International Journal of Operations Research and Information Systems
Volume 12 • Issue 4

A Technique for Securing Big Data Using K-Anonymization With a Hybrid Optimization Algorithm

Suman Madan, JIMS, Delhi, India
Puneet Goswami, SRM University, Haryana, India

ABSTRACT

The recent techniques built on cloud computing for data processing are scalable and secure, which increasingly attracts the infrastructure to support big data applications. This paper proposes an effective anonymization-based privacy-preservation model using k-anonymization criteria and grey wolf-cat swarm optimization (GWCSO) for attaining privacy preservation in big data. The anonymization technique is processed by adapting k-anonymization criteria for duplicating k records from the original database. The proposed GWCSO is developed by integrating grey wolf optimizer (GWO) and cat swarm optimization (CSO) for constructing the k-anonymized database, which reveals only the essential details to the end users by hiding the confidential information. The experimental results of the proposed technique are compared with various existing techniques based on the performance metrics, such as classification accuracy (CA) and information loss (IL). The experimental results show that the proposed technique attains an improved CA value of 0.005 and IL value of 0.798, respectively.

KEYWORDS

Anonymization, Big Data, Cat Swarm Optimization, Cloud Computing, Grey Wolf Optimization, Privacy Preservation

1. INTRODUCTION

The advancements in big data led to several opportunities for research in the upcoming years. The Big data is adapted for discovering knowledge using different sectors of society. The big data contains vast data, which is generated through the digital processes and shared among several individuals through webs. The big data has spanned the way for making the decisions in a right way. The decision support has motivated several users to keep the data online (Xuezhen, et al., 2014). Due to the sharing of data, several concerns related to security are generated. The ability to store the personal information is a major issue in the context of privacy-preservation (Karle & Vora, 2017). As the big data handles the data of a large number of users, the privacy is an important task, which needs to be accomplished for protecting the data (Yang, et al., 2014), (Youke, et al., 2020). Numerous applications are designed for allowing the users to access the data with trust management (Denglong et al., 2020). The privacy and security is a major challenge in big data. The big data is not accepted if privacy and security are not addressed. The scalability (S. Atiewi et al., 2020) is another major issue when the conventional preservation technique is adapted in big data. In spite of several techniques developed for privacy preservation, most of them cannot efficiently preserve the privacy as they fail to handle different attacks (Antony & Antony, 2016). The big data requires large storage and computational power for

DOI: 10.4018/IJORIS.20211001.oa3

preserving the data. Hence, it adapts a large distributed system for storing data at various locations and for easy retrieval (Geetha, et al., 2017).

As preserving the privacy is an important issue in processing the big data, it affects academia as well as the IT industry. The important aspect while sharing data is to preserve the privacy and simultaneously provide the data utility. The purpose of extracting the useful data from large datasets is to obtain a data, which is not misused. Several techniques are devised for privacy preservation, but most of them are ineffective for addressing the problems related to security while privacy preservation (Thanamani, 2017). The priory used privacy preservation techniques can be categorized into two phenomena. The first phenomenon is hiding the identity of the user and the second phenomenon is the preservation of user's important data. The big data needs to consider communication overhead and computational cost due to its large volume, velocity, and variety (Guan & Si, 2017). The information transfer can be secured if the privacy of the database is preserved. The parameters considered for the privacy preservation while processing big data are categorized as integrity, controllability, preservability, and confidentiality. The performance of various algorithms based on privacy preservation is increased due to its outstanding behavior to protect the big data. However, the technique based on privacy preservation ignores the accessing of data by untrustworthy users due to data loss, and data leakage. The privacy can be preserved using input privacy and output privacy. The performance of the anonymization based algorithms can be improved if optimization based algorithms are adapted (Tang, et al., 2016).

Various privacy preservation techniques are developed for preserving the privacy in various stages, such as data generation, data storage, and data processing. Encryption, (Abdulatif, et al., 2020) and Cryptography is one of the techniques used for securing data, but these algorithms are too complex and are not easily understandable (Dijk & Juels, 2010). The data anonymization methods considered for privacy preservation are K-Anonymity (Sweeney, 2002 ; Bayardo & Agrawal, 2005), L-Diversity (Sedayao, 2012 ; Sei, *et al.,* 2016), TCloseness (Li, *et al.,* 2007), Notice and Consent (Cate & Schönberger, 2012). Differential privacy (Salido, 2012) is one of the privacy preservation methods, which are widely accepted by several users. This method enables the analysts to refine the required information from the database having personal information by providing protections individually (Gosain & Chugh, 2014). The data anonymization and multidimensional anonymization methods are important for preserving data among the existing methods based on anonymization. The scalability and cost are the issues, which are not addressed by the existing anonymization methods while handling the big data. The multidimensional anonymization scheme is popular from the existing anonymization techniques due to its low data distortion. The scalability is an issue, which is addressed by the existing techniques by combining spatial indexing or sampling technologies (Zhang, et al., 2013). Numerous techniques adapted for privacy preservation contribute to protecting the confidential information of the users. In (Sreedhar & Umamaheshwari, 2014), a two-phase top-down specialization approach is devised for anonymizing huge data by employing a MapReduce framework to solve the problems of big data in the context of cloud computing (Anshi Singh, and Anjani Rai, 2020). In (Priyanka, *et al.,* 2014), a Top DownScale (TDS) approach is developed for preserving the privacy. However, the technique failed to manage huge datasets, creating several issues. Several optimization methods (Ratre & Pankajakshan, 2017 ; Dhumane & Prasad, 2017 ; Nipanikar, *et al.,* 2017 ; Shelke & Prasad, 2018 ; Krishnamoorthy & Asokan, 2014) are utilized for privacy preservation to optimize the results.

This paper proposes a privacy preservation model using k-anonymization criteria and GWCSO for achieving secure communication in the cloud platform while transmitting the data to the end user. The challenges faced by the existing privacy preservation algorithms are addressed for designing effective anonymization based privacy preservation model using k-anonymization criteria and GWCSO for secure communication. This method adapts k- anonymization constraint for attaining k-duplicate records from the original records. The proposed privacy preservation model uses the GWCSO algorithm for constructing a k-anonymized database that satisfies k-anonymization criteria considering fitness as a minimization function. The aim of the proposed preservation model is to gather

the information from the data owners and store it in the information database that is transformed to the k-anonymized database by applying k-anonymization technique over the database for attaining secure communications with the end users. The original data is converted to secured data with a set of operations to anonymize the data. The data is anonymized in such a way that the confidential information is hidden from the end users without revealing the original data.

The major contributions of the proposed technique for attaining privacy preservation are illustrated as follows:

- Designing a k-anonymized database using k-anonymization criteria and GWCSO for duplicating k records to achieve an effective privacy preservation mechanism.
- Developing a hybrid optimization model, named GWCSO, by integrating GWO with CSO algorithm to construct the k-anonymized database, for initiating a secure communication among the end users.

The paper is organized as follows: Section 1 elaborates the introduction about privacy preservation system and its contribution towards publishing data. Section 2 describes the literature survey based on privacy preservation by analyzing eight research papers and the challenges faced by the existing techniques. Section 3 explains the proposed privacy preservation method for attaining a secured communication. Section 4 explains the results and discussion of the proposed k-anonymization and GWCSO with other existing methods, and finally, section 5 concludes with a summary.

2. MOTIVATION

2.1 Related Works

In this section, various researches developed in the literature for privacy preservation are presented.

Shalin Elizabeth S *et al.* (2015) developed a technique, named Two-Phase Top-Down Specialization (TPTDS), for obtaining effective privacy preservation. The TPTDS technique solves the issues of big data. The original database adapts an anonymization technique for converting the original data to a protected data and is considered as a multidimensional problem. The privacy preserved database is constructed using the taxonomy tree and generalization mechanisms. However, the technique is not suitable for on-click elasticity.

