Resbee Publishers

Multimedia Research (MR)

Received 20 April, Revised 28 June, Accepted 23 July

Hybrid Weed-Particle Swarm Optimization Algorithm and C-Mixture for Data Publishing

Yogesh R kulkarni
Vel Tech University
Chennai, Tamil Nadu, India
yogeshrk2015@gmail.com

Senthil Murugan T
Dept of Computer Science and Engineering
Vel Tech University
Chennai, Tamil Nadu, India

Abstract: From the experts and researchers, data publishing is the center of attention in the latest technology, which receives great interest. The idea of data publishing faces a large number of security problems chiefly, while any trusted organization presents data to the third party, personal information requires not to be revealed. Hence, to keep the data privacy, this work presents a method for privacy preserved collaborative data publishing by exploiting the Weed and Particle Swarm Optimization algorithm (W-PSO) for that a C-mixture parameter is utilized. The parameter of C-mixture improves data privacy if the data does not assure privacy constraints, like l-diversity, m-privacy and k-anonymity. The least fitness value is controlled which is based upon the least value of the widespread information loss and the least value of the average equivalence class size. The minimum value of the fitness assures the utmost utility and the least privacy. Simulation is performed by exploiting the adult dataset and the proposed method is superior to the conventional algorithms regarding the widespread information loss and the average equivalence class metric and attained minimum values.

Keywords: Data Publishing; Data Privacy; C-Mixture; Fitness Value; Privacy Constraints

1. Introduction

Nowadays a volatile increment of crowd-sourced data is increasing with the speedy advancement of mobile devices and networking, (for instance real-time traffic at Waze, check-in data at Foursquare, commonsense and air quality at Sensor Map) from numerous users. In real-time, these crowd-sourced data can be comprehensive and mined using machine learning applications to recognize precious information and the additional advantage of people's life, for instance, popular restaurant recommendation, real-time traffic analysis, and navigation [2]. For data mining reasons lately progressively agencies (for example companies and governments) are publishing the crowd-sourced data to the public. On the other hand, the capable benefits of data mining and publishing are at the hazard of revealing responsive information to data miners. An up to date study [1] exhibited that with a few exterior information, the human mobility data attained from users can be connected back to an individual. With the rising apprehension of private data publishing, privacy leakage methods are immediately necessary to keep the sensitive information of individuals [4].

By several administrative systems, data publishing sets the phase for the data users to perform widespread researches with different determinations. For instance, for analysis, banks publish their data consequently hence economists examine the data and make decisions in view of that. For pharmaceutical researchers and world health organizations, hospitals publish their data. In the present epoch, data publishing is compulsory for researchers and analysts. For making decisions it is a requirement and additional growth in several areas. The published data is considered confidential information and sensitive regarding the individuals (i.e. data owners) besides the quasi-identifiers information and personally identifiable. Moreover, the data publishing in its innovative structure is an open threat [1] and [3] to an individual's privacy such as suppression and a generalization, incognito.

For data analysis and publishing, the epidemic of privacy connected events has encouraged an extended line of study in privacy notions, like l-diversity, t-closeness, and k-anonymity, to name some [9], [10] and [11]. A table assures k-anonymity if each quasi-identifier attribute in the table is impossible to tell apart from at least k-1 other quasi-identifiers attributes; such a table is termed a k-anonymous table. Whereas k-anonymity secures individuality disclosure of individuals by connecting attacks, it is inadequate to put off attribute disclosure with side information. By integrating the released data with side information, it creates it probable to deduce the probable sensitive attributes equivalent to an individual. For an individual, once the association among the sensitive and identifier attributes is exposed, it may mischief the distribution and the individual of the complete table. In [12], to pact with this problem, l-diversity was developed. L-diversity needs that the sensitive attributes enclose as a minimum well-indicated value in every correspondence class. As explained in [13], l-diversity has 2 main issues. One is that it restricts the adversarial knowledge, whereas it is probable to obtain knowledge of a sensitive attribute from usually presented worldwide distribution of the attribute. One more issue is that all attributes are implicit to be definite that assumes the adversary either obtains all the information or obtains nothing for a sensitive attribute.

In spite of the flourishing privacy-protection of using differential privacy to data publishing, the majorities of conventional systems rely on a trusted server to combined the crowdsourced data and agitate the true aggregated statistics prior to their publishing [1]. The conventional architecture of statistics publishing and data collection shows in a trusted server, whereas users upload data openly to the trusted server that subsequently achieves statistic publishing and computation [18]. On the other hand, a server will almost certainly be hacked and become untrusted. In this scenario, the untrusted server must not be permitted to store or receive the raw data directly from users, or else the sensitive and identity information of individuals will be revealed. Additionally, the untrusted server must not be permitted to aggregate on the crowdsourced data; if not the true aggregated statistics will be revealed that will additionally reveal sensitive information of individuals [2]. Regrettably, a conventional differentially private data publishing model relying on the trusted server cannot protect the privacy of individuals anymore while the server is untrusted [3].

