Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Privacy Preserving Using K Member Gaussian Kernel Fuzzy C Means and Self Adaptive Honey Badger for Online Social Networks

Sivasankari K[1,*], Umamaheswari K M[2]

[1] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai 600089, Tamil Nadu, India
[2] Department of Computing Technologies, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur 603203, amil Nadu, India

ARTICLE INFO

Article history:
Received 9 November 2023
Received in revised form 28 March 2024
Accepted 14 April 2024
Available online 22 May 2024

Keywords:
Anonymization; Gaussian kernel; Optimization; Information loss; Social Networks

ABSTRACT

An online social network (OSN) gives users a strong platform to communicate and exchange information. Protecting publicly published data from individual identification is the primary problem in sharing social network databases. The most popular method for protecting privacy is anonymizing data, which involves deleting or altering some information while maintaining as much of the original data as feasible. This work presents a combination anonymizing algorithm, which is based on k member Gaussian kernel fuzzy c means clustering and self-adaptive honey badger optimization technique (KGKFCM-SAHBO). As part of the suggested anonymization process, the various users are divided into C clusters, each of which has at least K users, using a K-member Gaussian kernel fuzzy c means clustering technique. After that, the primary clusters are further optimized using the self-adaptive honey badger optimization approach (SAHBO) to further anonymize the data and network graph. Using the Yelp dataset, the experimental findings demonstrate the efficiency of the proposed model and evaluate a number of parameters, including execution time, degree of anonymization, and information loss. The experimental results show that the proposed strategy reduced information loss and improved the degree of anonymization when compared to existing methodologies.

1. Introduction

Social networks may be mined and studied to answer a variety of fascinating issues, such how groups change over time or how opinions spread. Concerns regarding the privacy of users of social networks are growing as they are released. Understanding a social network and its activity better while safeguarding the privacy of its members is the goal of privacy-preserving analysis [1]. Maintaining privacy, the first thing that springs to mind when hearing this word is what type of privacy will be protected. When one organization transfers user data to another organization for a specified reason, privacy preservation is required. For instance, Canatics is a Canadian business. In order to detect fraud from insurance claimants, Canatics must gather all available insurance company

* *Corresponding author.*
E-mail address: sivasank1@srmist.edu.in

https://doi.org/10.37934/araset.45.2.2537

data. However, doing so may result in the disclosure of some sensitive or private information about insurance holders to Canatics. As a result, safeguarding an individual's personal data becomes a top research priority in privacy-preserving data mining [2].

Safeguarding an individual's personal information becomes a major research concern in privacy-preserving data mining [3]. Many data privacy models, such as l-diversity, t-closeness, p-sensitive, and k-anonymity, can be used to anonymize sensitive data. The k-anonymity model is one of the most often used techniques for privacy protection [4]. K-anonymity is a well-liked method for ensuring privacy. By requiring that every record in the released data be identical to at least k −1 other records in terms of a set of characteristics known as quasi-identifiers, the k-anonymity model ensures privacy [5].

Applying Fuzzy C-Means (FCM) to social network analysis faces several challenges. The high dimensionality of social network data, the dynamic nature of relationships, and the presence of noise and outliers pose difficulties for FCM. The algorithm's sensitivity to initialization, scalability concerns with large-scale networks, and the inherent difficulty in interpreting fuzzy membership values add complexities. Additionally, validating and evaluating clusters in the context of social networks, handling missing or incomplete data, addressing various types of relations, and managing computational complexity contribute to the overall challenges. To overcome these issues, it is crucial to explore alternative clustering methods, adapt FCM to specific social network characteristics, and consider hybrid approaches that integrate fuzzy clustering with complementary techniques.

Eliminating any explicit identifiers is the first step in the anonymization process. According to the pre-established quasi-identifiers, each record must be identical to at least k-1 other records. The K-anonymity model's relative conceptual simplicity and efficacy have led to substantial research recently as a potential definition of privacy in data publishing [6]. The goal of applying K-anonymity algorithms to data is to stop attackers from identifying users through connection assaults, but we also need to make sure that the published data is as close to the source data as possible [7]. However, locating the optimal K-anonymity data is an NP-hard task. It would take a considerable amount of computations to get the published dataset for a large dataset with high attribute dimensions that fulfills the K-anonymity model by generalization [10].

2. Related Works

A framework was proposed by Kiran and Shirisha [11] and its performance was assessed in two ways. It initially maintained the accuracy of the data mining model. Second, it minimizes data loss while maintaining the privacy of the original data. Removing or transferring personally identifying information was the primary reason for adopting k-anonymity.

Srijayanthi and Sethukarasi [12] have investigated a clustering-based privacy-preserving strategy that used feature selection techniques in addition to anonymization. The suggested model is divided into two stages: feature selection and anonymization. In the first stage, Spearman's correlation coefficient was used to eliminate any redundant characteristics from the dataset and symmetrical uncertainty (SU) was used to identify the relevant features. The utility preserved anonymization (UPA) technique was used to preserve privacy in the second phase. Additionally, in order to facilitate the formation of clusters for anonymization, the suggested approach lowers the dimensionality of the data. Real-time datasets are used in the experimental study to confirm the suggested method's efficacy. The outcomes demonstrate good accuracy (up to 98%) and high sensitivity (up to 98.63%), which enables us to assert effective attribute selection for anonymization. It is thus demonstrated that the suggested strategy successfully eliminates the unnecessary features, reducing the complexity of clustering.

