G.K. Shailaja* and C.V. Guru Rao

Opposition Intensity-Based Cuckoo Search Algorithm for Data Privacy Preservation

https://doi.org/10.1515/jisys-2018-0420
Received October 19, 2018; previously published online June 14, 2019.

Abstract: Privacy-preserving data mining (PPDM) is a novel approach that has emerged in the market to take care of privacy issues. The intention of PPDM is to build up data-mining techniques without raising the risk of mishandling of the data exploited to generate those schemes. The conventional works include numerous techniques, most of which employ some form of transformation on the original data to guarantee privacy preservation. However, these schemes are quite multifaceted and memory intensive, thus leading to restricted exploitation of these methods. Hence, this paper intends to develop a novel PPDM technique, which involves two phases, namely, data sanitization and data restoration. Initially, the association rules are extracted from the database before proceeding with the two phases. In both the sanitization and restoration processes, key extraction plays a major role, which is selected optimally using Opposition Intensity-based Cuckoo Search Algorithm, which is the modified format of Cuckoo Search Algorithm. Here, four research issues, such as hiding failure rate, information preservation rate, and false rule generation, and degree of modification are minimized using the adopted sanitization and restoration processes.

Keywords: PPDM, sanitization, restoration, key extraction, cuckoo search, opposition intensity.

1 Introduction

Nowadays, the quantity of information produced and conveyed among governments, institutions, and other firms has extremely risen [1, 13]. In addition, with the speedy improvement of data-mining (DM) tools, hidden interactions among the data could not be exposed with ease to find out the first choice of users. Consequently, a concern that has come up is that data established by DM schemes may totally expose private, sensitive, or confidential information (e.g. home addresses, permanent account number, credit card, and social security numbers data).

To deal with these issues, methods for privacy-preserving data mining (PPDM) [6, 15, 18, 24] were introduced. They comprise database perturbation to sanitize it. PPDM approaches convert the data to maintain privacy. PPDM [12, 25, 30] offers better privacy throughout the mining stage, and it moreover desires to regard the privacy problems of data post-processing and pre-processing. It deals with the issues met by a person or organization when sensitive data are misused or lost. Therefore, the data require to be updated, and thus other persons will not get any suggestion of the private information. Simultaneously the effectiveness of the data has to be conserved [7, 26].

Several techniques have been implemented to hide sensitive patterns, which emerge in binary databases like frequent item sets and association rules [8, 19, 22]. Generally, these techniques delete item sets or transactions from the original database to decrease the confidence or support of the sensitive patterns throughout the process of sanitization. Numerous secure protocols have been presented so far for machine learning and data-mining schemes [14, 29] for clustering, decision tree classification, neural networks, association rule mining (ARM), and Bayesian networks [2, 4]. The major concern of these schemes is to preserve the sensitive data [20, 23] of parties, while they achieve valuable knowledge from the entire dataset. A major concern in DM

***Corresponding author: G.K. Shailaja,** Department of IT, Kakatiya Institute of Technology and Science-Warangal, Telangana 506015, India, e-mail: gkshailaja68@gmail.com
C.V. Guru Rao: Department of CSE, S.R. Engineering College, Warangal, Telangana 506371, India

[5, 16] is the process of finding out frequent item sets, and subsequently association rules [21] are frequently exploited in numerous areas. The majority of the PPDM [9, 11] schemes exploit a transformation that minimizes the convenience of the underlying data when it is applied to data-mining algorithms. Anyhow, these schemes are not capable of adjusting the accuracy and privacy efficiently: "when one is preserved, the other appears to suffer" [10, 28].

This paper contributes secure PPDM, comprising two phases, namely, data sanitization and data restoration, which are the processes after extracting association rules from the original database. In both phases, key extraction plays a major role, which is selected optimally using Opposition Intensity-based Cuckoo Search Algorithm (OI-CSA). Here, four research challenges such as hiding failure (HF) rate, information preservation (IP) rate, false rule generation (FR), and DM are minimized using the proposed sanitization and restoration processes. In addition, the proposed OI-CSA model is compared with the traditional algorithms such as Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Differential Evolution (DE), and Cuckoo Search Algorithm (CSA), and the results are obtained. The paper is organized as follows. Section 2 portrays the related works and reviews done under this topic. Section 3 portrays the suggested objectives for data sanitization and restoration: proposed model, Section 4 explains the key encoding: an improved optimization algorithm, Section 5 demonstrates the results, and Section 6 concludes the paper.

2 Literature Review

2.1 Related Works

In 2018, Afzali and Mohammadi [1] have suggested a data anonymization scheme for fitting big DM. The distinctive characteristics of big data make it essential to consider every rule as an ARM with an appropriate degree. In addition, the proposed methods were compared with the conventional methods, and the quickness of the DM process was offered.

In 2016, Li et al. [13] have focused on privacy-preserving mining on vertically partitioned databases. To ensure privacy they have designed an efficient homomorphic encryption scheme and a secure comparison scheme. They have also proposed a cloud-aided frequent itemset mining solution, which was used to build an association rule mining solution. These solutions are designed for outsourced databases that allow multiple data owners to efficiently share their data securely without compromising on data privacy. These solutions leak less information about the raw data than most existing solutions. Finally, the proposed scheme was compared with other conventional schemes, and the resource utilization at the data owner end was found to be very low.