The main contribution of this paper is to propose the W-PSO algorithm, which is the hybridization of the weed approach and the PSO approach which presents the fittest report with a maximum degree of privacy protection. The proposed algorithm presents in secure data publishing when preserving privacy.

2. Literature Review

In 2018, M.H. Afifi et al. [1], developed a new multi-variable privacy quantification and characterization method. On the basis of this method, it had the ability to examine the posterior and prior adversarial beliefs regarding attribute values of individuals. In privacy characterization, they had examined the sensitivity of any identifier. Subsequently, they had exhibited which privacy must not be estimated on the basis of one metric. Moreover, they had presented 2 diverse metrics for quantification of distribution leakage, privacy leakage, and entropy leakage. They had examined a few of the majority renowned Privacy-Preserving Data Publishing (PPDP) methods like –diversity, k-anonymity, l and t-closeness by exploiting aforesaid metrics. On the basis of the presented metrics and framework, and the conventional PPDP strategies were determined and it had some restrictions in privacy characterization. Moreover, the presented privacy characterization and measurement model gives the superior understanding and assessment of these models. Hence, it presents an establishment for the plan and study of PPDP strategies.

In 2018 Chao Yan et al [2], worked on the Item-based collaborative filtering (ICF) method, which had extensively employed to create service recommendations in the big data environment. On the other hand, the ICF model only carried out well while the data for service recommendation decision-making were saved in a physically centralized way. However, in the distributed environment, it frequently not succeeds to suggest suitable services to a target user wherever the involved multiple parties were unenthusiastic to discharge their data to each other owing to privacy apprehensions. Taking into consideration of this disadvantage, they had enhanced the conventional ICF model by combining the locality-sensitive Hashing (LSH) method, to understand reliable and secure data publishing. Additionally, by combining the published data with small privacy across different platforms, suitable services were suggested on the basis of the recommended recommendation method called ICFLSH.

In 2016, Dingqi Yang et al [3], presented PrivRank, which was a continuous and customizable privacy-preserving social media data publishing structure protecting users over inference attacks whereas enabling personalized ranking-based recommendations. Its main proposal was to incessantly confuse user action data so that the privacy leakage of user-specified private data was reduced in a given data distortion budget that bounds the ranking loss incurred from the data obfuscation procedure to preserve the efficacy of the data for enabling recommendations. An experiential assessment on both real-world and synthetic datasets exhibits that the proposed framework can competently present efficient and continuous protection.

In 2018, SabaYaseen et al [4], worked on data publishing, privacy, and utility, which were necessary for data users and owners correspondingly that cannot exist at the same time. This incompatibility places the data privacy researchers in a requirement to discover newer and consistent privacy-preserving tradeoff models. Data providers such as numerous private and public organizations (for instance banks and hospitals) publish microdata of individuals for several research reasons. Publishing microdata might cooperation with the privacy of individuals. Moreover, they had presented three new generalization approaches such as Divisors Based Generalization Hierarchies (DBGH), Conventional Generalization Hierarchies (CGH), and Cardinality-Based Generalization Hierarchies (CBGH).

In 2018, Zhibo Wang et al [5], explained the issue of real-time crowd-sourced statistical data publishing with sturdy privacy protection in an untrusted server. Moreover, they presented a new distributed agent-based privacy-preserving model, named DADP which set up a new level of multiple agents among the untrusted server and the users. Rather than directly uploading the check-in information to the untrusted server, a user can arbitrarily choose one agent and upload the check-in information to it with the anonymous association technology. Every agent aggregates the received crowd-sourced data and perturbs the aggregated statistics locally with the Laplace model. From all the agents the perturbed statistics were additionally integrated together to form the complete perturbed statistics for publication.

In 2017, Brijesh B. Mehta and Udai Pratap Rao [6], worked on big data, which was processed and collected by exploiting different tools and sources that lead to privacy problems. Privacy-preserving data publishing models like k- and l-diversity, anonymity, and t-closeness were exploited to de-identify the data. On the other hand, the probability of re-identification was forever residual present as data was gathered from multiple sources. Because of the outsized volume of data, less generalization or suppression was needed to attain a similar level of privacy that was as well called as large crowd effect, though it was forever challenging to handle such large data for anonymization. MapReduce handles a large volume of data and distributes the data into the lesser chunks across the multiple nodes; as a result, the complete benefit of a large volume of data was attained. Hence, the scalability of privacy-preserving models becomes a challenging area of research. Here, a method called Scalable K-Anonymization (SKA) was proposed by exploiting MapReduce for privacy-preserving big data publishing.

In 2019, TAO WU et al [7], developed an active learning model which chooses the majority representative relations to be perturbed, therefore regulating the structural predictability of graphs, i.e., removing as small as probable relations to challenge the regularity level of graphs, that forms the foundation of inference attack models. Particularly, with the supposition that the substructure with superior regularity level encloses more regular equivalence components and has more equivalent paths supplied for the random walk processes, random walk-based relation significance measuring method was presented to recognize the representative relations.