Kaur *et al.,* [13] have created an enhanced version of k-degree anonymization on social network graphs. This version, known as NeuroSVM, preserves the average path length of the graph while drastically reducing the addition of noisy nodes and noisy edges. A few criteria, including average path length, precision, recall, F-measure, and information loss, were used to assess the suggested method. Experimental evidence has demonstrated that, in comparison to current procedures, the suggested technique has reduced average path length distortion. The suggested method reduced information loss and had an accuracy rate of over 75%.

A new approach for achieving k-anonymity through more efficient clustering was developed by Chavhan and Challagidad [14]. Most clustering algorithms require more processing power to handle the data, but a better cluster array would be produced if the initial centers found were compatible with the data configuration. In this work, we offer a dissimilarity tree based approach for NCP and for finding a somewhat more accurate cluster and an improved initial centroid with reduced computation time. The anonymized dataset's overall information loss was, on average, 20% less than that of alternative methods, according to graphical results. Additionally, it can handle numerical and category features.

A multi-level privacy preserving k-anonymity, an enhanced protection model based on k-anonymity, has been presented by Weng and Chi [15]. It splits data into distinct groups and requires each group to satisfy its own privacy need. Additionally, we offer a workable solution that guarantees the property by employing clustering techniques. The assessment conducted on an actual dataset validates that the suggested approach possesses the benefits of greater privacy parameter setting flexibility and greater data value compared to conventional k-anonymity. Casino *et al.,* [16] have created a novel micro aggregation-based PPCF technique that simultaneously generates precise suggestions and distorts data to provide k-anonymity. The suggested approach perturbs data more effectively, according to experimental results, than the popular distortion method based on Gaussian noise addition.

A collaborative anonymization strategy has been presented by Wong *et al.,* [17] with the goal of boosting respondents' confidence during data gathering. In contrast to previous research, our methodology does not disclose the entire set of quasi-identifiers (QID) to the agency or other data collector both before and after the data anonymization procedure. QID can be both identifying and sensitive, thus we gave respondents the option to conceal critical QID characteristics from others. Our procedure makes sure that, prior to the responders sending their records to the agency, the intended protection level (k-anonymity) may be confirmed. Additionally, if a malicious agency alters the intermediate results or does not adhere to the protocol faithfully, we permit truthful respondents to indict it.

Yazdanjue *et al.,* [18] have developed social networks using k-anonymity. To begin with, maximize 1-NSIL while minimizing normalized structural information loss (NSIL) by using the particle swarm optimization (PSO) algorithm to optimize the segmentation technique in the k-anonymity approach. Even though having a better rate of convergence than the formerly used genetic algorithm (GA) technique, the PSO-based approach did not result in a lower NSIL score. Thus, they propose hybrid solutions depending on the GA and PSO algorithms in order to reach the NSIL value supplied by GA optimization while maintaining the high convergence rate acquired from the PSO approach. Finally, the edge generalization method is used depending on their connections to produce indistinguishable nodes. The simulated outcomes show how effectively the approach balances the maximal 1-NSIL with the pace of convergence of the algorithm.

Fu *et al.,* [19] have analyzed varietal de-anonymization in an SN model that is better realistic than previous work and is characterized by overlapping populations. They create a very well-objective function using MMSE that minimizes the anticipated number of users that are incompatible. They

demonstrate the NP-hardness of minimizing MMSE and successfully turn it into WEMP, which eliminates the conflict between complication and optimality: (i) Under moderate conditions, WEMP asymptotically delivers a minimal mapping error made possible by increasing overlapping intensity; (ii) WEMP can be algorithmically addressed using CBDA, which precisely identifies the WEMP optimum. Several tests further support the CBDA's efficiency in overlapping populations.

3. Proposed Methodology

Globally, the number of people using online social networks (OSNs) has increased dramatically, particularly after the COVID-19 epidemic. OSNs are now an essential component of many people's everyday lives. Hence, the need for privacy regulations to shield users from harmful sources grows with the dependence on OSNs. The most popular method for protecting privacy is anonymizing data, which involves deleting or altering some information while maintaining as much of the original data as feasible. The primary limitation of current anonymity approaches is their inability to fend off resemblance attacks and attribute/link disclosure. Additionally, they have a significant level of information loss in the publicly available database. Furthermore, they can only be used for mixed social networks that contain both graph and data matrices because they have been offered for anonymization at the graph or data matrix level. This study presents a hybrid anonymizing approach based on self-adaptive honey badger optimization technique and k member Gaussian kernel fuzzy c means clustering (KGKFCM-SAHBO) to solve these shortcomings. This approach simultaneously performs the anonymization procedure at the matrix and graph levels, greatly reducing information loss. The first stage, the stage of cluster optimization, and the stage of privacy preservation comprise the three stages of the suggested paradigm. The input OSN data normalization and initial clustering using k member kernel fuzzy c means clustering (KGKFCM) are carried out in the first stage. Using the self-adaptive honey badger optimization (SAHBO) technique, the clusters are optimized during the cluster optimization phase. The clusters are further improved during the privacy preservation phase to guarantee l-diversity and t-closeness. The following part provides a full explanation of the proposed model, while the suggested framework is depicted in Figure 1.