R. Sreedhar *et al.* (Sreedhar & Umamaheshwari, 2014) developed a model for anonymization and named it as Optimal balancing scheduling. The model adopts the MapReduce framework for maximizing the scalability parameter. This model can be used for improving re-anonymization, and the issues related to data locality are solved. Besides, the method failed to address the problems of security while privacy preservation.

Xuyun Zhang *et al.* (2014) developed an anonymization technique and presented it as a local recording problem. The technique utilized several sensitive attributes of the database for preserving the privacy. This technique adapts MapReduce and k-means mechanisms for building the model for privacy preservation.

David Rebollo-Monedero *et al.* (2017) designed a model, called the probabilistic-based anonymity model for addressing the privacy-related issues. This technique adapts a different assumption of trusts for removing the distortions from the database. Moreover, this model improves the utility of the database by maximizing the privacy.

Chi Lin *et al.* (2016) developed a scheme, namely differential privacy protection scheme, for sensitive big data. For reducing the errors, a tree structure is built for providing long-range queries. The histogram is transformed to a complete binary tree by Haar Wavelet transformation. However, the method is not secure and needs further study on applying differential privacy scheme for protecting the flow data in body sensor networks.

Yoon-Su Jeong and Seung-Soo Shin (2016) developed a security management scheme for allowing users to access big data frameworks from varying networks using a key that is known to user and server. This key links the Big Data and user attribute information for protecting the privacy of big data users. This scheme is safer due to its guaranteed signal and provides assurance to the users as it uses hash-chained bit sequence values so that the information is not exposed to third-party users.

A. K. Ilavarasi and B. Sathiyabhama (2017) designed a model, which integrates the anonymization method in a learning algorithm for mitigating the overheads caused by the data mining mechanism. The utility of data is applied with the transformation for analyzing the workloads using anonymization technique. However, the method needs to be extended for sequential partitioning structures and must revise the algorithm for functioning the ensemble classifier, and the over anonymity issue needs to be addressed.

Yongjiao Sun *et al.* (2016) developed a technique, named splitting anonymization, for pointing out the contradictions composed by privacy and utility parameters. It protects social network data, which is not known to attackers. This technique refuses the direct attack, and these mechanisms are secure as compared to indirect attacks, which are more harmful than the direct attacks. However, for attaining a meaningful level of anonymity, the vertices of the graph need to be tuned so that it does not reflect as original graph.

Benjamin C. M. *et al.* (2007) have introduced a k-anonymization solution for classification. This method preserved the classification structure. Also, here, the quality of classification was preserved even for highly restrictive anonymity requirements.

Xuyun Zhang *et al.* (2015) analyzed the local-recoding problem for big data anonymization against proximity privacy breaches and attempted to identify a scalable solution to this problem. The authors presented the proximity privacy model with allowing semantic proximity of sensitive values and multiple sensitive attributes, and model the problem of local recoding as a proximity-aware clustering problem. This method improved the scalability and the time-efficiency of local-recoding anonymization, the capability of defending the proximity privacy breaches.

2.2 Challenges

Several challenges discovered from the existing techniques are as listed as follows,

- Handling of big data is a complicated task due to its large size. It is a challenging task to build a scalable algorithm for privacy preservation (Geetha, *et al.,* 2017).
- The coupling of data in cloud platform discards the traditional privacy protection measures in the cloud and can lead to economic loss or severe loss to data owners and introduces several privacy concerns to third parties in the cloud (Zhang, *et al.,* 2014).
- The existing anonymization techniques lack the compatibility for combining the analytical tools and platforms. Hence, these techniques need smart applications for processing the number of customers (Eliabeth & Sarju, 2015).
- Due to the large size of big data, it needs to be outsourced to the cloud for easy retrieval. Hence, the physical control on the data is lost, which produces insecurity in the cloud and causes damage while building privacy preservation algorithms (Mehmood, *et al.,* 2016).
- The privacy preservation algorithm is designed by considering homogeneous data sources with high probability. Hence, the algorithm is more complex if it deals with heterogeneous data sources (Geetha, *et al.,* 2017).
- The privacy algorithms need advanced features for attaining improved privacy. Hence, the algorithm is subjected to deal with complex mathematical formalisms for achieving a convex optimization (Rebollo-Monedero, *et al.,* 2017).

3. PROPOSED PRIVACY PRESERVATION MODEL USING K-ANONYMIZATION AND GWCSO

This section elaborates the proposed privacy preservation model by adapting k-anonymization criteria and GWCSO algorithm. Figure 1 illustrates the schematic diagram of the proposed privacy preservation model. The data owners and the end data users are connected to the cloud servers for exchanging and storing their confidential data in the cloud platform and to provide required information through cloud computing. Initially, multiple data owners store their data in the information database. The focus is to maximize the data privacy and data utility by providing the balance between data protection and easy data retrieval to minimize the delay. An algorithm, GWCSO, is developed for effective privacy preservation to ensure secure communication among the end users and information providers. The basic functionality of the proposed privacy preservation model depends on the anonymization process, in which the users get an assurance that the data is safe in the k anonymized database. In the proposed privacy preservation model, the components used are data owners, data users, central server, data publisher, and information database and k-anonymized database. The data users request the data from the original database, and the proposed privacy preservation model protects the data contained in the database. Initially, the records are collected from data owners and stored in the original database. Then, the records are analyzed to hide the key information for attaining more privacy, and the record is categorized using k- anonymization criteria and applies GWCSO for constructing a k-anonymized database. Finally, the data users send the request, and the response queries and the data analyst responds by retrieving the data from the database.

3.1 K-anonymized database model using k- anonymization technique

3.1.1 Information database

The information database stores the information gathered from various data owners. The data contained in the database has different records and attributes. Assume the information database contains a number of records with the varying number of attributes. The data contained in each record is expressed as follows,

$$D = \left\{ p_r \right\}; \quad 1 \le r \le I \tag{1}$$

where, p_r denotes the data present at r^{th} record, and I denotes the total number of records in the database. The record in the database is depended on attribute K. Hence, the data p_r with the total K attributes is represented as follows,

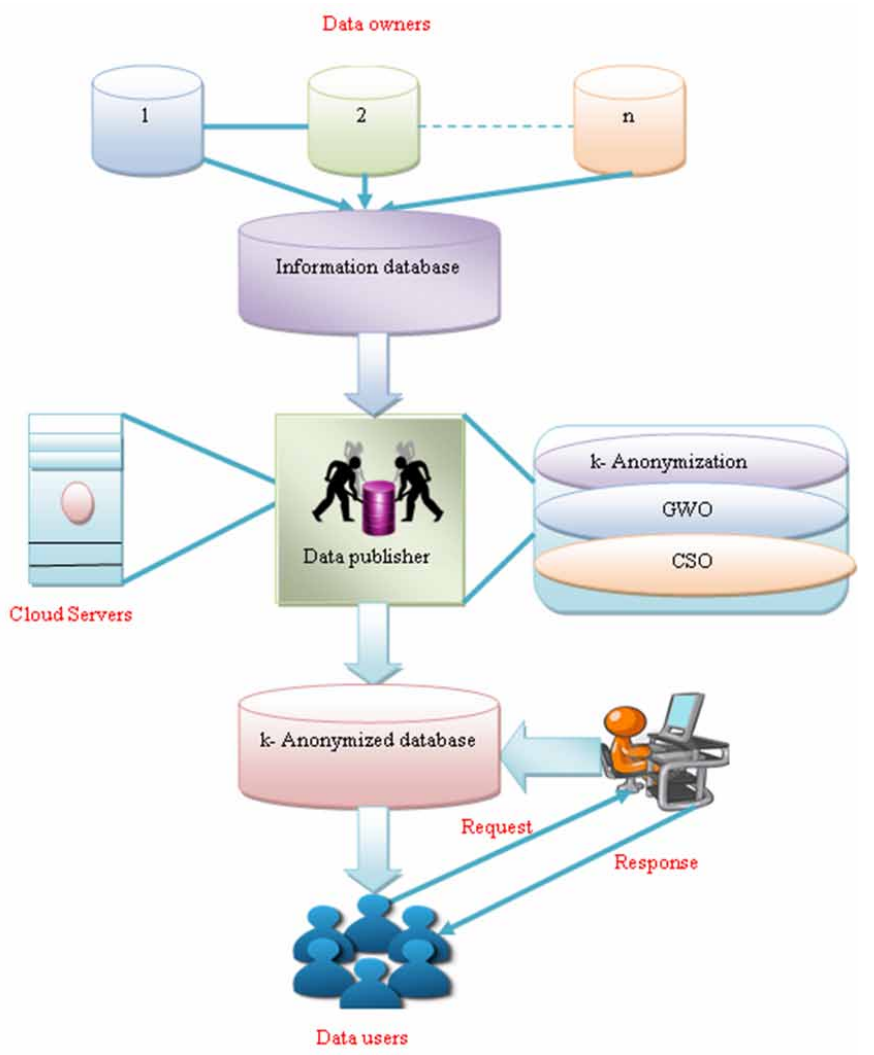
$$p_r = \left\{ t_h \right\}; \quad 1 \le h \le K \tag{2}$$