In 2017, LU Qiwei et al [8], worked on a privacy-preserving trajectory data publishing, and the majority of them suppose the attacks with similar adversarial backdrop knowledge. In fact, different users had different privacy requirements. Such a non-personalized privacy supposition does not meet the personalized privacy requirements; for now, it loses the possibility to attain improved efficacy by the enchanting benefit of differences of users' privacy requirements. Moreover, they had examined the personalized trajectory k-anonymity principle for trajectory data publication. In particular, they had investigated and recommended a complete model that presents privacy-preserving services.

3. Description of Secure Data Publishing

Data publishing has numerous benefits in several fields mostly, in the medical field, the data with respect to the diseases are proposed to a gathering of research board, and they have the duty of deciding the availability, nature, effects, and the impacts of the disease. With the occurrence of publishing the studies to the researchers, the publisher must not reveal any of the personal information of the individuals with the research board. That is to say, the privacy of the individual information is protected. To assure the entire privacy of the individual reports, this work exploits three privacy constraints and the C-mixture which improves the privacy measures. Additionally, the W-PSO method is proposed which determines the optimal solution for publishing the privacy assured data. In this paper, the privacy parameters are proposed and explained as follows:

3.1 Privacy Preserved Data Publishing Model

The main objective of the privacy preserved data publishing is to create the data smaller specific in order that the privacy of the individual is preserved and concurrently, the functionality of the data is ensured. For providing the data, there is a set of service providers to the trusted third party and these service providers present a set of data every containing of the quasi-identifiers, service providers name, sensitive and non-sensitive attributes. Eq. (1) is used for the set of data or reports developed by the service providers.

$$E = \{E_p \in S_p ; 1 \leq p \leq N\} \tag{1}$$

In eq. (1), N indicates the total count of providers gives out the reports and p indicates the number of providers included in proposing the reports to the third party. E_p indicates the reports proposed using the p^{th} service provider and S_p indicates the p^{th} service provider. The eq. (1) represents the individual reports published by the service provider. Eq. (2) is used for the data published by the service provider contains the quasi-identifier, attributes.

$$E = \{S, C_p, C_2, C_0, qi_p, qi_2, qi_Q, SA_p, SA_2, SA_n\} \tag{2}$$

In eq. (2), E indicates the data report of the service provider, qi is the quasi-identifier, SA indicates the sensitive attribute, S indicates the service provider name, C indicates the common attribute. The trusted third party publishes the report proposed by the service provider via the adaptations done to E and the report to be published is indicated as E^*. The report E^* is altered by considering all the attacks which are done to recognize the individual information. The published information is preserved therefore any attack was done on the published data never reveal the individual information or present any identity of the private data. The security to the data from revealing the originality of the report with the individual information is preserved by the subsequent process called l-diversity, k-anonymity, and m-privacy. The procedure of security is improved exploiting the generalization procedure that creates the data minimum specific. Fig. 1 indicates the basic model of the privacy preserved data publishing scheme. From the individuals, the data are preserved so that those records need privacy. The data from any government or any organization -related is shared with any trusted third party when assuring entire security to the data. The word privacy describes that the third party does not have any clue in relating the individual data.