In 2016, Lin et al. [15] introduced two novel techniques based on the Maximum Sensitive Utility (MSU), which were established to reduce the effects of the sanitization procedure for hiding sensitive high-utility itemsets (SHUIs). Accordingly, the introduced schemes were modeled to remove SHUIs proficiently or lessen their utilities by means of the conceptions of minimum and maximum utility.

In 2018, Chamikara et al. [6] presented an effectual data stream perturbation technique, known as P2RoCAl, that provides improved data utility when distinguished with its contenders. Moreover, the classification accurateness of P2RoCAl data streams was found adjacent to those of the data streams, which were original. Finally, from simulation results, the P2RoCAl was found to offer better resilience in contrast to data reconstruction attacks.

In 2016, Tripathi [24] adopted an antidiscrimination approach like discrimination discovery and prevention in the DM. There were primarily two kinds of adopted models, namely, direct and indirect discrimination. The former existed in the conditions when decisions were attained depending on the sensitive attributes, while the latter exists in the circumstances when decisions were attained depending on the non-sensitive parameters.

In 2016, Upadhyay et al. [25] implemented a geometric data perturbation (GDP) technique by means of data partitioning and 3D rotations. Accordingly, attributes were alienated into three groups, and every group of attributes revolved around varied pairs of axes. Finally, the investigational assessment shows that the

implemented technique delivers worthy privacy preservation outcomes and data utility when distinguished from the traditional techniques.

In 2016, Komishani et al. [12] adopted a scheme known as preserving privacy in trajectory data (PPTD), a new methodology for data preservation depending on the perception of privacy. In addition, it intends to strike stability among the contradictory objectives of data privacy and data utility in agreement with the privacy necessities. From the simulation results, PPTD was found to be proficient for preserving personalized privacy.

In 2015, Yun and Kim [30] proposed a fast perturbation process depending on a tree structure that more rapidly carries out database perturbation procedures for avoiding sensitive information from being revealed. In addition, wide-ranging experimental results were performed for the offered method and traditional schemes by means of both real and synthetic datasets, and the results have offered improved and faster runtime when compared with other conventional schemes.

In 2017, Mehta and Rao [18] proposed techniques such as k-anonymity, l-diversity, and t-closeness, which were utilized to de-identify the data; however, the chances of reidentification are always possible because data were collected from multiple sources. Moreover, it is tough to handle large data for anonymization; MapReduce technique was introduced to handle large volume of data. It distributes large data into smaller chunks across the multiple nodes. Therefore, scalability of privacy-preserving techniques becomes a challenging area of research. The authors explored this area, and they introduced an algorithm named scalable k-anonymization using MapReduce for privacy-preserving big data publishing. Finally, they compared their technique with existing techniques in terms of running time that results into a remarkable improvement.

2.2 Review

Table 1 shows the methods, features, and challenges of conventional techniques based on the privacy data preservation techniques. At first, Fuzzy logic was adopted in [1], which minimizes unwanted side effects and offers better implementation on a huge amount of data. However, there was no contemplation on information loss. In addition, ARM was suggested in [13] that provides reduced complexity, and it also achieves high level of security; anyhow, there was less consideration on the utility of the proposed scheme. In [15], MSU-MAU and MSU-MIU were proposed, which speed up the evaluations, and along with that, they offer bias

Table 1: Review on State-of-the-Art Privacy Data Preservation Techniques.

Author [citation]	Adopted methodology	Features	Challenges
Afzali and Mohammadi [1]	Fuzzy logic	– Minimizes unwanted side effects – Better implementation on a huge amount of data	– No contemplation on information loss
Li et al. [13]	ARM	– Reduced complexity – Achieves high level of security	– No consideration on utility of the proposed scheme
Lin et al. [15]	MSU-MAU and MSU-MIU	– Speed up the evaluations – Offers optimal bias	– Require effective consideration on flexible fitness function
Chamikara et al. [6]	kNN	– Better classification of privacy-preserving data – Improved accuracy	– No contemplation on effectiveness of sampling methods
Tripathi [24]	ARM	– Reduced complexity – Achieves high level of security	– No description of carrying out clustering
Upadhyay et al. [25]	GDP	– Enhanced privacy preservation – High variance	– Decrease in variance increases the chance of attacks
Komishani et al. [12]	PPTD	– Eliminates critical moving points – Minimizes the attacks	– No consideration of PPTD with various sensitive attributes
Yun and Kim [30]	FPA	– Evades the privacy breaches – Better runtime and scalability	– No contemplation on combination with web mining

among the generated side effects and preservation. However, they require effective consideration on flexible fitness function. Similarly, *k*-nearest neighbor (kNN) was adopted in [6], which presents better classification of privacy-preserving data and improved accuracy, but there was no contemplation on effectiveness of sampling methods. In [24], ARM was suggested, which provides reduced complexity, and moreover, it achieves a high level of security. However, there was no description of carrying out clustering. In addition, GDP was adopted in [25], which offered enhanced privacy along with high variance, but the decrease in variance increases the chance of attacks. Accordingly, PPTD was proposed in [12], which eliminated critical moving points together with the minimization of attacks. Anyhow, there was no consideration of PPTD with various sensitive attributes. Finally, fast perturbation algorithm (FPA) was presented in [30], which evades the privacy breaches, and it also provides better runtime and scalability. However, there was no consideration on combination with web mining. There, these limitations have to be considered for improving the PPDM techniques effectively in the current research work.

