 pages, 2010.

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

[26] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceeding of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, University of California Press, 1967.

[27] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceeding of the 18 th annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007.