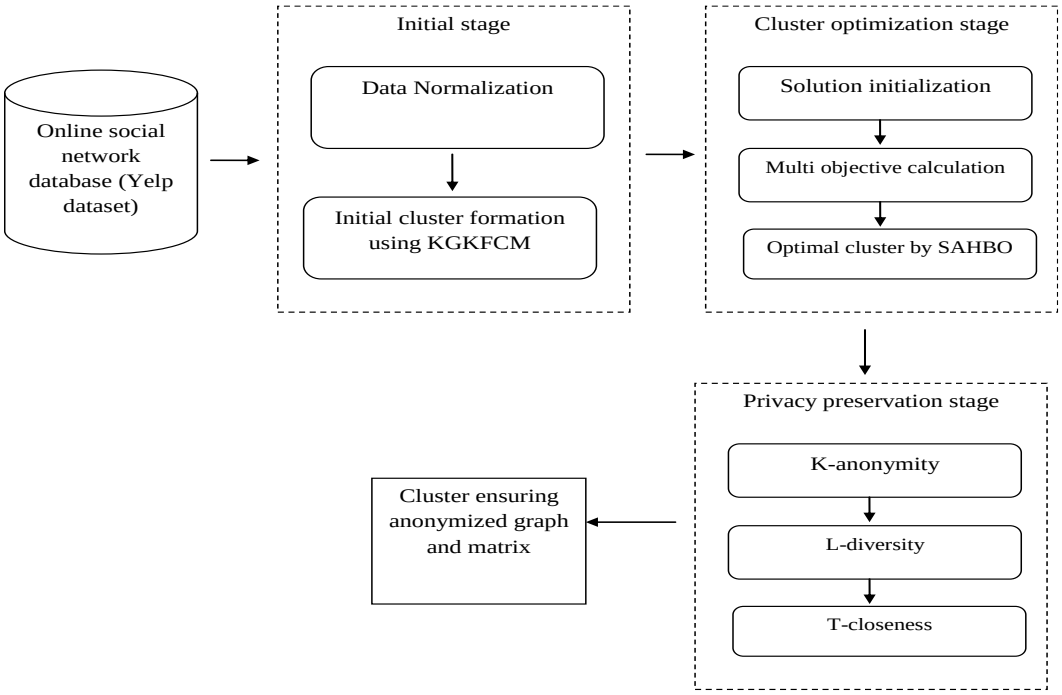

Fig. 1. Proposed framework for privacy preserving with KGKFCM

A social network's features typically consist of matrix properties (user personal attributes) and graph properties (user edges). Let's have a look at the OSN data input, which are shown as a graph G with vertices V and edges E. The social network users are represented by V, while a link or connection between two users or vertices is represented by E. The network of the suggested approach includes x vertices, and each vertex has y associated properties. The main objective of the suggested approach is to preserve privacy for social network graph G while ensuring that all of its elements (E, V, and A) are anonymized.

3.1 Initial Stage

First, the raw OSN data is normalized, and the k member Gaussian kernel fuzzy c means clustering (KGKFCM) is used to create the first clusters. The users, along with their attributes and the edges connecting them, make up the input raw OSN data. Therefore, in order to create the sets of nodes and their accompanying characteristics by carrying out the statistical operations, the data normalization step is necessary. After the noise is eliminated and each attribute is given a weight through data normalization, the information is prepared and stored. In this case, normal qualities have a weight of 0.5, meaning that grouping them with sensitive traits is more significant. Given a weight of one, the sensitive attributes are prioritized above the regular attributes for clustering purposes. Improving the quality of OSN data is the aim of the data normalization stage.

3.2 Clustering Using K Member Gaussian Kernel Fuzzy C Means Clustering (KGKFCM)

Using the k member Gaussian kernel fuzzy c means clustering (KGKFCM), the normalized inputs of vertices (V) and attributes (A) provided to the clustering algorithm aid in the accurate computation of cluster centroids and associated cluster members (CMs). The fuzzy c means (FCM) algorithm is a popular method for clustering data. The FCM algorithm uses sample points to create whole and sub vector spaces based on distance measurements. However, the convergence is imprecise and sluggish when non-linear data is divided. A kernel-based fuzzy C-Means approach that uses kernel functions in place of Euclidean distance was used to solve this issue. For various conditions, the Euclidean distance can be substituted to select different kernels. For clustering, a Gaussian kernel works well since it allows for the formulation of necessary requirements. Below is a step-by-step explanation of the GKFCM method.