3 Suggested Objectives for Data Sanitization and Restoration: Proposed Model

3.1 Objective Function

The proposed model OI-CSA intends to attain the objective function for preserving data as given by Eq. (1).

$$\min F = \max(F_1, F_2, F_3, F_4) \tag{1}$$

In Eq. (1), F_1, F_2, F_3, and F_4 are the objectives, which demonstrates the importance of the corresponding function as defined in the section below.

$$F_1 = \frac{f_1}{\max(f_1) \forall \text{ iterations}} \tag{2}$$

$$F_2 = \frac{f_2}{\max(f_2) \forall \text{ iterations}} \tag{3}$$

$$F_3 = \frac{f_3}{\max(f_3) \forall \text{ iterations}} \tag{4}$$

$$F_4 = \frac{f_4}{\max(f_4) \forall \text{ iterations}} \tag{5}$$

In Eq. (2), F_1 denotes the normalized HF rate, f_1 denotes the HF rate, where $\max(f_1)$ is considered as the worst f_1 of all iterations. In Eq. (3), F_2 denotes the normalized IP rate, and f_2 denotes the IP. In Eq. (4), F_3 denotes the normalized IP rate; f_3 denotes the FR rate. In Eq. (5), F_4 denotes the normalized DM rate; f_4 denotes the DM. Here the original database is considered as O, and the sanitized database is indicated as O'.

HF rate denoted by f_1 is defined as the fraction of sensitive rules which is depicted in O' as revealed by Eq. (6). Here, the count of sensitive rules available in O' is described as $f_1 = |B' \cap SRs|$. In Eq. (6), B signifies the association rule produced prior to sanitization, B' denotes the association rules obtained from O', and SRs denotes the sensitive rules.

$$f_1 = \frac{|B' \cap SRs|}{|SRs|} \tag{6}$$

IP rate denoted by f_2 is described as "the rate of non-sensitive rules which are concealed in O'". It is the reciprocal of information loss as revealed in Eq. (7).

$$f_2 = 1 - \frac{|B - B'|}{|B|} \tag{7}$$

FR denoted by f_3 is described as "the rate of artificial rules produced in O'" which is demonstrated by Eq. (8).

$$f_3 = \frac{|B - B'|}{|B'|} \tag{8}$$

DM denoted by f_4 is portrayed as the count of modifications carried out in O' from O as demonstrated by Eq. (9), in which *dist* points out the Euclidean distance found between O and O'.

$$f_4 = dist(O, O') \tag{9}$$

3.2 Proposed Architecture

The overall framework of the adopted OI-CSA scheme is portrayed by Figure 1. Initially, the data preservation comprises two major processes, such as data sanitization and data restoration. In data sanitization process, a key is generated to preserve the sensitive data in a protective approach. The key has to be generated such that it must hide the sensitive data effectively for which OI-CSA is deployed for optimal key generation. The authorized person at the receiver side could then restore the sanitized data by exploiting the same key. This is because to sanitize and restore the data faster, a symmetric key is used by the sender as well as the receiver.

3.3 Sanitization Process

Binarization of O and pruned key matrix A_2 are performed during sanitization. The resultant key matrix in binarized form is consecutively provided onto the rule hiding process, in which an XOR function is performed with binarized form of O with identical matrix dimensions and added up with one that generates the O', as specified by Eq. (10). Moreover, O' obtained from sanitization process attains *SRs* and association rules following sanitization B'. Similarly, O extracts the relative association rules prior to sanitization B for attaining the above mentioned objectives. The structural design of the proposed sanitization process is illustrated by Figure 2.

$$O' = (A_2 \oplus O) + 1 \tag{10}$$

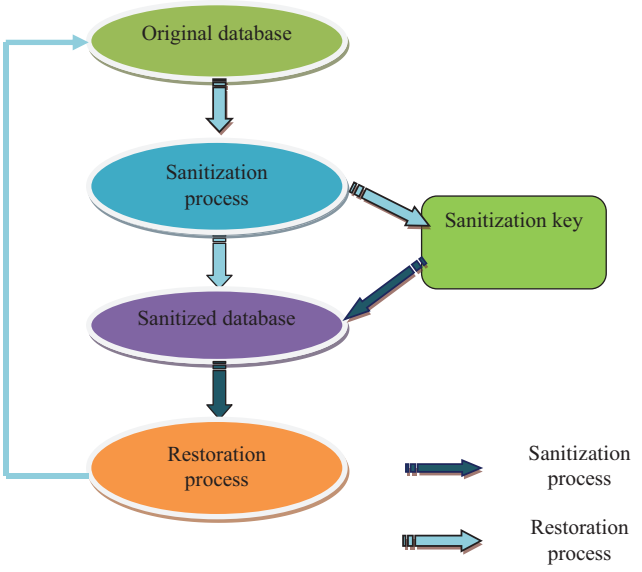

Figure 1: Overall Framework of the Proposed Data Preservation Model.