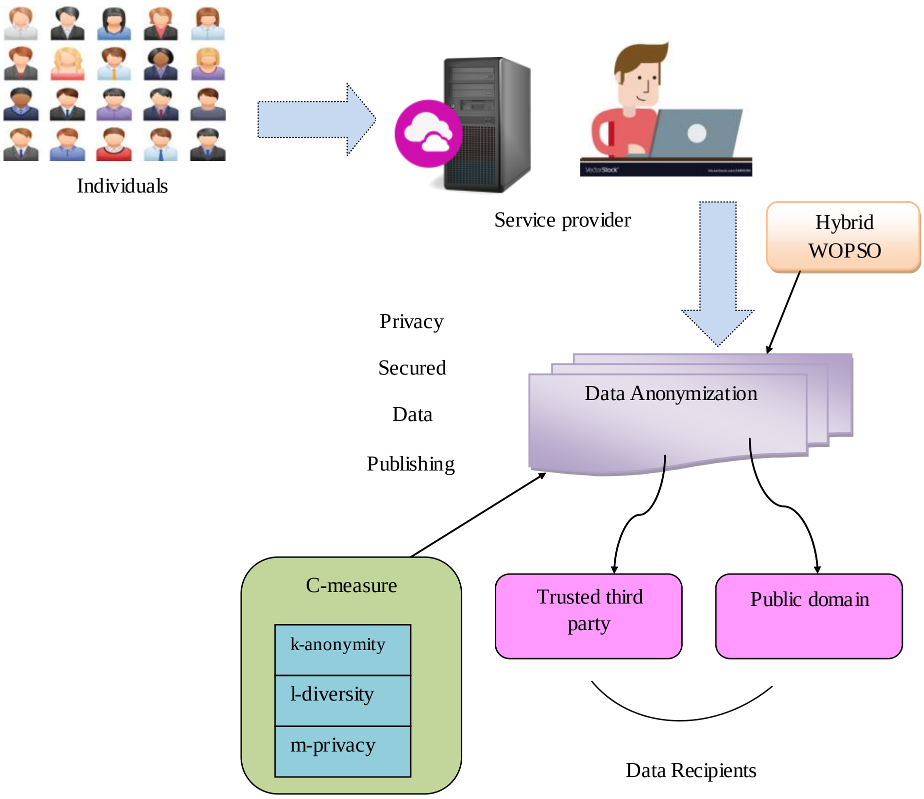

Fig. 1. *Schematic illustration of data publishing*

3.2 Improving Data Privacy Using the C-mixture

For a report E comprising the quasi-identifiers SA , the k-anonymity indicates that the report E fulfills k-anonymity regarding the identifier SA if and only if, for each series in $E[SA]$, there is a minimum j count of incidents in $E[SA]$. The k-anonymity experiences generalization to make $(j-1)$ records that are less-informative and indistinguishable. The procedure of creating a report inexplicit protects the report from connecting an individual. Nevertheless, k-anonymity solitary cannot assure entire security to the data as it is presented backdrop attacks and agonizes from homogeneity. The k-anonymity solitary cannot present the privacy to the data hence, the other privacy measures such as m-privacy and l-diversity are additionally exploited in this work. For assuring entire privacy, this work exploits the anonymization model by exploiting the three processes, such as l-diversity, k-anonymity, and m-privacy. To assure the privacy of the individual data rigidly the mixture of the 3 processes together is done. The privacy of the data is preserved by exploiting a parameter C called the C-mixture developed in the subsequent section.

C-mixture refers to the measure of the privacy of the published data which alters the 3 privacy measures. Any data which is published must fulfill the constraints to assure data privacy. Consider a report E comprising the quasi-identifiers SA, subsequently; the report E must fulfill the C-mixture. The following are the circumstances to fulfill the C-mixture. i) In every group, there must be less of $B\%$ duplicate records. ii) For sensitive attributes of each group, it is obligatory to have $B\%$ well-described values. iii) For each group, there must be $B\%$ service providers. Eq. (3) represents the C-mixture, and the privacy constraint is conceived.

$$[j = T * B]\,[g = k * B]\,[h = S * B] \tag{3}$$

In eq. (3), T indicates the total number of records and the total number of records is eight, B indicates to the index which indicates the C-mixture. k indicates the number of class attributes and the value is 2, S indicates the number of the service providers and the service providers value is 2.

The privacy constraint, such as the l-diversity, m-privacy, and k-anonymity are improved on the basis of the C-mixture. The C value deviates in the value range $0 \le B \le 1$. In [14], the sample input data that consists of the provider name, quasi-identifiers and the attributes. The privacy is sustained by the generalization procedure via the c-mixture table. While the data is published, it is obligatory to secure the privacy of the data and for assuring the security, the C-mixture idea is enabled. The privacy constraints, such as g, j, and h are enhanced on the basis of the C-mixture. The data before publishing is checked to decide the privacy, and only the data that have ideal privacy is published. If the privacy measure of the data is not fulfilled, subsequently the privacy of the data is enhanced before publishing the data. For example, let us explain the application of the privacy measures in assuring privacy to the data which is published. While the C -the mixture is located at 0.56, the privacy measures, j, g, and h values attained are 4, 2, and 2 correspondingly.

4. Hybrid Weed- PSO Model for Assuring Privacy in Data Publishing

The proposed algorithm for collaborative data publishing by exploiting a security-enabled approach by the W-PSO is presented in this section. The W-PSO is a hybridization method consisting of the weed approach and the PSO method, which has numerous benefits while comparing with the conventional Weed Optimization and the conventional PSO. The importance of the hybridization method is that it has a superior convergence rate and it evades converging to the local optimum. The global optimum is attained by exploiting the PSO method. The execution cost and convergence time are less while comparing with the other optimization issues.

4.1 Fitness Calculation

The main objective of the fitness model is to determine the optimal data which have a maximum degree of privacy for the reason of data publishing. The fitness model is on the basis of the generalized information loss (GenILoss) and the average equivalence size of the class (C_{avg}). Once the privacy constraints are understandable, the fitness is analyzed to low the usefulness and increase privacy. The usefulness of the data and the fitness metric (GenILoss) are inversely proportional, therefore, to maintain an increased value of the usefulness, the value of (GenILoss) is maintained less and (C_{avg}) must endure less to maintain the data privacy. Hence, eq. (4) represents fitness constraints.