where, t_h denotes the attribute value at h^{th} position. The attributes contained in the information database are categorized as Numerical attributes and Categorical attributes.

3.2.2 k- Anonymization criterion

In K-anonymization, the optimal solution is selected based on various criteria. It prevents the various attacks by modifying the microdata which is released for business or research purposes. This is done by applying generalization and suppression techniques to the microdata. Also, it helps us in releasing a huge amount of data so that it can be used for business or research-related work by various organizations, ensures the privacy of no individual is being put in danger due to the released data by protecting released information against inference and linking attacks. The k-anonymization criterion

Figure 1. Schematic diagram of the proposed privacy preservation model with the k- Anonymity and proposed GWCSO



is applied on the information database for generating a k-anonymized database with secured information. This technique aims to find a set of clusters from the given records such that each cluster contains k records. The selection of clusters should be done in a systematic way such that the total records and the size of each cluster is k and must satisfy the criteria of k anonymization. The technique minimizes the intracluster distances for maximizing the distance between two consecutive records and to minimize the information loss. The following steps are applied on the information database to obtain k anonymized database.

Initially, the information database is clustered with different sizes of clusters and values. Then, the attributes contained in the database adapt generalization for improving the privacy of data owners and the important attributes of the database. These steps are responsible for generating an anonymization value, and the term is called anonymization constant, which is described by the users. The database clustering is based on the value k. The proposed GWCSO algorithm is used for constructing a k -anonymized database using certain criteria. The duplicate records for each quasi-

attribute are contained in the k-anonymized database. The proposed GWCSO is responsible for fulfilling the requirements of k-anonymization technique, and the generalization is based on level L. The greater value of L indicates that the data is more generalized. The privacy of the database can be increased by improving the generalization, but the utility is decreased. Hence, a generalization level must be selected in an optimal manner for balancing the utility and privacy.

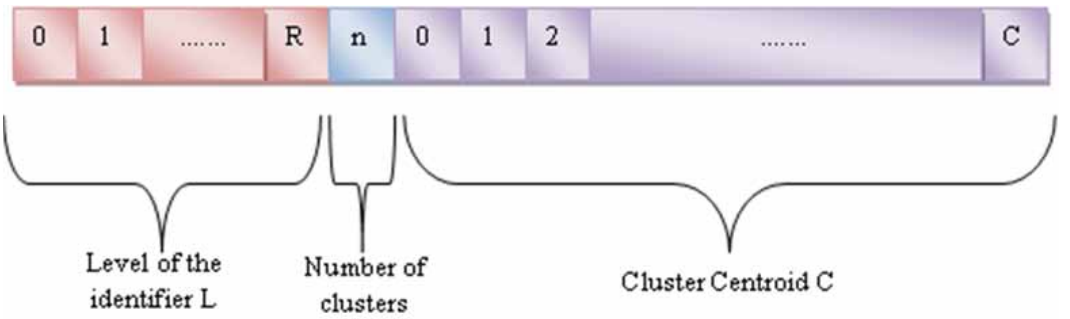
3.2.3 Proposed GWCSO for privacy preservation

This section presents the proposed GWCSO developed for the privacy preservation. The solution encoding, fitness function and the algorithm of the proposed GWCSO are explained below.

i) Solution encoding

Solution encoding plays an important role in determining the optimal solution in optimization problems. The solution encoding should help the optimization algorithms to find the optimal solution from a set of solutions. The solution of the proposed k-anonymization represents a number of solutions, from which the best solution is chosen using the derived fitness function. The solution encoding using proposed k-anonymization is depicted in figure 2. The solution encoding is represented with a combination of three parameters, which are identifier level, the total number of clusters, cluster Centroid C. The optimization problem identifies the identifier level, the clusters, and the cluster centroids.

Figure 2 Solution encoding



The clusters in the database are represented as,

$$C = n \times K \tag{3}$$

where, K denotes the attributes in the database, and n denotes the total number of clusters. The attributes are defined as a combination of both the numerical and the categorical category. The attributes contained in the database are represented as,

$$K = \beta^Q + \beta^S + \beta^J + \alpha^T + \alpha^U + \alpha^V \tag{4}$$

where, β^Q denotes the numerical quasi-attribute, β^S denotes the numerical sensitive attribute, and the term β^J denotes numerical information provider name, $\alpha^T, \alpha^U and \alpha^V$ denote the categorical attributes.

ii) Fitness evaluation

After solution encoding, the fitness function is evaluated for the privacy preservation model. The fitness function is devised using the privacy and the utility parameters. Here, the fitness function is considered as a minimization function, which means both privacy and utility should be smaller. The fitness function is represented as,

$$\underset{\min}{Fitness}(\sigma) = \frac{1}{2}\left[M + N\right] \tag{5}$$

where, M denotes the privacy parameter, and N denotes the utility of the cluster σ from the database. The privacy parameter is formulated from (Li, et al., 2012). Here, the fitness is considered as a minimization function, so the value of privacy taken is 0 for satisfying the k-anonymization criterion. The privacy is formulated as,

$$M(\sigma) = \begin{Bmatrix} 0\ ; & If\ k-anonymization\ is\ satisfied \\ 1\ ; & otherwise \end{Bmatrix} \tag{6}$$

The utility parameter (Kabir, et al., 2011) for evaluating the fitness function is considered as a measure of IL occurred while implementing the privacy preservation process. Assume χ denotes a set of data with f numeric quasi-identifiers with $\beta_1, \beta_2, \ldots, \beta_f$ and w categorical quasi-identifiers $\alpha_1, \alpha_2, \ldots, \alpha_w$. Let ν divides the set of data as $\nu = \{\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_n\}$. Here, $\upsilon\alpha_z$ denotes a taxonomy tree defined for the domain α_z.