Step 1: The FCM algorithm's objective function and Gaussian kernel variant are shown in Eq. (1).

$$ObjectiveFunction = \sum_{i=1}^{n}\sum_{j=1}^{m} D_{ij}^{o}(1 - GK(U_i - CC_j)) \quad (1)$$

$$GK(U_i, CC_j) = exp(-||U_i - CC_j||^2/\sigma^2) \quad (2)$$

Where $o$any is real number higher than 1, $D_{ij}^{o}$is the degree of membership of U_i in the cluster,$U_i$represents the users, $j$is the d-dimensional measured data, $CC_j$is the d- dimension centre of the cluster.

Step 2: Determine the centers of fuzzy clusters CC_j

$$CC_j = \frac{\sum_{i=1}^{n} D_{ij}^{o} GK(U_i, CC_j) U_i}{\sum_{i=1}^{n} D_{ij}^{o} GK(U_i, CC_j)}; j = 1,2.... m \quad (3)$$

Step 3: Use to determine the fuzzy membership function D_{ij}

$$D_{ij} = \frac{(1-GK(U_i, CC_j))^{-1/(o-1)}}{\sum_{i=1}^{n}(1-GK(U_i, CC_j))^{-1/(o-1)}}; j = 1,2.... m \quad (4)$$

The input dataset has been grouped using the aforementioned procedure. The suggested model uses clustering on pre-processed OSN data in order to attain k-anonymity. Nevertheless, the suggested model is unable to attain k-anonymity using the GKFCM. Some clusters with zero or less than K members are the result of the clustering procedure. There's a chance that many clusters have more than K members. Therefore, for clusters in the K anonymization privacy preserving strategy, this method is unsuccessful. Here, a unique K-member GKFCM is presented in order to allay these worries. First, the random cluster centre is chosen in KGKFCM, and then all of the users are grouped into the first clusters using GKFCM. Subsequently, clusters that meet the K anonymization criteria and have a member count more than or equal to K are designated as G1, while other clusters are designated as G2. Cluster splitting and merging are then carried out to ensure that every cluster achieves a balanced cluster while also establishing the K condition. Furthermore, it guarantees that the K-anonymity K anonymization requirement is satisfied.

3.3 Cluster Optimization Stage

Here, self-adaptive honey badger optimization (SAHBO) is presented to further anonymize the data when all clusters satisfy the K anonymization criterion. The honey badger algorithm (HBA) mimics how honey badgers' forage. The honey badger has two methods to find food sources: it either follows the honeyguide bird or smells and digs. We refer to the first situation as the "digging mode" and the second as the "honey mode." Using the sensing skills of the previous phase, it locates the prey, and once there, it searches the surroundings to find the best spot for digging and capturing it. Using the honeyguide bird as a guide in the last set, the honey badger finds the beehive directly. Although the classic HBO has demonstrated its superiority in terms of speed of convergence, quality of solution, and exploration-exploitation balance, it still has poor search accuracy. This uses the levy flight approach to get around such problems. The Levy flight mechanism is incorporated to increase HBO's search area. Here is an explanation of the steps in the self-adaptive HBO algorithm.

3.3.1 Solution initialization

The number of clusters and the clustering assignment vector are first determined by random selection of features, data attributes, and edges, both of which are generated from the KGKFCM. The equation below displays the initial solution format is shown in Eq. (5).

$$Solution(H) = \{C, CV, F, D, E\} \quad (5)$$

3.3.2 Fitness function

Cost and penalty are included in the proposed model's multi-objective fitness function. Three goals, including average distortion ratio, cluster balance, and cluster error, are included in the cost function.

The total number of restrictions that have not been satisfied is used to determine the punishment function. The fitness function with many objectives is is shown in Eq. (6).

$$fitness = min(\ Cost \times (1 + penalty) \tag{6}$$

3.4 Calculation of Intensity

Intensity may be correlated with both the honey badger's distance from its prey and its concentration.

$$In_i = \xi_1 . \frac{(H_i - H_{i+1})^2}{4\pi (H^{prey} - H_i)^2} \tag{7}$$

H^{prey} represents the present location of prey and $\xi_1$denotes the random number between 0 and 1.

3.5 Density Factor Updating

The smooth transition from exploration to exploitation is ensured by the density factor, which regulates time-varying randomness.

$$d_f = \lambda . e^{\frac{-t}{T}} \tag{8}$$

λ denotes a constant value which is greater than 1 and *t* denotes the current iteration and *T* represents the maximum iteration.

3.6 Prey Position Updating

There are two phases to the update process: the digging phase and the honey phase.

Digging phase: The honey pot shifts in a heart-shaped pattern in the direction of the food as you dig.

$$H^{new} = H^{prey} + f_g . \gamma . I_n . H^{prey} + f_g . \xi_2 . d_f . (H^{prey} - H_i) . |cos(2\pi \xi_3) . (1 - cos(2\pi \xi_4))| \tag{9}$$

$$f_g = \begin{cases} 1, \xi_5 \leq 0.5 \\ -1, Otherwise \end{cases} \tag{10}$$

γ denotes the ability of the honey badger to get food and ξ_2,ξ_3,ξ_4,ξ_5 ∈ [0,1] and $f_g$denotes the search direction.