$$F(P) = \lambda * \text{GenILoss}(P) + \chi * C_{avg}(P) \tag{4}$$

The usefulness and the data privacy is sustained by the fewer values of the (GenILoss) and the (C_{avg}). The eq. (4) is subjected to three conditions as stated below:

$$\left.\begin{matrix} j \ge U_{ano}(C, E) \\ g \ge U_{ano}(C, E) \\ h \ge U_{pri}(C, E) \end{matrix}\right\} \tag{5}$$

where,

$$\text{GenILoss}(P) = \frac{1}{N \times Q} \times \sum_{h=1}^{Q} \sum_{n=1}^{N} \frac{V_{hn} - L_{hn}}{V_h - L_h} \tag{6}$$

$$C_{avg}(P) = \frac{N}{\left|SAq_{iQ}\right| * j} \tag{7}$$

In eq. (6), L_h and V_h indicates the lower and the upper bounds of the h^{th} quasi-identifier, SA indicates the sensitive identifier, $U_{ano}(C,E)$ indicates the model which decides the duplicate records. In the generalized interval, the lower and the upper bounds present is denoted as L_{hn} and V_{hn} correspondingly. N states the total count of records and Q states the total count of quasi-identifiers available. By utilizing the function $U_{pri}(C,E)$ the number of service providers is calculated and $U_{div}(C,E)$ decides the number of the defined sensitive attributes.

4.2 Conventional Weed Optimization Algorithm (WOA)

In conventional WOA methods, every weed as a constituent of a population otherwise colony of possible solutions in that the weed locations comprise the decision variables of an optimization issue [15]. Moreover, the weeds are permitted to reproduce on the basis of their quality (that is on their objective model values) in the colony. It states that improved the weed quality, the higher the number of seeds it generates. A seed is an enhanced solution occurs from a conventional weed. While the method initiates the weeds are in an inappropriate environment and try to deal out their seeds in a bulky space to look for a more appropriate environment. This step of the WOA searches for the best space close to the best solution or point. From this time onwards the weeds deal out their seeds in a close-up range that brings the newly created weeds randomly near to the best position or global optimal solution of the optimization issue being resolved.

The conventional WOA steps are as follows:

a) A first p -dimensional population $Q_{initial}$ of weeds is produced and distributed arbitrarily in space.

b) Reproduction: In this step, the present weeds create seeds taking into consideration those weeds with the optimal and worst qualities. S_{max} and S_{min} state correspondingly the number of weeds with optimal and worst qualities. S_{max} and S_{min} are user-chosen. Eq. (8) is used for seed production.

$$S_i = \frac{F_i - F_{min}}{F_{max} - F_{min}} \times (S_{max} - S_{min}) + S_{min} \tag{8}$$

In eq. (8) S_i indicates the number of seeds created by weed i, F_i = the value of the objective model of weed i, In the colony of weeds F_{min} = minimum value of the objective model. In the colony of weeds F_{max} = maximum value of the objective model. Eq. (8) designates seed chooses from the fittest weeds, which, consecutively, more probable than not will acquiesce a better population of weeds, and rapidly and so forth until attainment a convergence reason. This selection process of the conventional WOA imitates the evolutionary model of the endurance of the fittest.

c) Adaptation, seed distribution, and randomness: The seeds are distributed arbitrarily with a 0-mean normal distribution. According to eq. (9), their population's S.D is minimized from a first (high) defined value to a last (low) defined value in every generation.

$$\delta_{itr} = \frac{(itr_{max} - itr)^n}{itr_{max}}(\delta_{initial} - \delta_{final}) + \delta_{final} \tag{9}$$

In eq. (9), δ_{itr} = the value of the Standard Deviation (SD) in the current iteration of the WOA, itr_{max} = a maximum number of iterations (that is the production of seeds), itr = iteration number, $\delta_{initial}$ = initial SD, δ_{final} = last SD and $n = b$ non-linear module (non-linear modulation index) that is chosen by the user.

d) Competitive exclusion: The exclusion of evolutionarily unwanted weeds begins subsequent to a small number of iterations when the number of weeds in the colony goes beyond its utmost probable number (Q_{max}).The exclusion is effected by steps sketched over to maintain the number of weeds in the population in restrictions. This process is iterated until the end of the algorithm.

4.3 Conventional Particle Swarm Optimization Algorithm (PSO)

In [16], the PSO approach was developed enthused by the social behavior of birds and fish that live in groups. The PSO approach is used to particles, therefore it is called Particle Search Optimization. Each particle's value presented to the objective model is computed based on its location in the decision space. Subsequently, any particle chooses a direction to move along on the basis of an amalgamation of its current location, the optimal location it has ever in use, and on the location of other particles that currently engage the optimal locations in the population of particles. One step of the PSO approach is finished once all the particles in the current population have stimulated. Until the method attains the maximum iteration number steps are repeated.