Consider a cluster σ, which divides a set of data into numerical and categorical attributes, where $\beta_{y_{\max}}$ and $\beta_{y_{\min}}$ denote the largest and the smallest values of records in σ and $G\beta_{y_{\max}}$ and $G\beta_{y_{\min}}$ be the largest and the smallest values of record in χ with respect to the numeric attribute, where β_y is $(y = 1,2,\ldots,f)$ and $\bigcup\alpha_z$ be a union set of values in σ with respect to categorical attribute α_z (z=1,2...w). The utility is given for each categorical and numerical attribute and is represented as,

$$N = IL(\sigma) = \left|\sigma\right|.\left(\sum_{y=1}^{f}\frac{\beta_{y_{\max}} - \beta_{y_{\min}}}{G\beta_{y_{\max}} - G\beta_{y_{\min}}} + \sum_{z=1}^{w}\frac{X(\wedge(\bigcup\alpha_z))}{X(\upsilon\alpha_z)}\right) \tag{7}$$

where, σ is the number of records in ν, $\nu\bigcup\alpha_z$ denotes the subtree rooted at the lowest common ancestor of every value in $\bigcup\alpha_z$ and $X(\upsilon)$ denotes the height of the taxonomy tree υ.

The total information loss of χ for n clusters is the sum of the information loss of each $(l = 1,2,\ldots,n)$ and is given by,

$$IL(G) = \sum_{l=1}^{n} IL(\sigma_l) \tag{8}$$

Therefore the total information loss is given by,

$$IL(\sigma_l) = \sum_{l=1}^{n} \left|\chi_l\right| . \left(\sum_{e=1}^{C} \frac{\beta_{y,y_{\max}} - \beta_{y,y_{\min}}}{G\beta_{y,y_{\max}} - G\beta_{y,y_{\min}}} + \sum_{f=1}^{W} \frac{X(\wedge(\bigcup \alpha_z))}{X(\upsilon\alpha_z)} \right) \tag{9}$$

3.3 Proposed GWCSO Algorithm

The proposed GWCSO algorithm is developed by the integration of GWO (Mirjalili, et al., 2014) and CSO (Bahrami, *et al.,* 2017) for constructing the database using the k-anonymization technique. GWO mostly depend on four search agents namely alpha, beta, delta and omega and are considered as best agents. The attacking of prey is usually guided by alpha and is responsible for making effective decisions about hunting, sleeping and so on. The algorithm has fast convergence and hence, more suitable for the big data. Moreover, GWO has the strongest ability to capture the exact position of prey and encircle them accordingly. The position of the wolves is updated by adding a new term with the inclusion of CSO algorithm for finding the optimal position. CSO follows the swarm intelligence along with the optimization algorithm for hunting the prey using its intelligent behavior. Hence, CSO and GWO are integrated to develop GWCSO for constructing the k-anonymized database for securing communication. The steps in the proposed GWCSO are depicted as follows,

Step 1: Initialization

Initially, the population of the algorithm is initialized. The entire population size of the algorithm is denoted $b.$ and is given by,

$$A = \left\{A_1, A_2, \ldots, A_j \ldots\ldots, A_b\right\} \tag{10}$$

where, A_b denotes the total population size where $1 \leq j \leq b.$

Step 2: Calculate the fitness of each search agent

After initializing the population, the fitness function of each value in the population is determined using the fitness function as given in equation (5). Among the best fitness values computed for each population, the best three values are determined.

Step 3: Updating the position according to the search agents

The grey wolves have the higher capability for recognizing the exact location of the prey and to surround them. The hunting behavior is guided by alpha, beta and delta wolves, which participate for finding the prey. The hunting behavior is formulated based on the location of the prey to find the best three solutions and oblige other search agents to update their positions according to the best search agents.

The encircling behavior of the GWO is represented as,

$$B = \mid F \cdot A_q(l) - \vec{A}(l) \mid \tag{11}$$

where, F denotes the coefficient vector, A_q represent the position vector of the prey and $\vec{A}$ denotes the position vector of the grey wolf. The coefficient vectors are represented as,

$$E = 2\vec{d} \cdot \vec{u}_1 - \vec{d} \tag{12}$$

$$F = 2 \cdot \vec{u}_2 \tag{13}$$

where, d denotes a random number, which linearly decreases from 2 to 0, u_1 and u_2 denote the random number that varies from 0 to 1.

According to GWO (Mirjalili, *et al.,* 2014), the updated position of grey wolves is given by,

$$A(l+1) = \frac{A_1 + A_2 + A_3}{3} \tag{14}$$

where, A_1 denotes the first search agent, A_2 denotes the second search agent, A_3 denotes the third search agent, and l represents the iteration.

For evaluating the optimal solution, a new term is added in the above equation to update the position that is close to the prey's position. The swarm intelligence concept induced with optimization algorithms can easily simulate the intelligent behavior of animals for catching the prey. Further, the positions of the search agents are updated.

In GWCSO, the update is performed using GWO algorithm by adding a new term A_4, which is the position update equation of the CSO algorithm. Hence, the position update of the proposed GWCSO is mathematically represented as,

$$A(l+1) = \frac{A_1 + A_2 + A_3 + A_4}{4} \tag{15}$$

where,

$$\vec{A}_1 = \vec{A}_\alpha - \vec{E}_1(\vec{B}_\alpha) \tag{16}$$

$$\vec{A}_2 = \vec{A}_\beta - \vec{E}_2(\vec{B}_\beta) \tag{17}$$

$$\vec{A}_3 = \vec{A}_\chi - \vec{E}_3(\vec{B}_\chi) \tag{18}$$

$$\vec{A}_4 = A(l) + Y(l+1) \tag{19}$$

where, A_α represent the best search agent, $, A_\beta$ expresses the second best search agent and A_χ denote the third best search agent for capturing the position of prey, whereas other wolves update their position in a random manner around the prey. E_1, E_2 and E_3 denote the coefficient vectors. The grey wolves encircle the prey during the hunt. Hence, the encircling behavior is denoted as B_α , B_β and B_χ

Equation (19) is the updated position of the CSO algorithm that is obtained by the position and the velocity at the current iteration. Once the prey is determined, the velocity of the cat is changed accordingly to catch the prey. Hence, the velocity of the cat is given by,

$$Y(l+1) = Y(l) + s_1 \times v_1 \left(A^* - A(l)\right) \tag{20}$$

where, v_1 is the constant and s_1 a random value in the range of [0,1]. The equation obtained after substituting equation (20) in equation (19) is given by,

$$\vec{A}_4 = A(l) + Y(l) + s_1 \times v_1 (A^* - A(l)) \tag{21}$$

As the best solution, A^*, of the CSO algorithm is equal to that of the GWO, the above equation can be represented as,

$$\vec{A}_4 = A(l) + Y(l) + s_1 \times v_1 (A_\alpha - A(l)) \tag{22}$$

4. RESULTS AND DISCUSSION

This section elaborates the results of the proposed anonymization technique with the existing techniques for effective privacy preservation.

4.1 Experimental Setup

The proposed privacy preservation model is implemented in JAVA framework. The NetBeans is considered as a development tool for the implementation. The system uses Windows 10 OS with 4GB RAM and i5 processor.

4.1.1 Dataset description

This technique employs the standard Adult dataset from the UCI Machine learning repository (Hettich & Merz, 1998) for analyzing the results. The adult dataset is taken from the 1994 Census database, and the extraction was done by Barry Becker. The total number of instances is 48842 with 14 numbers of attributes. The data set is characterized as a multivariate and attribute is characterized as an integer. The attributes considered are age, workclass, education, occupation, and native-country and so on. The adult dataset is considered as a standard dataset for the k-anonymization process. The task associated is classification, and the total web hits are 1201358.

4.1.2 Evaluation metrics

The evaluation metrics used for the proposed preservation model are explained as follows,

Information loss (IL): The exact result obtained after k- anonymization process for effective privacy preservation is defined as IL. The duplication of records leads to IL and thereby, minimizes the utility of the database.

Classification accuracy (CA): The CA is defined as a measure to define the database privacy. The proportion obtained for the correct classification of records in the cluster to the total number of clusters is termed as CA. The CA value should be greater for achieving better performance.

$$CA = \frac{N_c}{n} \tag{23}$$

where, N_c denotes the number of clusters to be correctly classified and n denotes the total clusters.