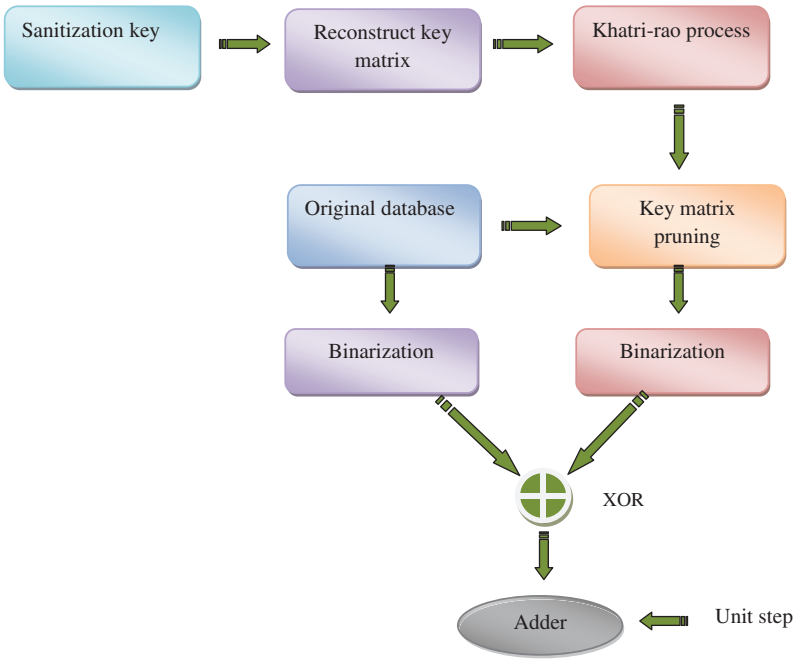

Figure 2: Sanitization Process of the Implemented Data Preservation Algorithm.

3.4 Key Generation

The key generation includes solution transformation process, where A, a key representation, is converted with the aid of the Khatri-rao product. Primarily, A is restructured into A_1 with matrix dimensions $\left[\sqrt{M''_O} \times O_{\max}\right]$ in which $O_{\max}$ indicates the highest transaction length and M_O denotes the number of transactions; the nearest highest perfect square of M_O is denoted by M''_O. For example, the restructured process of $A = \{1, 2, 1\}$ performs row-wise duplication and constructs the reconstructed key matrix, A_1 with dimension $\left[\sqrt{M''_O} \times O_{\max}\right]$ as revealed by Eq. (11).

$$A_1 = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}_{\left[\sqrt{M''_O} \times O_{\max}\right]} \tag{11}$$

Thus, the key matrix A_2 with dimensions $\left[\sqrt{M_O} \times O_{\max}\right]$ is attained by the Khatri-rao product of two identically restructured A_1 matrixes represented as $A_1 \otimes A_1$ in which the kronecker product is indicated by $\otimes$ and its dimensions are further reduced in terms of the dimension size of the original database. Based on the Khatri-rao product, the key generation process is carried out and generates a matrix with dimensions identical to O, which produces $A_{2\left[\sqrt{M_O} \times O_{\max}\right]}$. Finally, the process of rule hiding is done to obtain O' by hiding the sensitive rules. The optimal key generation is made by means of improved Cuckoo Search (CS) algorithm called OI-CSA.

3.5 Restoration Process

During the restoration process, the O' achieved from sanitization and A_2 from key generation technique could be binarized. The binarized S_d from the binarization block is minimized from the unit step. In the meantime, the database and key matrix, which is binarized, takes on an XOR function following the subtraction, and subsequently the restored database is extracted. The sanitizing key, A_2, is reconstructed by exploiting Eqs. (11), (10), and (1) and the proposed OI-CSA update. It is deployed to generate O' by which the lossless restoring

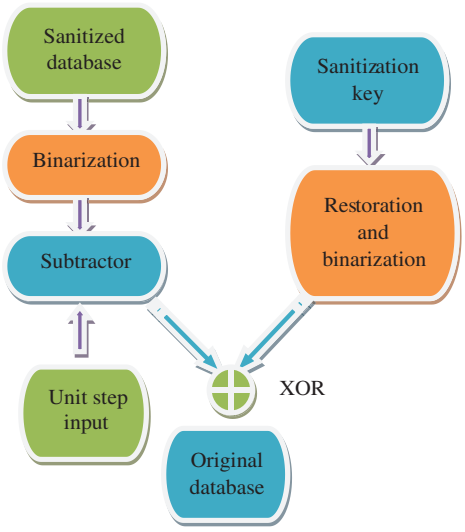

Figure 3: Restoration Process of the Proposed Data Preservation Algorithm.

could be carried out by Eq. (12), in which $\hat{O}$ indicates the restored data. The design of restoration process is given by Figure 3.

$$\hat{O} = (O' - 1) \oplus A_2 \tag{12}$$

4 Key Encoding: an Improved Optimization Algorithm

4.1 Key Encoding

The keys (chromosome) A used for the sanitization process are subjected to OI-CSA for encoding. The number of keys ranging from key A^1 to key A^M is optimized using OI-CSA, and the optimal key is identified. The solution-encoding process is illustrated by Figure 4. Here, the key length (chromosome) is assigned as $\sqrt{M''_O}$.