The PSO method starts with an arbitrarily produced population of particles. The preliminary stage of every particle is the optimal location the particle has ever engaged called P_{best}, and the optimal particle with the optimal objective model value is decided called G_{best} in the first iteration. In the subsequent iterations P_{best} and G_{best} are updated on the basis of the locations engaged by the particles in the population refereed using the values of their objective models. For each particle, a new velocity is computed using eq. (11) considering P_{best} and G_{best}. According to eq. (11) the new location for each particle is subsequently computed on the basis of the new velocity and current location of the particle.

$$U_i(t+1) = \tau\, U_i(t) + C_1 r_1 \left[P_{best_i}(t) - Y_i(t)\right] + C_2 r_2 \left[G_{best}(t) - Y_i(t)\right] \tag{10}$$

$$Y_i(t+1) = Y_i(t) + U_i(t+1) \tag{11}$$

In eq. (10), $U_i(t) =$ the velocity of a particle i in iteration t, $U_i(t+1) =$ velocity of a particle i in iteration $t+1$, $P_{best_i}(t)$ the optimal location that particle i has engaged until iteration t, $G_{best_i}(t)$ location of the optimal particle until iteration t, $Y_i(t+1) =$ location of a particle i in iteration $t+1$, $Y_i(t) =$ the current location of a particle i in the iteration t. Additionally, $\tau =$ indicates the inertia coefficient that is an index of convergence to local and global optima. C_1 and C_2 indicate the global learning and personal learning factors, correspondingly, and are user-defined. The parameters r_1 and r_2 indicate the numbers that are produced arbitrarily with a uniform distribution in the interval [0, 1] in each iteration.

4.4 Proposed Hybrid W-PSO Algorithm

In [17], the hybrid W-PSO algorithm is proposed by integrating the PSO and WOA method. The conventional WOA converges to the best area of solutions comparatively gradually. Nevertheless, subsequent to discovering the best area, the WOA algorithms have the global optimum precisely due to its efficacious and its diverse search ability. The PSO method that changes the particles' location within the search space is on the basis of the social-psychological of individuals to imitate the achievement of other individuals. In each iteration, the PSO method converges to the best area of solutions rapid with quite intentional particle movements. The proposed W-PSO shows a fast and precise convergence to the near-global optimal solution. The interface among the dispersion algorithm of the WOA method and the speed of the PSO method creates an effectual balance among global and local exploration of the issue space. Moreover, [17] showed the hybrid W-PSO method outcomes showing faster convergence to best optima than the PSO and WOA method could attain individually by examining them with mathematical benchmark models.

The hybrid W-PSO method steps are stated as below:

A preliminary population of weeds is arbitrarily produced and distributed in a p-dimensional space. For each weed, the objective model value is computed. The number of seeds that is new weeds or potential solutions produced by each weed is computed using Eq. (8).

The locations of weeds are considered regarding P_{best} and G_{best} as performed with the PSO method as stated. New solutions (seeds) are produced using eq. (10) and (11) used in the PSO method. The values of the objective model are computed for every generated weed. Their locations are changed with the P SO method. This is going after by seed generation with the WOA. Then, the PSO method updates the locations P_{best} and G_{best}. By the WOA, the reasonable exclusion is done. In a colony, while the number of weeds attains the permitted number the exclusion procedure is unnatural and the weeds which have lower objective model values are evaded from the colony.