Honey phase: A honey badger follows the honey guide bird during the honey phase to get to the beehive.

$$H^{new} = H^{prey} + f_g . \xi_6 . d_f . (H^{prey} - H_i) \tag{11}$$

where ξ_6 ∈ [0,1]

3.7 Levy Flight Updating

Levy flight is a random walk that conforms to the Levy probability distribution and has different step lengths. In order to prevent the population from getting stuck in a local optimum, it can be utilized to shift the solution. The most common method for producing steps from a Levy distribution is the Mantegna algorithm. The updated Levy flight-based solution is provided below,

$$step = \frac{n_1}{|n_2|^{1/\tau}} \tag{12}$$

$$n_1 \sim N(0, \sigma_{n_1}^2) and n_2 \sim N(0, \sigma_{n_2}^2) \tag{13}$$

$$\sigma_{n_1} = \left(\frac{\chi(1+\tau)Sin(\prod \cdot \tau/2)}{\chi((1+\tau/2)\cdot\tau\cdot 2^{(\tau-1)/2}}\right)^{1/\tau}; \sigma_{n_2} = 1 \tag{14}$$

τ denotes the constant value and $n_1$and $n_2$represents the normally distributed number and σ_{n_1}, σ_{n_2} represents the standard deviation values χdenotes the gamma distribution.

3.8 Termination Criteria and Privacy Preservation Stage

The population updating procedure is repeated up until the final iteration of the algorithm. The optimal solution, which consists of a collection of variables that depict the optimal clustering of the social network data table, is obtained when the SAHBO run is finished. Three restrictions are imposed by the proposed KGKFCM-SAHBO to defend the anonymized dataset from identity, attribute disclosure, and similarity assaults: K-anonymity, L-Diversity, and T-Closeness.

K-anonymity: When at least K-1 samples from the feature set cannot be identified from any sample, the adjusted table meets the K anonymity criterion.

L-Diversity: When there are at least L distinct values for each sensitive attribute inside a cluster of users, that cluster is referred to as L diverse.

T-Closeness: If the gap between each sensitive attribute's distribution in this cluster and the attribute's distribution in the global dataset is less than a desired threshold T, the cluster of users is referred to as T-closeness.

The suggested clustering technique, which is based on the aforementioned procedure, guarantees the privacy preservation of vertices/users and their attributes, with the least amount of sensitive data loss and the least amount of computational work.

4. Result and Discussion

The results of the experimental work for the proposed model's performance analysis are explained in this section. To run the experiments using MATLAB on Windows 10 with an Intel 13 processor and 4GB RAM. We conducted experiments on an actual Yelp dataset from Jure [20] in order to evaluate the effectiveness of each strategy. The Yelp dataset contains a set of user reviews, in which each user has numerous connections and access to the profiles of those connections. The degree of anonymization, information loss, and execution time are used to assess the performance of the suggested model.

Degree of anonymization: The degree of the user and the cluster to which it is assigned are the same and described as shown in Eq. (15).

$$DOA = deg\,r\,ee(CC_{u_i}) \times i \quad (15)$$

CC_{u_i} denotes the degree of anonymization of user u_i that belongs to cluster CC.

Three state-of-the-art techniques are used to compare the experimental outcome of the suggested model. First, PSO-GA, a hybrid swarm-intelligence-based OSN clustering method from Yazdanjue *et al.,* [21] was used; second, for privacy protection, l-diversity enhanced equi-cardinal (LECC) clustering by Siddula *et al.,* [22] was used. K means clustering and one pass algorithm-based anonymization constituted the third technique from Gangarde *et al.,* [23]. These techniques from Nabilah *et al.,* [24] and Hamrelaine *et al.,* [25] were chosen because, due to their clustering approach, they are both closely related to the proposed model and because it has been suggested that they can be used to achieve OSN anonymization. The experimental findings are displayed in Figure 2.