4.2 Cuckoo Search Algorithm

CS algorithm is a method that is established depending on the reproduction of cuckoos [17]. Usually, cuckoos lay their eggs in the nests of erstwhile cuckoos with the expectation of their babies grown up by alternative parents. Certain periods exists, when the cuckoos find out that the eggs in their nests do not possess to them, in such circumstances, the unfamiliar eggs are push out from the nests and are deserted. This approach is dependent on the subsequent three conditions:

1. Every cuckoo chooses a nest arbitrarily and lays an egg in it.
2. The desired nests with more eggs would be considered for subsequent generation.
3. For a predetermined quantity of nests, a host cuckoo can find out another egg with a probability $P\varepsilon[0, 1]$. Under such conditions, the host cuckoo can either push the egg or construct a nest in another place.

The final condition can be estimated by substituting a fraction of the n host nests with novel ones. The fitness or quality F_i of a solution can merely be equivalent to the value of the objective function. From the execution

Figure 4: Keys for Encoding.

point, the demonstration which is followed is that every egg in a nest indicates a solution, and every cuckoo can lay only one egg, i.e. one solution. Moreover, no reverence can be performed among an egg, a cuckoo, or a nest. The objective is to exploit the novel and capable cuckoo egg (i.e. solution) to substitute a worst solution in the nest. CS approach is very effectual for global optimization issues as it sustains a balance among global random walk and the local random walk. The balance among global and local random walks is adjusted by a switching constraint $P\varepsilon[0, 1]$. The global and local random walks are demonstrated by Eqs. (13) and (14) correspondingly. Accordingly, in Eq. (13), X_i^t and X_k^t denotes the present positions chosen by arbitrary permutation, β indicates the positive step size scaling factor, X_i^{t+1} points out the subsequent position, s denotes the step size, $\otimes$ indicates the element-wise product of two vectors, F signifies the heavy side function, P symbolizes a variable that is exploited to switch among global and random walks, and ε denotes an arbitrary variable from a uniform distribution. Accordingly, from Eq. (14), $N(s, \tau)$ indicates levy distribution exploited to describe the step size of an arbitrary walk.

$$X_i^{t+1} = X_i^t + \beta\, s \otimes F(P - \varepsilon) \otimes \left(X_i^t - X_k^t\right) \tag{13}$$

$$X_i^{t+1} = X_i^t + \beta\, N(s, \tau) \tag{14}$$

4.3 OI-CSA

The conventional CS algorithm is an uncomplicated and efficient global optimization algorithm; anyhow, it could not be exploited directly to resolve multimodal optimization issues. Hence, the traditional CS approach is improved by modifying the opposition intensity denoted by γ as given in Eq. (15). In Eq. (14), $X_i^{(w)}$ indicates the worst solution, X_i^t denotes the current solution, and γ varies from 0 to 1.

$$X_i^{t+1} = X_i^t + \beta\, N(s, \tau) - \gamma\left[X_i^{(w)} - X_i^t\right] \tag{15}$$

The fundamental phases of the OI-CSA algorithm depending on their conditions is given by Algorithm 1.

Algorithm 1: Pseudo code of OI-CSA.

Objective function, $f(X) = X = (X_1, X_2 \ldots X_D)^T$
Generate initial population of n host nests $Xi(i = 1, 2 \ldots . n)$
While ($t <$ *Max Generation*) or terminate the process
Determine a cuckoo (assume i) arbitrarily by Lévy distribution;
Evaluate its quality/fitness; F_i
Choose a nest among n (assume j) arbitrarily;
Evaluate its quality/fitness; F_j
If ($F_i > F_j$)
Update the solution using Eq. (13);
Else
Determine γ
Update the solution using Eq. (15)
End if
Novel nests are built at novel locations.
Maintain the best solutions;
Order the solutions and determine the best;
End while
Post processing

5 Results and Discussions

5.1 Simulation Procedure

The proposed OI-CSA method for privacy preservation of sensitive data was simulated in JAVA, and the results were obtained. The analysis was carried out using four datasets, namely, T10, Chess, Retail, and T40. Moreover, the results were compared with conventional models such as PSO [31], GA [27], DE [32], and CSA [3] algorithms. In addition, the results that were obtained by varying the γ value were also described for the four adopted datasets.

5.2 Performance Analysis

The performance analysis of the proposed OI-CSA model for four datasets is given by Figure 5. From Figure 5A, it can be noted that the presented model for chess dataset in terms of F_1 is 17.36%, 14%, 12.24%, and 6.99% better than the PSO, GA, DE, and CSA designs. Also, for F, the suggested scheme is 0.76%, 0.49%, 0.28%, and 0.23% superior to the PSO, GA, DE, and CSA algorithms. Also, from Figure 5B, a retail dataset was implemented, where the implemented scheme for F_2 is 1.89%, 3.02%, 2.64%, and 2.64% better than the PSO, GA, DE, and CSA algorithms. Also, for F, the proposed OI-CSA model is 1.89%, 3.02%, 2.64%, and 2.64% superior to the PSO, GA, DE, and CSA algorithms. Similarly, from Figure 5C, the performance analysis for the T40