4.1.3 Comparative methods

The comparative methods used for the evaluation are k-anonymization (Fung, et al., 2007), k-Diversity (Eliabeth & Sarju, 2015), K-anonymization + GA (Applied Genetic Algorithm (GA) in (Fung, *et al.,* 2007)), Duplicate-Divergence-Different properties enabled Dragon Genetic Algorithm (DDDG), k-anonymization + Dragon+ PSO (Applied Dragonfly Algorithm and PSO in (Fung, *et al.,* 2007)) are analyzed and compared with the proposed k-anonymization + GWCSO. The performance of each model is evaluated using the information of the adult dataset.

4.2 Comparative Analysis

This section illustrates the comparative analysis of the proposed k-anonymization + GWCSO algorithm. The values of k are varied as 2 and 4 for analyzing the performance of the comparative methods.

4.2.1 Comparative analysis based on k=2

The analysis of the comparative techniques for k=2 is depicted in figure 3. The analysis based on IL with a varying number of clusters is shown in figure 3a. When the cluster size is 2, the values of IL for existing techniques, like k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.525, 0.289, 0.250, 0.088, 0.057, and 0.012. Similarly, for the cluster size 4, the IL values measured by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and proposed k- anonymization + GWCSO are 0.225, 0.158, 0.134, 0.057, 0.005, and 0.005. From the above data, it is seen that the proposed k- anonymization + GWCSO has fewer IL values when cluster size is varied from 2 to 5. The analysis based on CA metric is depicted in figure 3b. When cluster size is 5, the corresponding CA values obtained by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and proposed k- anonymization + GWCSO are 0.633, 0.633, 0.711, 0.792, 0.797, and 0.798. Similarly, for the cluster size 3, the CA values calculated by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.614, 0.633, 0.645, 0.756, 0.759, and 0.766. The proposed k-anonymization + GWO have attained an improved classification accuracy of 0.798 at cluster size=4.

Figure 3a. Comparative analysis of various models for k=2 a) IL

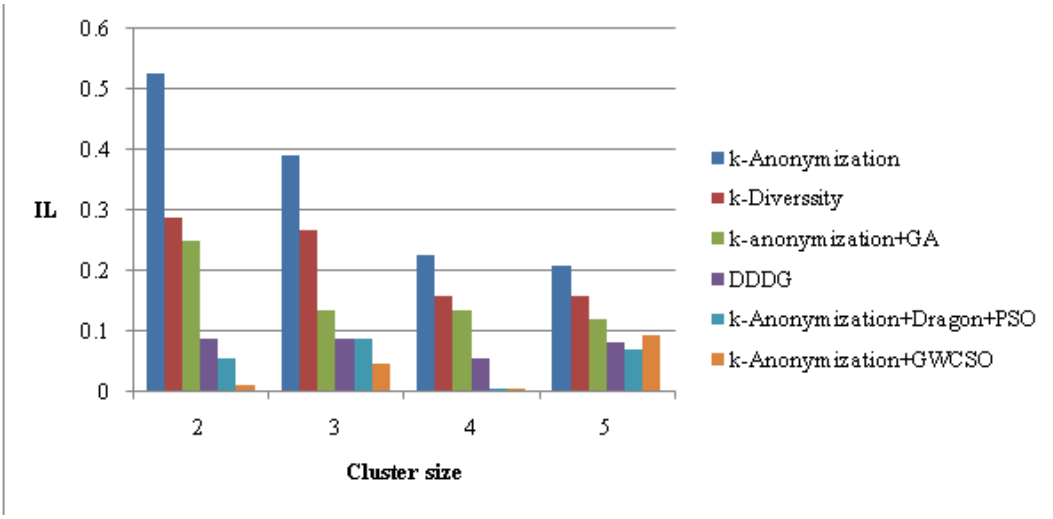
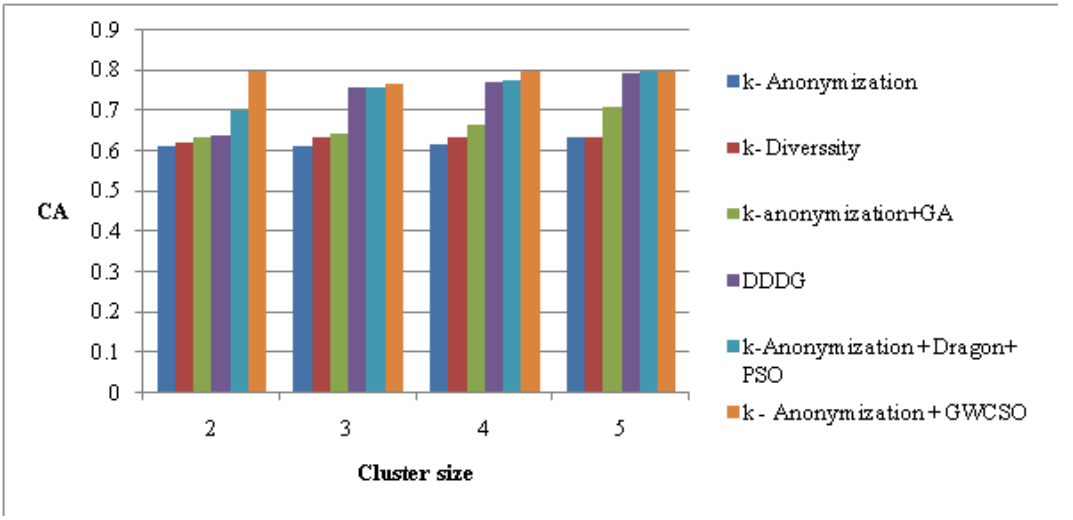


Figure 3b. Comparative analysis of various models for k=2 b) CA



4.2.2 Comparative analysis based on k=3

The comparative analysis of the proposed k-anonymization + GWCSO algorithm for k=3 is depicted in figure 4. The analysis based on IL with varying number of clusters is shown in figure 4a. When cluster size is 4, the values of IL for existing techniques, like k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.249, 0.114, 0.095, 0.021, 0.018, and 0.015. Similarly, for the cluster size 3, the IL values measured by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and proposed k- anonymization + GWCSO are 0.349, 0.203, 0.098, 0.044, 0.043, and 0.011. From the above data, it is seen that the proposed k- anonymization + GWCSO have less IL values when cluster size is varied from 2 to 5. The analysis based on CA metric is depicted in figure 4b. When cluster