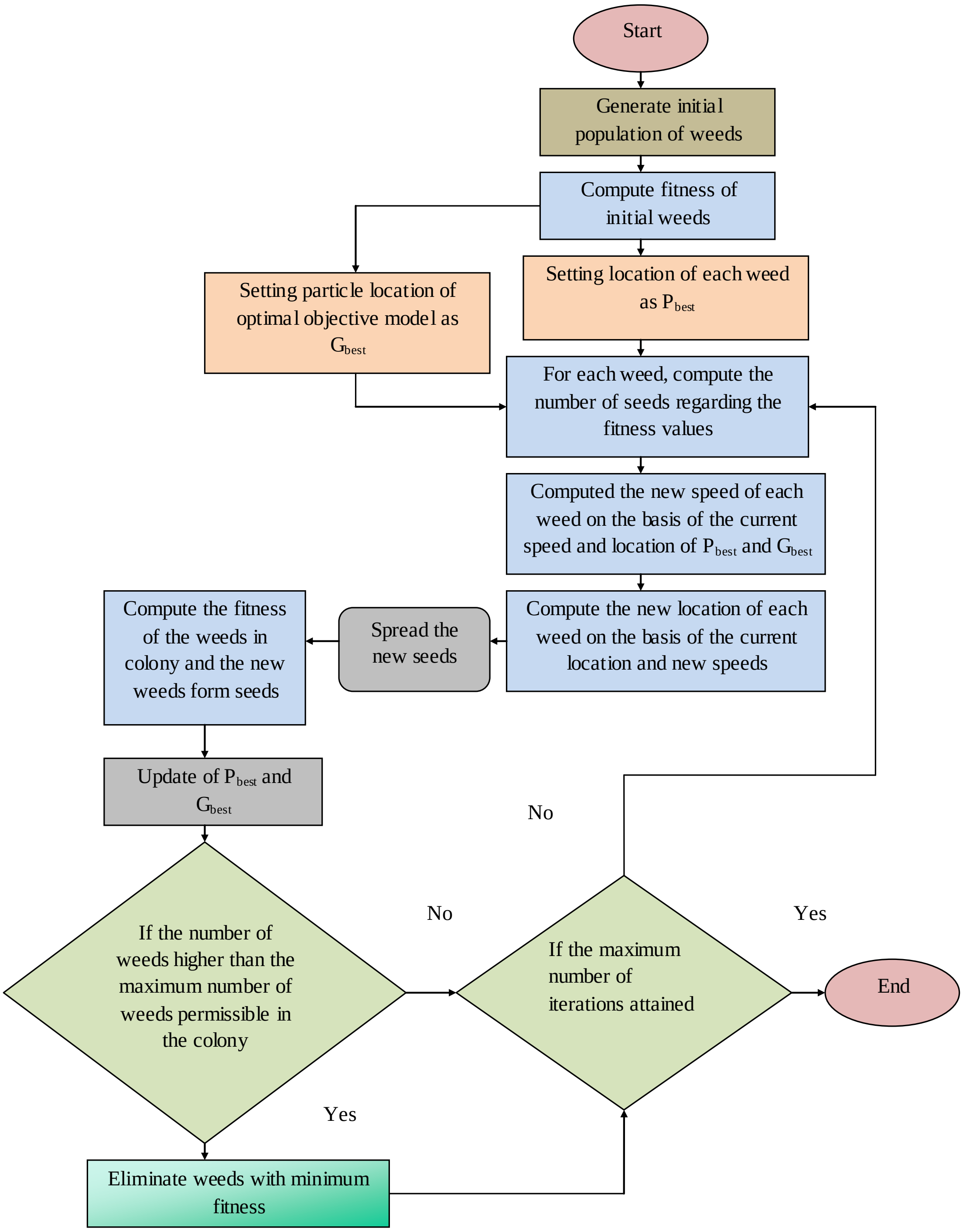
Start
Generate initial population of weeds
Compute fitness of initial weeds
Setting particle location of optimal objective model as G_{best}
Setting location of each weed as P_{best}
For each weed, compute the number of seeds regarding the fitness values
Computed the new speed of each weed on the basis of the current speed and location of P_{best} and G_{best}
Compute the fitness of the weeds in colony and the new weeds form seeds
Spread the new seeds
Compute the new location of each weed on the basis of the current location and new speeds
Update of P_{best} and G_{best}
No
If the number of weeds higher than the maximum number of weeds permissible in the colony
No
If the maximum number of iterations attained
Yes
End
Yes
Eliminate weeds with minimum fitness

Fig. 2. *Flowchart of the proposed hybrid W-PSO algorithm*

5. Results and Discussions

5.1 Experimental Procedure

In this section, the evaluation of the proposed W-PSO method was analyzed and analyzed with the conventional models regarding the performance metrics. The performance metrics obtained for the evaluation of the performance of the proposed model. Here, the metrics used for analysis such as Generalized Information loss, Average equivalence class size metric. In this paper, the dataset exploited is the adult data set, 1996 or Census Income dataset [14] that is generated from the census bureau database. It comprises of the 48,842 instances and 14 attributes constituting both categorical and integer attributes namely, age, work class, fnlwgt, education, education-num, marital status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per week, native-country. The performance of the proposed model is performed regarding the performance metrics to show the efficiency of the proposed model.

5.2 Performance Evaluation

The performance analysis of the proposed W-PSO approach with conventional algorithms such as Genetic Algorithm (GA), Grey Wolf Optimization (GWO) and Particle Swarm Optimization (PSO) is presented in this section. In Fig 3, the performance evaluation of the proposed algorithm regarding the GenILoss is presented. Here, the C-mixture value is varied from 0.2, 0.25, 0.3 and 0.35 correspondingly. From Fig 3, it is obvious that the proposed method value minimizes with the maximizing C values however for the conventional approach, the value of GenILoss is maximized.

Fig. 4 exhibits the performance analysis of the proposed model regarding the algorithms such as conventional algorithms. Here, it is obvious the value minimizes with the raise in the C-mixture value and the proposed model value is less while comparing with the conventional algorithms.