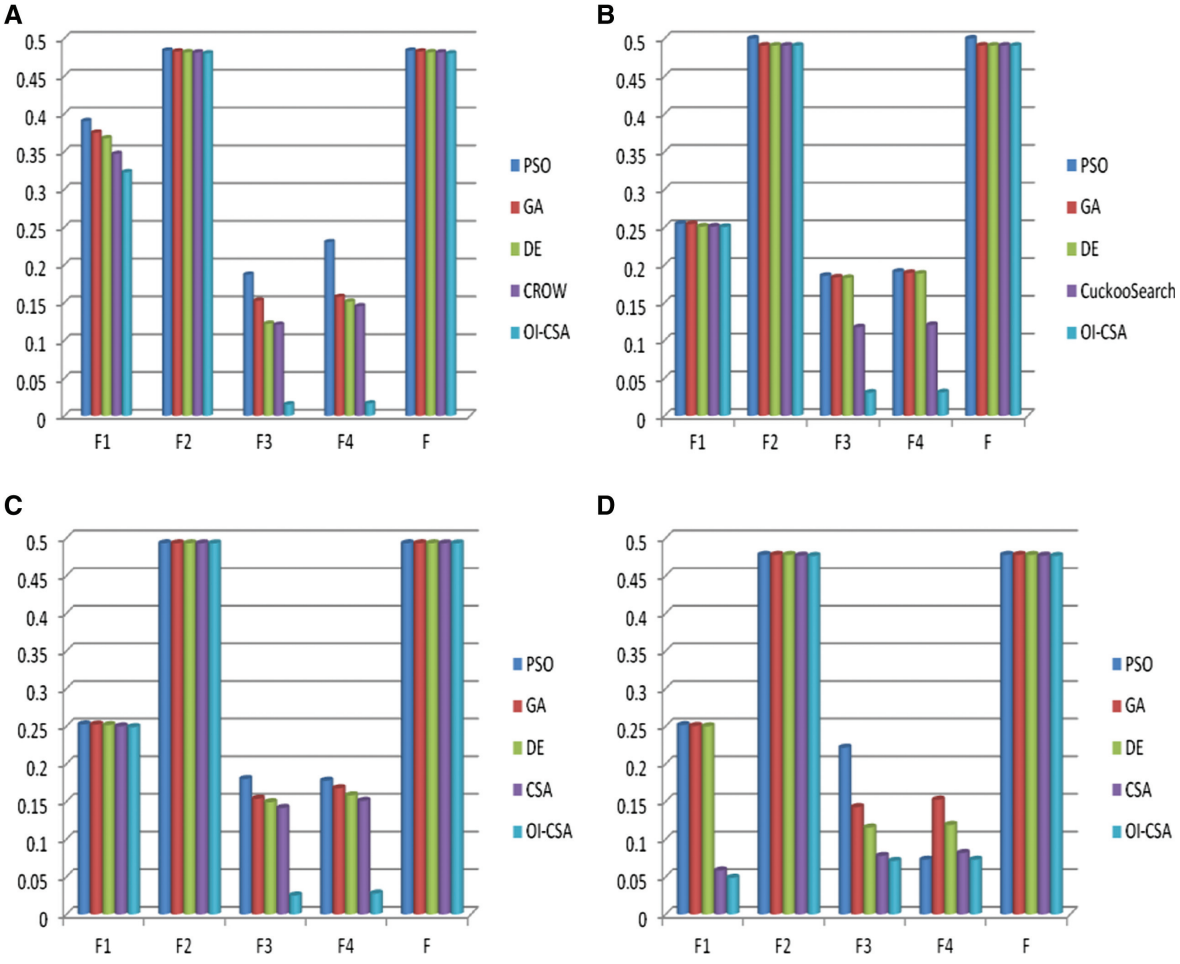

Figure 5: Performance Analysis of the Proposed and Conventional Approaches for Data Preservation Algorithm Using Satabase (A) Chess, (B) Retail, (C) T40 and (D) T10.

dataset can be obtained, where the suggested design for F_3 is 86.08%, 83.7%, 83.2%, and 82.29% better than the PSO, GA, DE, and CSA methods. For F, the presented method is 2.36%, 1.47%, 0.89%, and 0.29% superior to the PSO, GA, DE, and CSA algorithms. Moreover, from Figure 5D, the T10 dataset can be attained, where the F_4 analysis for OI-CSA model is 52.38%, 38.98%, and 10.94% better than the GA, DE, and CSA approaches. Also, from Figure 5D, the performance analysis for the T10 dataset for F can be attained, where the implemented scheme is 0.32%, 0.31%, 0.29%, and 0.1% superior to the PSO, GA, DE, and CSA schemes. Therefore, the enhancement of the proposed OI-CSA model has been substantiated successfully.

5.3 Effect on Varying γ

The effect of varying γ using four datasets, namely, chess, retail, T40, and T10, is demonstrated by Tables 2–5, respectively. The values of γ were varied from 0.0, 0.2, 0.5, 0.7, to 1. Accordingly, from Table 2, the normalized HF rate, F_1 has attained the values of 0.24875, 0.01890625, 0.24575, 0.24775, and 0.020227273 for $\gamma = 0.0$, $\gamma = 0.2$, $\gamma = 0.5$, $\gamma = 0.7$, and $\gamma = 1$, respectively, using the chess dataset. From Table 3, the normalized rate, F_2, has attained the values of 0.490418, 0.487253, 0.490492, 0.490418, and 0.487179 for $\gamma = 0.0$, $\gamma = 0.2$, $\gamma = 0.5$, $\gamma = 0.7$, and $\gamma = 1$, respectively, using the retail dataset. In addition, from Table 5, the normalized IP rate, F_3, has attained the values of 0.025027958, 0.019158457, 0.04072148, 0.015477353, and 0.025347075 for

Table 2: Effect of Varying γ Using Chess Dataset.