Figure 4a. Comparative analysis of the various models for k= 3 a) IL

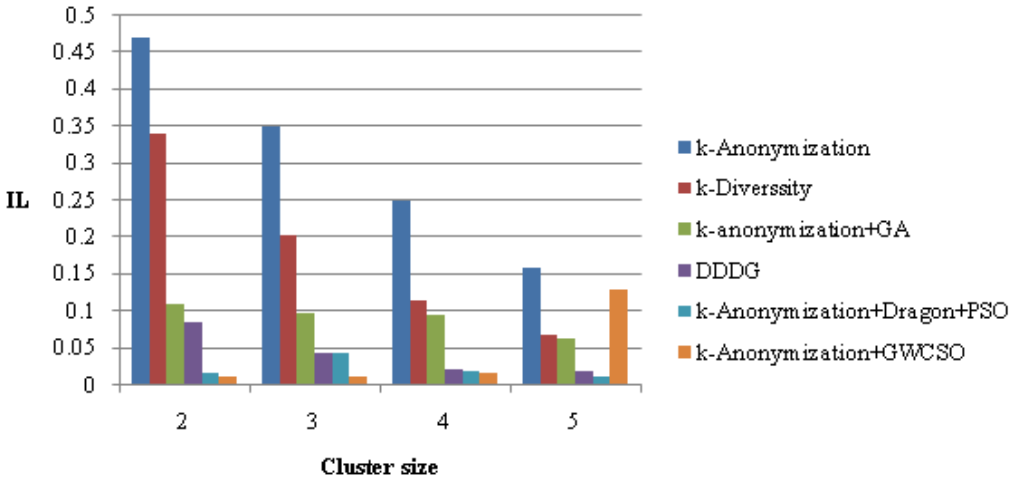
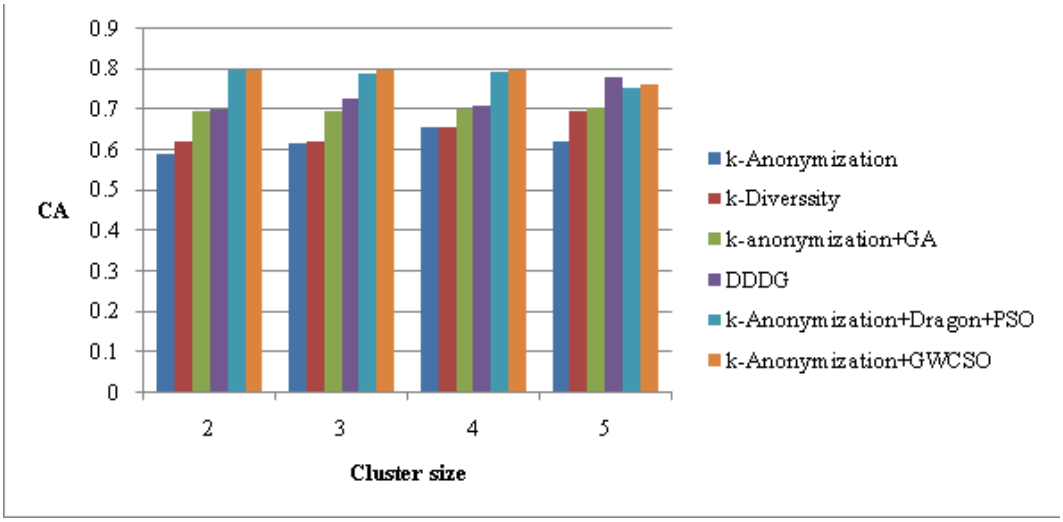


Figure 4b. Comparative analysis of the various models for k=3 b) CA



size is 2, the corresponding CA values obtained by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and proposed k- anonymization + GWCSO are 0.593, 0.622, 0.696, 0.703, 0.797, and 0.798. Similarly, for the cluster size 5, the corresponding CA values calculated by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.621, 0.695, 0.706, 0.782, 0.753, and 0.764.

4.2.3 Comparative analysis based on k=4

The comparative analysis of the proposed k-anonymization + GWCSO algorithm on the basis of IL and CA metric for k=4 is depicted in figure 5. The analysis based on IL parameter with varying

Figure 5a. Comparative analysis of the various models for k= 4 a) IL

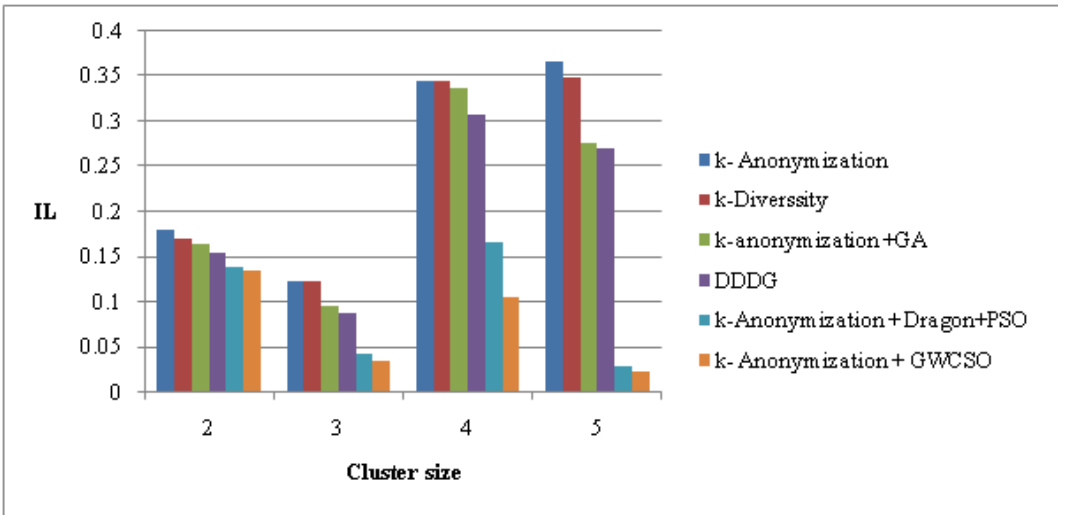
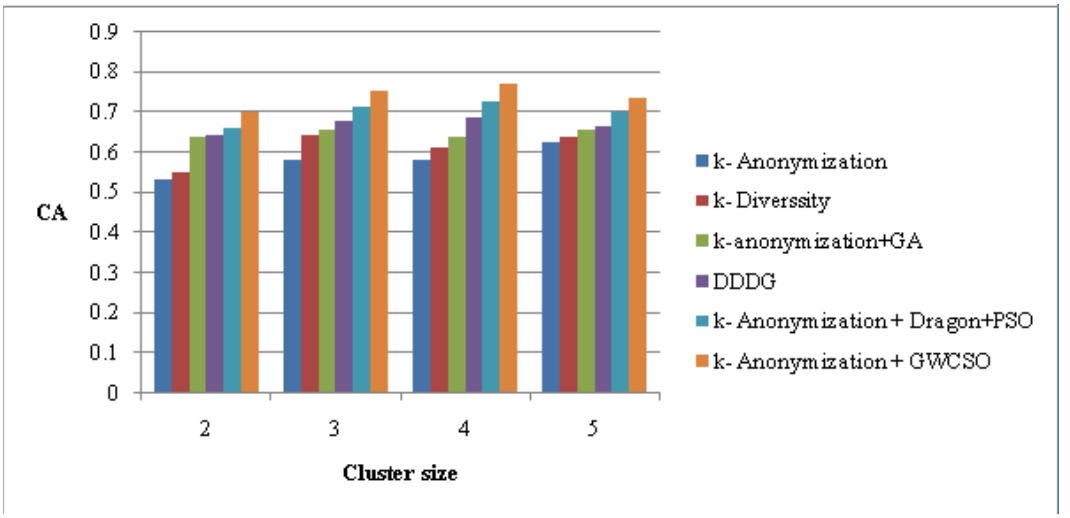


Figure 5b. Comparative analysis of the various models for k=4 b) CA



number of clusters is depicted in figure 5a. When the cluster size is 3, the IL values measured by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.123, 0.123, 0.096, 0.087, 0.043, and 0.034. Similarly, for the cluster size 2, the IL values calculated by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.180, 0.171, 0.164, 0.154, 0.138, and 0.136. The analysis with CA metric for varying number of clusters is depicted in figure 5b. When the cluster size is 5, the CA values measured by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.625, 0.638, 0.656, 0.666, 0.701, and 0.736. Similarly, for cluster size 3, the CA values measured by k-anonymization, k- diversity, k-anonymization + GA, DDDG,