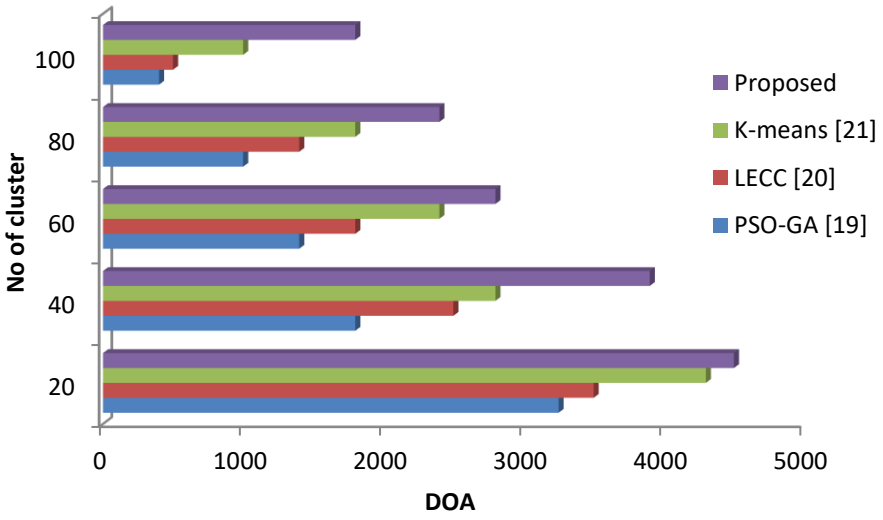

Fig. 2. Comparison of degree of anonymization

In comparison to state-of-the-art approaches, the suggested method achieves a higher degree of anonymization when analysing Figure 2. The first finding of this result is that anonymization reduced as the cluster size increased. The key explanation for this is that while the fraction of at least K-anonymous users in each cluster declines as cluster size increases, the small number of clusters sustains the large number of K-anonymous users. The suggested model's DOA for a cluster size of 20 is 4500, whereas the current methods and provide DOA values of 3250, 4300, and 3500, respectively. Similarly, the DOA value for cluster sizes 40, 60, 80, and 100 is determined by the suggested approach. The suggested strategy performed better than the performances of the current methods. Figure 3 shows the plot of the suggested comparison of information loss.

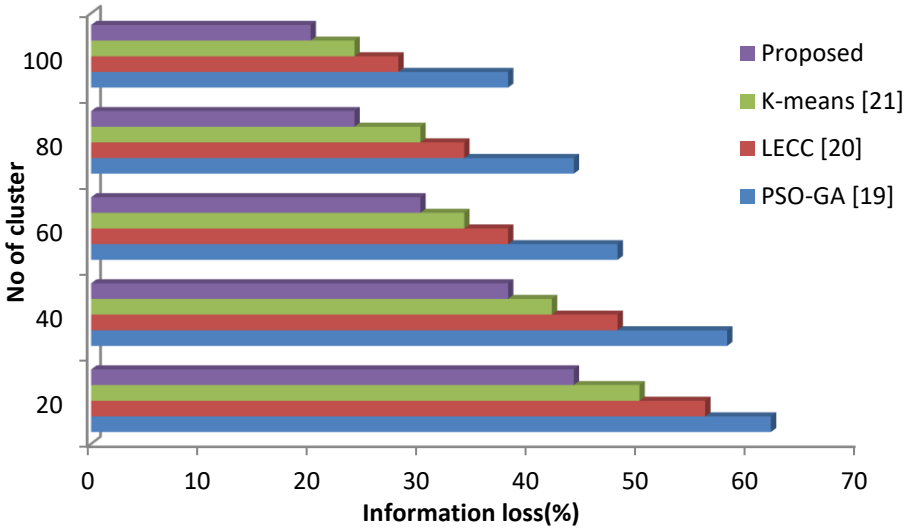

Fig. 3. Comparison of information loss

Figure 3 illustrates the results of the proposed information loss for various cluster sizes. For example, for cluster size 20, the information loss of the proposed model is 44%, while the information loss of the current technique from Jure [20] is 62%, and the information loss of the current methods from Yazdanjue *et al.,* [21] and Siddula *et al.,* [22] is 56% and 50%, respectively. The suggested model offers the least amount of information loss when compared to the current approach. When compared to the current method, the information loss of the suggested model is at a minimum of 38% for cluster sizes of 40. The information loss of the suggested model is 30% for cluster sizes up to 60, while the information loss of the current methods from Jure [20], Yazdanjue *et al.,* [21], and Siddula *et al.,* [22] is 48%, 38%, and 34%, respectively. When compared to the current method, the proposed model's information loss for cluster sizes of 80 is a minimum of 24%. The information loss of the suggested model is 20% for cluster sizes up to 100, while the information loss of the current methods from Jure [20], Yazdanjue *et al.,* [21], and Siddula *et al.,* [22] is 38%, 28%, and 24%, respectively. When compared to the current method, it is evident from the findings that the proposed model has the least amount of information loss. A greater number of clusters guarantees a lower number of users who are at least K-anonymous, while growing cluster sizes result in less loss of sensitive data. As a result, for a large number of clusters as opposed to a small number of clusters, this guarantees a minimal loss of critical information.

Figure 4 illustrates the results of the proposed execution times for various cluster sizes. For example, for a cluster size of 20, the proposed model takes 140s to execute, while the current approach from Jure [20] takes 225s, and the existing methods from Yazdanjue *et al.,* [21] and Siddula *et al.,* [22] both take 160s. The suggested model has a shorter execution time than the current procedure. The suggested model's execution time, for a cluster size of 40, is 180 s, which is the minimum required time for the current approach. The suggested model takes 220 seconds to execute for a cluster size of 60, however the current approach from Jure [20] takes 275 seconds, and the current methods from Yazdanjue *et al.,* [21] and Siddula *et al.,* [22] take 240 and 246 seconds to execute. The suggested model's execution time, for a cluster size of 80, is 275 s, which is the minimum required time for the current approach. The suggested model takes 300 seconds to execute for a cluster size of 100, whereas the current approach from Jure [20] takes 350 seconds, and the previous methods from Yazdanjue *et al.,* [21] and Siddula *et al.,* [22] take 300 and 310 seconds to execute. The findings make it evident that, in comparison to the current approach, the suggested model has the shortest execution time. The experimental results show that the proposed strategy reduced

information loss and improved the degree of anonymization when compared to existing methodologies.