Methods	$\gamma = 0.0$	$\gamma = 0.2$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 1$
F_1	0.24875	0.01890625	0.24575	0.24775	0.020227273
F_2	0.492828071	0.487003676	0.492828287	0.492857266	0.490466263
F_3	0.025027958	0.019158457	0.04072148	0.015477353	0.025347075
F_4	0.027649747	0.020470537	0.04542156	0.016355949	0.026856822
F	0.492828071	0.487003676	0.492828287	0.492857266	0.490466263

Table 3: Effect of Varying γ Using Retail Dataset.

Methods	$\gamma = 0.0$	$\gamma = 0.2$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 1$
F_1	0.25025	0.014156	0.25075	0.25025	0.013402
F_2	0.490418	0.487253	0.490492	0.490418	0.487179
F_3	0.030634	0.02366	0.052324	0.101194	0.035404
F_4	0.031084	0.024052	0.053499	0.103217	0.036264
F	0.490418	0.487253	0.490492	0.490418	0.487179

Table 4: Effect of Varying γ Using T40 Dataset.

Methods	$\gamma = 0.0$	$\gamma = 0.2$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 1$
F_1	0.24875	0.01890625	0.24575	0.24775	0.020227273
F_2	0.492828071	0.487003676	0.492828287	0.492857266	0.490466263
F_3	0.025027958	0.019158457	0.04072148	0.015477353	0.025347075
F_4	0.027649747	0.020470537	0.04542156	0.016355949	0.026856822
F	0.492828071	0.487003676	0.492828287	0.492857266	0.490466263

Table 5: Effect of Varying γ Using the T10 Dataset.

Methods	$\gamma = 0.0$	$\gamma = 0.2$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 1$
F_1	0.048591	0.24925	0.032639	0.05025	0.24825
F_2	0.47599	0.477248	0.473262	0.459936	0.477248
F_3	0.070969	0.028772	0.047274	0.023253	0.087213
F_4	0.072494	0.030037	0.052792	0.025071	0.09723
F	0.47599	0.477248	0.473262	0.459936	0.477248

$\gamma = 0.0$, $\gamma = 0.2$, $\gamma = 0.5$, $\gamma = 0.7$, and $\gamma = 1$, respectively, using the T40 dataset. Also, from Table 5, using the T10 dataset, the normalized FR rate, F_4, has acquired the values of 0.072494, 0.030037, 0.052792, 0.025071, and 0.09723 for $\gamma = 0.0$, $\gamma = 0.2$, $\gamma = 0.5$, $\gamma = 0.7$, and $\gamma = 1$, respectively. Therefore, the superiority of the presented OI-CSA approach has been validated effectively.

6 Conclusion

The paper has presented a novel PPDM method, which comprises two phases like data sanitization and data restoration, which were started after the association rules generation. Accordingly, in both the sanitization and restoration processes, the key extraction has a major role that was chosen optimally by means of OI-CSA. Here, four research objectives, namely, HF rate, IP, FR, and DM were reduced using the adopted sanitization and restoration processes. In addition, the proposed scheme was compared with conventional approaches such as the GA, PSO, DE, and CS, and the optimal results were attained for the proposed scheme. From the performance analysis, it can be noted that the proposed model for the chess dataset on considering the overall cost function was 0.76%, 0.49%, 0.28%, and 0.23% superior to the PSO, GA, DE, and CSA algorithms. On considering the performance analysis for the retail dataset, it can be noted that the proposed OI-CSA model was 1.89%, 3.02%, 2.64%, and 2.64% superior to the PSO, GA, DE, and CSA algorithms. Thus, the enhancement of the adopted OI-CSA technique has been confirmed in an effective manner. In future, in order to secure keys, a key management technique is also considered. Key management is a significant step in protecting the generated key and transferring it to the authenticated receiver. Since this paper considers a general platform of data privacy protection, the key management has not been considered. However, appropriate key management protocols/mechanisms need to be considered based on the applications, communication link, communication protocols, and sender/receiver characteristics.