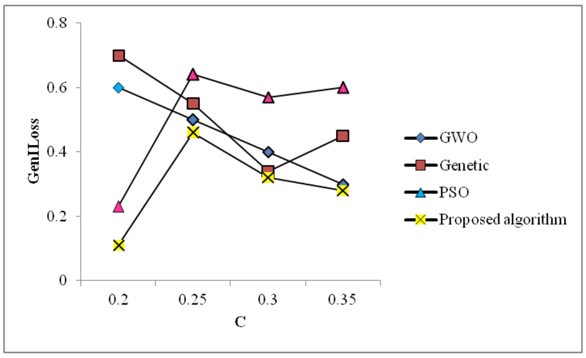

Fig. 3. *Performance analysis of the proposed approach concerning GenIloss*

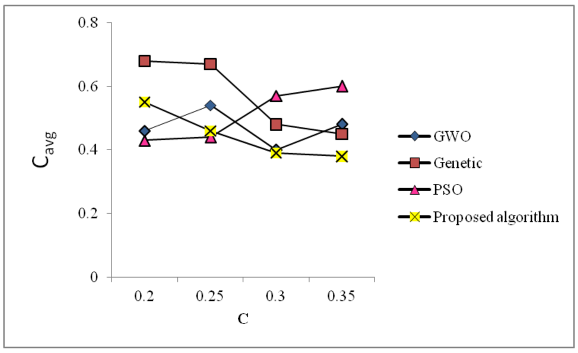

Fig. 4. *Performance analysis of the proposed approach concerning the Cavg*

6. Conclusion

In this paper, the proposed W-PSO algorithm improves the privacy of the data and therefore, and the privacy preservation was done via three privacy constraints, like m-privacy, l-diversity, and the k-anonymity. These privacy constraints based upon the C-mixture value. At first, the privacy constraints were examined for all the records to examine if the privacy constraints are fulfilled. While it was realized that the privacy measures of the data were not fulfilled, the value of C-mixture was modified. The modification if the value of C-mixture was performed by exploiting the meta-heuristic optimization approach, W-PSO algorithm, which decides the optimal record that preserves the privacy. The record with enhanced privacy was chosen for publishing hence that the data chosen does not present any probability for connecting the individual information. The performance was evaluated utilizing the adult dataset that was regarding the possibility and GenILoss, which enables the minimum value of fitness.

References

[1] M. H. Afifi, K. Zhou and J. Ren, "Privacy Characterization and Quantification in Data Publishing," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 9, pp. 1756-1769, 1 Sept. 2018.

[2] C. Yan, X. Cui, L. Qi, X. Xu and X. Zhang, "Privacy-Aware Data Publishing and Integration for Collaborative Service Recommendation," IEEE Access, vol. 6, pp. 43021-43028, 2018.

[3] D. Yang, B. Qu and P. Cudré-Mauroux, "Privacy-Preserving Social Media Data Publishing for Personalized Ranking-Based Recommendation," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 3, pp. 507-520, 1 March 2019.

[4] S. Yaseen et al., "Improved Generalization for Secure Data Publishing," IEEE Access, vol. 6, pp. 27156-27165, 2018.

[5] Z. Wang et al., "Privacy-Preserving Crowd-Sourced Statistical Data Publishing with An Untrusted Server," IEEE Transactions on Mobile Computing, vol. 18, no. 6, pp. 1356-1367, 1 June 2019.

[6] B. B. Mehta and U. P. Rao, "Privacy preserving big data publishing: a scalable k-anonymization approach using MapReduce," IET Software, vol. 11, no. 5, pp. 271-276, 10 2017.

[7] T. Wu, G. Ming, X. Xian, W. Wang, S. Qiao and G. Xu, "Structural Predictability Optimization Against Inference Attacks in Data Publishing," IEEE Access, vol. 7, pp. 92119-92136, 2019.

[8] Q. Lu, C. Wang, Y. Xiong, H. Xia, W. Huang and X. Gong, "Personalized Privacy-Preserving Trajectory Data Publishing," Chinese Journal of Electronics, vol. 26, no. 2, pp. 285-291, 3 2017.

[9] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[10] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in Security & Privacy, pp. 111–125, 2008.

[11] M. G¨ otz, S. Nath, and J. Gehrke, “Maskit: Privately releasing user context streams for personalized mobile applications,” in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD ’12, (New York, NY, USA), pp. 289– 300, ACM, 2012.

[12] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “`-diversity: Privacy beyond k-anonymity,” ACM Trans. Knowl. Discov. Data, vol. 1, Mar. 2007.

[13] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in ICDE, pp. 106–115, 2007.

[14] Adult Data Set, 1996. from <https://archive.ics.uci.edu/ml/datasets/Adult>.

[15] F. Mohamadi Monavar, N. Komjani and P. Mousavi, "Application of Invasive Weed Optimization to Design a Broadband Patch Antenna With Symmetric Radiation Pattern," in IEEE Antennas and Wireless Propagation Letters, vol. 10, pp. 1369-1372, 2011.

[16] Y. Song, Z. Chen and Z. Yuan, "New Chaotic PSO-Based Neural Network Predictive Control for Nonlinear Process," in IEEE Transactions on Neural Networks, vol. 18, no. 2, pp. 595-601, March 2007.

[17] Hajimirsadeghi, H., Lucas, C., 2009. A hybrid IWO/PSO algorithm for fast and global optimization. In: Institute of Electrical and Electronics Engineers Eurocon, SaintPetersburg, Russian, May18-23, pp. 1964–1971.

[18] V. Vinolin,"Breast Cancer Detection by Optimal Classification using GWO Algorithm"Multimedia Research, Volume 2, Issue2,April2019.