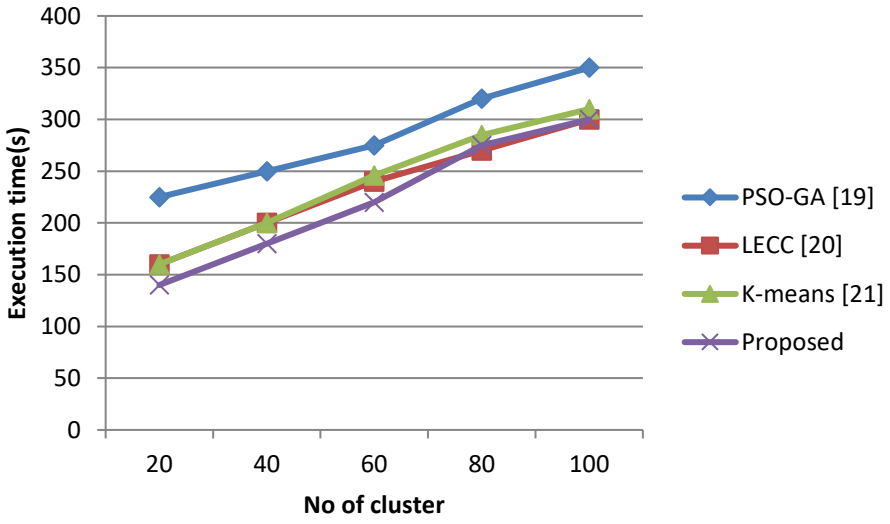

Fig. 4. Comparison of execution time

5. Conclusion

To achieve a high degree of anonymity and minimal loss of important structural information, a unique anonymization technique was presented for OSNs. To maintain privacy in social networks, a combined approach based on K-member Gaussian kernel fuzzy means clustering (KGKFCM) and the self-adaptive honey badger optimisation (SAHBO) algorithm has been developed. Using the Yelp dataset, the experimental findings demonstrate the efficiency of the proposed model and evaluate a few parameters, including execution time, degree of anonymization, and information loss. In order to assess the effectiveness of the suggested model, experiments were carried out with different cluster sizes. Compared to the current methods, the suggested model's average result increased anonymization by 22% and decreased information loss by 12%. Our goal for future work is to lead changes to the data/graph using an efficient method and we will use this method in real time applications.

References

[1] Skarkala, Maria E., Manolis Maragoudakis, Stefanos Gritzalis, Lilian Mitrou, Hannu Toivonen, and Pirjo Moen. "Privacy preservation by k-anonymization of weighted social networks." In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 423-428. IEEE, 2012. https://doi.org/10.1109/ASONAM.2012.75

[2] Vinod, D., Bharathiraja, N., Anand, M., & Antonidoss, A. (2021). An improved security assurance model for collaborating small material business processes. *Materials Today: Proceedings*, 46, 4077-4081. https://doi.org/10.1016/j.matpr.2021.02.611

[3] Marappan, R., Vardhini, P. H., Kaur, G., Murugesan, S., Kathiravan, M., Bharathiraja, N., & Venkatesan, R. (2023). Efficient evolutionary modeling in solving maximization of lifetime of wireless sensor healthcare networks. *Soft Computing*, 27(16), 11853-11867. https://doi.org/10.1007/s00500-023-08623-w

[4] Abbasi, Afsoon, and Behnaz Mohammadi. "A clustering-based anonymization approach for privacy-preserving in the healthcare cloud." *Concurrency and Computation: Practice and Experience* 34, no. 1 (2022): e6487. https://doi.org/10.1002/cpe.6487

[5] Wu, Yingjie, Zhihui Sun, and Xiaodong Wang. "Privacy preserving k-anonymity for re-publication of incremental datasets." In *2009 WRI World Congress on Computer Science and Information Engineering*, vol. 4, pp. 53-60. IEEE, 2009. https://doi.org/10.1109/CSIE.2009.549

[6] Jayanthi, E., Ramesh, T., Kharat, R. S., Veeramanickam, M. R. M., Bharathiraja, N., Venkatesan, R., & Marappan, R. (2023). Cybersecurity enhancement to detect credit card frauds in health care using new machine learning strategies. *Soft Computing*, 27(11), 7555-7565. https://doi.org/10.1007/s00500-023-07954-y

[7] FeiFei, Zhao, Dong LiFeng, Wang Kun, and Li Yang. "Study on privacy protection algorithm based on k-anonymity." *Physics Procedia* 33 (2012): 483-490. https://doi.org/10.1016/j.phpro.2012.05.093

[8] Majeed, Abdul, and Sungchang Lee. "Anonymization techniques for privacy preserving data publishing: A comprehensive survey." *IEEE access* 9 (2020): 8512-8545. https://doi.org/10.1109/ACCESS.2020.3045700