Bibliography

[1] G. A. Afzali and S. Mohammadi, Privacy preserving big data mining: association rule hiding using fuzzy logic approach, *IET Inform. Secur*. **12** (2018), 15–24.
[2] A. Anjum, T. Ahmed, A. Khan, N. Ahmad and N. Farooq, Privacy preserving data by conceptualizing smart cities using MIDR-Angelization, *Sustain. Cities Soc*. **40** (2018), 326–334.
[3] A. Askarzadeh, A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm, *Comput. Struct*. **169** (2016), 1–12.
[4] S. Bennati and E. Pournaras, Privacy-enhancing aggregation of Internet of Things data via sensors grouping, *Sustain. Cities Soc*. **39** (2018), 387–400.
[5] H. K. Bhuyan and N. K. Kamila, Privacy preserving sub-feature selection in distributed data mining, *Appl. Soft Comput*. **36** (2015), 552–569.
[6] M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe and I. Khalil, Efficient data perturbation for privacy preserving and accurate data stream mining, *Pervasive Mob. Comput*. **48** (2018), 1–19.
[7] Y. Dong and D. Pi, Novel privacy-preserving algorithm based on frequent path for trajectory data publishing, *Knowl.-Based Syst*. **148** (2018), 55–65.
[8] M. A. Ferrag, L. A. Maglaras, H. Janicke, J. Jiang and L. Shu, A systematic review of data protection and privacy preservation schemes for smart grid communications, *Sustain. Cities Soc*. **38** (2018), 806–835.
[9] F.-J. González-Serrano, Á. Navia-Vázquez and A. Amor-Martín, Training support vector machines with privacy-protected data, *Pattern Recogn*. **72** (2017), 93–107.
[10] R. Jiang, R. Lu and K.-K. R. Choo, Achieving high performance and privacy-preserving query over encrypted multidimensional big metering data, *Future Gener. Comput. Syst*. **78** (2018), 392–401.
[11] Y. Kokkinos and K. G. Margaritis, Confidence ratio affinity propagation in ensemble selection of neural network classifiers for distributed privacy-preserving data mining, *Neurocomputing* **150** (2015), 513–528.
[12] E. G. Komishani, M. Abadi and F. Deldar, PPTD: preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression, *Knowl.-Based Syst.* **94** (2016), 43–59.
[13] L. Li, R. Lu, K. K. R. Choo, A. Datta and J. Shao, Privacy-preserving-outsourced association rule mining on vertically partitioned databases, *IEEE Trans. Inf. Foren. Sec*. **11** (2016), 1847–1861.
[14] Y. Li, J. Yang and W. Ji, Local learning-based feature weighting with privacy preservation, *Neurocomputing* **174** (2016), 1107–1115.

[15] J. C.-W. Lin, T.-Y. Wu, P. Fournier-Viger, G. Lin and M. Voznak, Fast algorithms for hiding sensitive high-utility item sets in privacy-preserving utility mining, *Eng. Appl. Artif. Intel*. **55** (2016), 269–284.
[16] L. Liu, M. Kantarcioglu and B. Thuraisingham, The applicability of the perturbation based privacy preserving data mining for real-world data, *Data Knowl. Eng*. **65** (2008), 5–21.
[17] M. Mareli and B. Twala, An adaptive Cuckoo search algorithm for optimisation, *Appl. Comput. Inform*. **14** (2017), 107–115.
[18] B. B. Mehta and U. P. Rao, Privacy preserving big data publishing: a scalable *k*-anonymization approach using MapReduce, *IET Softw*. **11** (2017), 271–276.
[19] C. Modi, U. P. Rao and D. R. Patel, A survey on preserving privacy for sensitive association rules in databases, in: *International Conference on Business Administration and Information Processing*, vol. 70, pp. 538–544, Berlin, Heidelberg, 2010. https://doi.org/10.1007/978-3-642-12214-9_96.
[20] M. Prakash and G. Singaravel, An approach for prevention of privacy breach and information leakage in sensitive data mining, *Comput. Electr. Eng*. **45** (2015), 134–140.
[21] T. Pranav Bhat, C. Karthik and K. Chandrasekaran, A privacy preserved data mining approach based on *k*-partite graph theory, *Procedia Comput. Sci*. **54** (2015), 422–430.
[22] X. Qi and M. Zong, An overview of privacy preserving data mining, *Procedia Environ. Sci*. **12** (2012), 1341–1347.
[23] P. Sui and X. Li, A privacy-preserving approach for multimodal transaction data integrated analysis, *Neurocomputing* **253** (2017), 56–64.
[24] K. K. Tripathi, Discrimination prevention with classification and privacy preservation in data mining, *Procedia Comput. Sci*. **79** (2016), 244–253.
[25] S. Upadhyay, C. Sharma, P. Sharma, P. Bharadwaj and K. R. Seeja, Privacy preserving data mining with 3-D rotation transformation, *J. King Saud Univ. Comput. Inform. Sci*. **30** (2016), 524–530.
[26] S. Vennila and J. Priyadarshini, Scalable privacy preservation in big data a survey, *Procedia Comput. Sci*. **50** (2015), 369–373.
[27] T. D. Vrionis, X. I. Koutiva and N. A. Vovos, A genetic algorithm-based low voltage ride-through control strategy for grid connected doubly fed induction wind generators, *IEEE Trans. Power Syst*. **29** (2014), 1325–1334.
[28] A. Waqar, A. Raza, H. Abbas and M. K. Khan, A framework for preservation of cloud users' data privacy using dynamic reconstruction of metadata, *J. Netw. Comput. Appl*. **36** (2013), 235–248.
[29] R. Wei, H. Tian and H. Shen, Improving *k*-anonymity based privacy preservation for collaborative filtering, *Comput. Electr. Eng*. **67** (2018), 509–519.
[30] U. Yun and J. Kim, A fast perturbation algorithm using tree structure for privacy preserving utility mining, *Expert Syst. Appl*. **42** (2015), 1149–1165.
[31] J. Zhang and P. Xia, An improved PSO algorithm for parameter identification of nonlinear dynamic hysteretic models, *J. Sound Vib*. **389** (2017), 153–167.
[32] L. M. Zheng, S. X. Zhang, K. S. Tang and S. Y. Zheng, Differential evolution powered by collective information, *Inform. Sci*. **399** (2017), 13–29.