Figure 6a. Comparative analysis of the various models for k= 5 a) IL

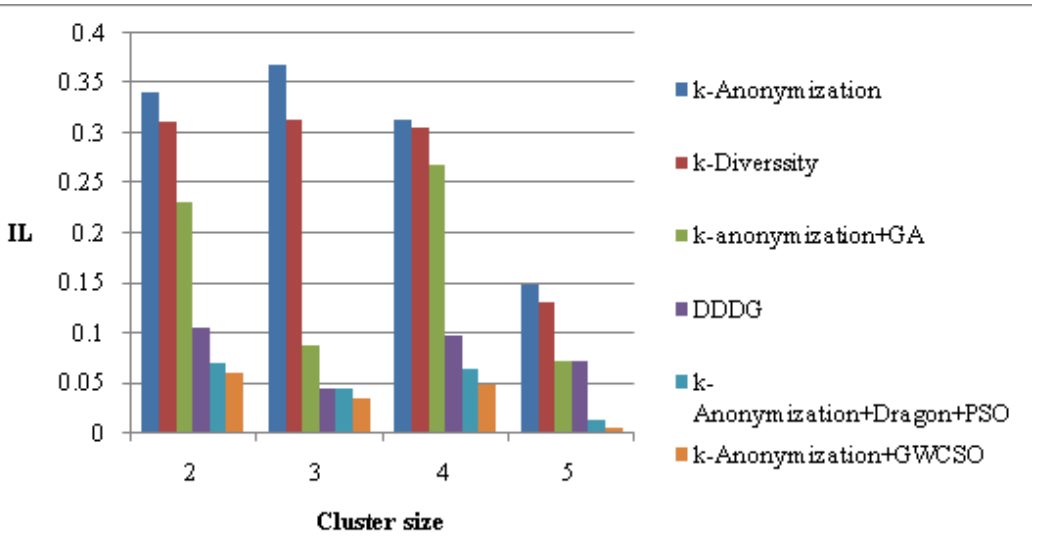
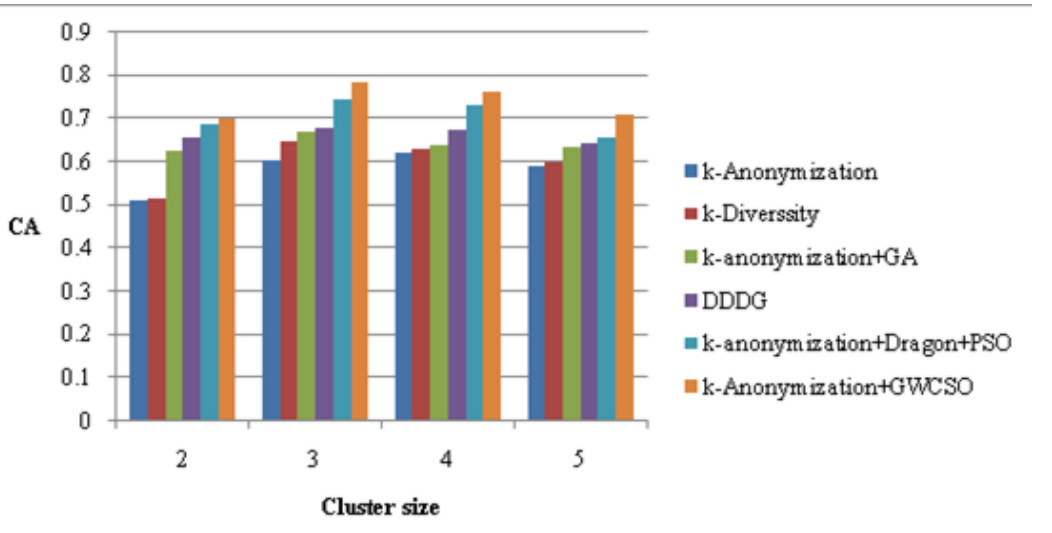


Figure 6b. Comparative analysis of the various models for k=5 b) CA



k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.581, 0.645, 0.656, 0.678, 0.714, and 0.753. The proposed k- anonymization + GWCSO has improved CA with value 0.770 for cluster size 4.

4.2.4 Comparative analysis based on k=5

The comparative analysis of the proposed k-anonymization + GWCSO algorithm on the basis of IL and CA metric for k=5 is depicted in figure 6. The analysis based on IL parameter with varying number of

clusters is depicted in figure 6a. When cluster size is 2, the IL values measured by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.340, 0.310, 0.232, 0.104, 0.070, and 0.061. Similarly, for the cluster size 4, the IL values calculated by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.313, 0.306, 0.267, 0.097, 0.065, and 0.048. The analysis with CA metric for varying number of clusters is depicted in figure 6b. When cluster size is 4, the CA values measured by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k-anonymization + GWCSO are 0.621, 0.632, 0.638, 0.674, 0.731, and 0.761. Similarly, for the cluster size 3, the CA values measured by k-anonymization, k- diversity, k-anonymization + GA, DDDG, k-anonymization + Dragon + PSO, and the proposed k- anonymization + GWCSO are 0.603, 0.648, 0.672, 0.679, 0.747, and 0.787.

4.3 Discussion

The discussion regarding the maximum performance attained by the existing techniques with the proposed technique for k=2, k=3,k=4, and k=5 is depicted in table 1. When the value of k is 2, the proposed k- anonymization with GWCSO algorithm has the IL and CA values as 0.005 and 0.798, respectively. When the value of k is 3, the proposed k-anonymization + GWCSO algorithm provides IL and CA value as 0.128 and 0.798. At k is 4, the proposed k- anonymization + GWCSO algorithm shows better performance with IL and CA value as 0.022 and 0.770, respectively. Similarly, when the value of k is 5, the proposed k- anonymization with GWCSO algorithm provides IL value as 0.006 and CA value as 0.787. The proposed k- anonymization with GWCSO algorithm shows superior performance than the existing methods with minimum IL value as 0.005 and maximum CA value as 0.798, at k=2. The proposed method resists several attacks, and it ensures that the secure communication is done among the end users. It is more suitable for handling the big data.

Table 1. Comparative discussion

k value	Evaluation metrics	Comparative models					
		k-anonymization	k-Diversity	k-anonymization + ga	DDDF	k-anonymization + Dragon +PSO	k-anonymization + GWCSO
k= 2	**IL**	0.208	0.157	0.121	0.0572	0.005	**0.005**
	CA	0.6332	0.6333	0.711	0.792	0.797	**0.798**
k=3	**IL**	0.159	0.067	0.062	0.020	0.013	**0.128**
	CA	0.621	0.695	0.706	0.782	0.797	**0.798**
k= 4	**IL**	0.123	0.123	0.096	0.087	0.029	**0.022**
	CA	0.625	0.645	0.656	0.689	0.727	**0.770**
k=5	**IL**	0.148	0.131	0.073	0.045	0.013	**0.006**
	CA	0.621	0.648	0.672	0.679	0.747	**0.787**

5. CONCLUSION

In this paper, an algorithm, named GWCSO is proposed for effective privacy preservation in big data. The proposed GWCSO is developed by modifying the update process of GWO using CSO algorithm and is employed for constructing the k-anonymization database, where k duplicate records

are created within the original database. The GWCSO is employed in the k-anonymized database for hiding the confidential information of the data owners in order to provide secure communication among the end users. The proposed algorithm generates the parameters required for constructing the k-anonymized database optimally, assuring the k-anonymization criteria. Hence, the privacy and the utility are the two metrics considered for computing the fitness of the solutions such that the best solution has maximum fitness. The proposed technique is implemented on JAVA platform and is experimented using the adult database. From the analysis, it is noted that the proposed technique provides maximum classification accuracy with less information loss. Moreover, when the value of k is 2, the value of IL and CA are 0.005 and 0.798, respectively.

REFERENCES

Karle, T., & Vora, D. (2017). Privacy preservation in big data using anonymization techniques. *Proceedings of International Conference on Data Management, Analytics and Innovation (ICDMAI)*, 340-343.

Alabdulatif, Khalil, & Yi. (2020). Towards secure big data analytic for cloud-enabled applications with fully homomorphic encryption. *Journal of Parallel and Distributed Computing, 137*.

Antony, P.J., & Antony, S.T. (2016). A Survey on Privacy Preservation in Big Data. *International Journal of Engineering Science Invention Research and Development, 3*(3).