[9] Kumar, B. Santhosh, T. Daniya, N. Sathya, and R. Cristin. "Investigation on privacy preserving using K-anonymity techniques." In *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-7. IEEE, 2020. https://doi.org/10.1109/ICCCI48352.2020.9104109

[10] Yan, Yan, Eyeleko Anselme Herman, Adnan Mahmood, Tao Feng, and Pengshou Xie. "A weighted k-member clustering algorithm for k-anonymization." *Computing* (2021): 1-23. https://doi.org/10.1007/s00607-021-00922-0

[11] Kiran, Ajmeera, and N. Shirisha. "K-Anonymization approach for privacy preservation using data perturbation techniques in data mining." *Materials Today: Proceedings* 64 (2022): 578-584. https://doi.org/10.1016/j.matpr.2022.05.117

[12] Srijayanthi, S., and T. Sethukarasi. "Design of privacy preserving model based on clustering involved anonymization along with feature selection." *Computers & Security* 126 (2023): 103027. https://doi.org/10.1016/j.cose.2022.103027

[13] Kaur, Harmanjeet, Nishtha Hooda, and Harpreet Singh. "k-anonymization of social network data using Neural Network and SVM: K-NeuroSVM." *Journal of Information Security and Applications* 72 (2023): 103382. https://doi.org/10.1016/j.jisa.2022.103382

[14] Chavhan, Kalpana, and Praveen S. Challagidad. "Anonymization Technique For Privacy Preservation In Social Networks." In *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, pp. 131-136. IEEE, 2021. https://doi.org/10.1109/ICEECCOT52851.2021.9708007

[15] Weng, Jui-Hung, and Po-Wen Chi. "Multi-level privacy preserving k-anonymity." In *2021 16th Asia Joint Conference on Information Security (AsiaJCIS)*, pp. 61-67. IEEE, 2021. https://doi.org/10.1109/AsiaJCIS53848.2021.00019

[16] Bharathiraja, N., Pradeepa, K., Murugesan, S., Hariharan, S., & Veeramanickam, M. R. M. (2022, December). A Novel Framework for Cyber Security Attacks on Cloud-Based Services. In 2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP) (pp. 1-4). IEEE.

[17] Wong, Kok-Seng, Nguyen Anh Tu, Dinh-Mao Bui, Shih Yin Ooi, and Myung Ho Kim. "Privacy-preserving collaborative data anonymization with sensitive quasi-identifiers." In *2019 12th CMI Conference on Cybersecurity and Privacy (CMI)*, pp. 1-6. IEEE, 2019. https://doi.org/10.1109/CMI48017.2019.8962140

[18] Yazdanjue, Navid, Mohammad Fathian, and Babak Amiri. "Evolutionary algorithms for k-anonymity in social networks based on clustering approach." *The Computer Journal* 63, no. 7 (2020): 1039-1062. https://doi.org/10.1093/comjnl/bxz069

[19] Nagu, B., Arjunan, T., Bangare, M. L., Karuppaiah, P., Kaur, G., & Bhatt, M. W. (2023). Ultra-low latency communication technology for Augmented Reality application in mobile periphery computing. *Paladyn, Journal of Behavioral Robotics*, 14(1), 20220112.

[20] Jure, Leskovec. "SNAP Datasets: Stanford large network dataset collection." *Retrieved December 2021 from http://snap. stanford. edu/data* (2014). https://doi.org/10.1145/2740908.2744708

[21] Yazdanjue, Navid, Mohammad Fathian, and Babak Amiri. "Evolutionary algorithms for k-anonymity in social networks based on clustering approach." *The Computer Journal* 63, no. 7 (2020): 1039-1062. https://doi.org/10.1093/comjnl/bxz069.

[22] Siddula, Madhuri, Yingshu Li, Xiuzhen Cheng, Zhi Tian, and Zhipeng Cai. "Anonymization in online social networks based on enhanced equi-cardinal clustering." *IEEE Transactions on Computational Social Systems* 6, no. 4 (2019): 809-820. https://doi.org/10.1109/TCSS.2019.2928324

[23] Gangarde, Rupali, Amit Sharma, Ambika Pawar, Rahul Joshi, and Sudhanshu Gonge. "Privacy preservation in online social networks using multiple-graph-properties-based clustering to ensure k-anonymity, l-diversity, and t-closeness." *Electronics* 10, no. 22 (2021): 2877. https://doi.org/10.3390/electronics10222877.

[24] Nabilah, Nur Amira, Cheng Yee Ng, Nauman Riyaz Maldar, and Fatin Khalida Abd Khadir. "Marine hydrokinetic energy potential of Peninsular Malaysia by using hybrid site selection method." *Progress in Energy and Environment* (2023): 1-10. https://doi.org/10.37934/progee.26.1.110

[25] Pradeepa, K., Bharathiraja, N., Meenakshi, D., Hariharan, S., Kathiravan, M., & Kumar, V. (2022, December). Artificial Neural Networks in Healthcare for Augmented Reality. In 2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP) (pp. 1-5). IEEE. https://doi.org/10.1109/CCIP57447.2022.10058670