Atiewi, S., Al-Rahayfeh, A., Almiani, M., Yussof, S., Alfandi, O., Abugabah, A., & Jararweh, Y. (2020). Scalable and Secure Big Data IoT System Based on Multifactor Authentication and Lightweight Cryptography. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 113498–113511.

Bahrami, M., Bozorg-Haddad, O., & Chu, X. (2017). Cat Swarm Optimization (CSO) Algorithm. *Advanced Optimization by Nature-Inspired Algorithms,* 9-18.

Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. *Proceedings of International Conference on Data Engineering (ICDE'05)*, 217-228.

Cate, F.H., & Schönberger, V.M. (2012). *Notice and Consent in a World of Big Data.* Microsoft Global Privacy Summit Summary Report and Outcomes.

Denglong, Lv., & Zhu, S. (2020). *Achieving secure big data collection based on trust evaluation and true data discovery* (Vol. 96). Computers & Security.

Dhumane, A. V., & Prasad, R. S. (2017). Multi-objective fractional gravitational search algorithm for energy efficient routing in IoT. *Wireless Networks*, 1–15.

Dijk, M. V., & Juels, A. (2010). On the impossibility of cryptography alone for privacy-preserving cloud computing. *Proceedings of 5th USENIX conference on Hot topics in security*, 1-8.

Eliabeth, S. S., & Sarju, S. (2015). Big data Anonymization Using One Dimensional and Multidimensional Map Reduce Framework on Cloud. *International Journal of Database Theory and Application*, *8*(6), 253–262.

Fung, B. C. M., Wang, K., & Yu, P. S. (2007). Anonymizing Classification Data for Privacy Preservation. *IEEE Transactions on Knowledge and Data Engineering*, *19*(5), 711–725.

Geetha, M.C.S., Selvakumar, N., & Jose, W.W. (2017). Analyzing The Privacy Preserving Using Big Data Techniques. *International Journal of Innovative Research in Science and Engineering, 3*(4).

Gosain, A., & Chugh, N. (2014). Privacy Preservation in Big Data. *International Journal of Computers and Applications*, *100*(17).

Guan, Z., & Si, G. (2017). Achieving privacy-preserving big data aggregation with fault tolerance in smart grid. *Digital Communications and Networks*, *3*(4), 242–249.

Hettich, C. B. S., & Merz, C. (1998). *UCI Repository of machine learning databases*. https://archive.ics.uci.edu/ml/datasets.html

Ilavarasi, A. K., & Sathiyabhama, B. (2017). An evolutionary feature set decomposition based anonymization for classification workloads Privacy Preserving Data Mining. *Cluster Computing*, *20*(4), 3515–3525.

Jeong, Y. S., & Shin, S. S. (2016). An efficient authentication scheme to protect user privacy in seamless big data services. *Wireless Personal Communications*, *86*(1), 7–19.

Kabir, E., Wang, H., & Bertino, E. (2011). Efficient, systematic clustering method for k-anonymization. *Acta Informatica*, *48*(1), 51–66.

Krishnamoorthy, N., & Asokan, R. (2014). *Optimized Resource Selection to Promote Grid Scheduling Using Hill Climbing Algorithm*. Academic Press.

Li, C. (2020). Information processing in Internet of Things using big data analytics. *Computer Communications*, *160*, 718–729.

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *Proceedings of 23rd International Conference on Data Engineering*, 106 – 115.

Li, T., Li, N., Zhang, J., & Molloy, I. (2012). Slicing: A New Approach for Privacy Preserving Data Publishing. *IEEE Transactions on Knowledge and Data Engineering*, *24*(3), 561–574.

Lin, C., Wang, P., Song, H., Zhou, Y., Liu, Q., & Wu, G. (2016). A differential privacy protection scheme for sensitive big data in body sensor networks. *Annales des Télécommunications*, *71*, 465–475.

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. *IEEE Access: Practical Innovations, Open Solutions*, *4*, 1821–1834.

Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, *69*, 46–61.

Nipanikar, S.I., Deepthi, V.H., & Kulkarni, N. (2017). *A sparse representation based image steganography using Particle Swarm Optimization and wavelet transform*. Academic Press.

Priyanka, Z., Nagaraju, K., & Venkateswarlu, Y. (2014). Data Anonymization Using Map Reduce on Cloud based A Scalable Two-Phase Top-Down Specialization. *International Journal on Recent and Innovation Trends in Computing and Communication*, *2*(12), 3879–3883.

Ratre, A., & Pankajakshan, V. (2017). Tucker visual search-based hybrid tracking model and Fractional Kohonen Self-Organizing Map for anomaly localization and detection in surveillance videos. *Imaging Science Journal*, 1–16.

Rebollo-Monedero, D., Forné, J., Soriano, M., & Allepuz, J. P. (2017). p-Probabilistic k-anonymous micro aggregation for the anonymization of surveys with uncertain participation. *Information Sciences*, *382*, 388–414.

Salido, J. (2012). *Differential privacy for everyone*. White Paper, Microsoft Corporation.

Sedayao, J. (2012). *Enhancing cloud security using data anonymization.* White Paper, Intel Corporation.

Sei, Y., Okumura, H., Takenouchi, T., & Ohsuga, A. (2016). Anonymization of Sensitive Quasi-Identifiers for l-diversity and t-closeness. *IEEE Transactions on Dependable and Secure Computing*.

Shelke, P. M., & Prasad, R. S. (2018). An improved anti-forensics JPEG compression using Least Cuckoo Search algorithm. *Imaging Science Journal*, *66*(3), 169–183.

Singh, A., & Rai, A. (2020). Enhancement of Big Data Security in Cloud Environment. *European Journal of Molecular & Clinical Medicine*, *7*(4).

Sreedhar, R., & Umamaheshwari, D. (2014). Big-Data Processing With Privacy Preserving Map-Reduce Cloud, *International Journal of Innovative Research in Science*, *Engineering and Technology*, *3*(1), 343–350.

Sun, Y., Yuan, Y., Wang, G., & Cheng, Y. (2016). Splitting anonymization: A novel privacy-preserving approach of social network. *Knowledge and Information Systems*, *47*(3), 595–623.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 557–570.

Tang, J., Cui, Y., Li, Q., Ren, K., Liu, J., & Buyya, R. (2016). Ensuring Security and Privacy Preservation for Cloud Data Services. *ACM Computing Surveys*, *49*(1).

Thanamani, A. S. (2017). Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data. *Advances in Computational Sciences and Technology*, *10*(2), 247–253.

Wu, Y., Huang, H., Wu, N., & Wang, Y. (2020). *An incentive-based protection and recovery strategy for secure big data in social networks* (Vol. 508). Information Sciences.

Xuezhen, H., Jiqiang, L., Zhen, H., & Jun, Y. (2014). A New Anonymity Model for Privacy-Preserving Data Publishing. *Communications System Design,* 47-59.

Yang, J., Liu, Z., Jia, C., Lin, K., & Cheng, Z. (2014). New Data Publishing Framework in the Big Data Environments. *Proceedings of IEEE International Conference on P2P*, *Parallel, Grid, Cloud, and Internet Computing,* 363-366.

Zhang, X., Dou, W., Pei, J., Nepal, S., Yang, C., Liu, C., & Chen, J. (2014). Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud. *IEEE Transactions on Computers*, *64*(8), 2293–2307.

Zhang, X., Dou, W., Pei, J., Nepal, S., Yang, C., Liu, C., & Chen, J. (2015). Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud. *IEEE Transactions on Computers*, *64*(8), 1–14.

Zhang, X., Yang, C., Nepal, S., Liu, C., Dou, W., & Chen, J. (2013). A map reduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud. *Proceedings of Third International Conference on Cloud and Green Computing*, 105-112.