Hindawi
Advances in Fuzzy Systems
Volume 2022, Article ID 7783196, 9 pages
https://doi.org/10.1155/2022/7783196

Research Article

Clustering by Hybrid K-Means-Based Rider Sunflower Optimization Algorithm for Medical Data

A. Jaya Mabel Rani[1] **and A. Pravin**[2]

[1]*Research Scholer, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India*
[2]*Associate Professor, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India*

Correspondence should be addressed to A. Jaya Mabel Rani; ajayamabelrani@gmail.com

Received 10 November 2021; Revised 29 December 2021; Accepted 8 January 2022; Published 7 March 2022

Academic Editor: Bekir Sahin

Currently, medical data clustering is a very active and effective part of the research area to take proper decisions at the medical field from medical data sets. But medical data clustering is a very challenging issue due to limitless receiving data, vast size, and high frequencies. To achieve this and improve the performance with fast and effective clustering, this paper proposes a hybrid optimization technique, namely, the K-means-based rider sunflower optimization (RSFO) algorithm for medical data. In this research, initially, the data preprocessing phase has been carried out to clean the current input medical data, and then in the second phase, important features are chosen with the help of the Tversky index with holoentropy. Finally, medical data clustering has been carried out by using hybrid K-means-based rider sunflower optimization (RSFO) algorithm. RSFO is designed to produce optimum clustering centroid, which is the combination of two optimization techniques, such as rider optimization algorithm (ROA) and sunflower optimization (SFO). This hybrid algorithm can get the advantages of both K-means and RSFO technique and avoid premature convergence of K-means algorithm and high computation cost of optimization technique. K-Means clustering algorithm is used to cluster the medical data by using an optimum centroid. The efficiency of the proposed K-means-based rider sunflower optimization algorithm is examined with a heart disease data set and analyzed based on three different performance metrics.

1. Introduction

Today, there are many clustering models under the data mining, which is the sub-branch of artificial intelligence [1]. The main objective of the recent researchers is to project and provide the model in a very fast and effective manner for any type of data. This paper is also designed based on fast and effective clustering than existing models. For effective and efficient clustering of medical data, this paper proposed hybrid data clustering, which can avoid the premature convergence of K-means algorithm and can reduce the computation cost of optimization technique. Today, medical data are in the form of high dimensional with heterogeneous data, remote sensing data, geographical data, and a huge volume of real-time applications with hidden data [2]. For this type of medical data, there needs proper data analysis to take proper analysis decision by the medical personnel. At the time-optimal feature, the selection is also an important task for a proper decision-making system in the sense of good quality prediction. Medical data also have missing data, data redundancy, incomplete data, and data inconsistency, So there is in need of proper data preprocessing technique for the current input data. To manage all these difficult processing, we need proper intelligent techniques under machine learning and data mining, which is simply called medical data mining. The main objective is to cluster the data from a large amount of databases. There are different categories of clustering models such as partitioning clustering [3], density-based clustering [4], optimization-based clustering, and fuzzy-based (soft and clustering) and hybrid-based clustering algorithms [5, 6]. This proposed paper also used a hybrid optimization-based clustering algorithm to produce optimum clustering centroid with a faster clustering solution, which integrates the advantages of traditional

clustering with optimization-based clustering centroid. Each hybrid optimization method has its style and procedures for clustering the data. To inherit the advantages of both traditional and optimization-based clustering, this paper proposed the K-means-based rider sunflower optimization (RSFO) [7] algorithm. RSFO is the combination of both rider optimization (ROA) [8] and Sunflower optimization (SFO) [9], which is simply called rider sunflower optimization-based clustering. Here, K-means is the most popular and partition-based traditional clustering algorithm. But, it has some disadvantages like being very sensitive for initializing and premature convergence [10, 11]. The proposed integrated RSFO reduced the complexity of analyzing data, so obviously it takes less time and memory. At the same time, optimization-based clustering techniques take more computation cost due to a greater number of computing steps [12, 13]. This proposed model handles three steps, such as data preprocessing, feature selection by Tversky index with holoentropy [14, 15], and the final step of medical data clustering K-means rider sunflower optimization algorithm.

1.1. The Major Contribution of the Research Paper. This paper proposed global optimization technique, to find cluster centroid by using RSFO, and the number of clusters is defined by the user. This research work is structured in the following manner: Section 1 explains the basic introduction part, Section 2 gives challenges of medical data clustering, Section 3 explains the proposed method of medical data clustering, Section 4 explains experimental results and discussion with performance metrics, and Section 5 provides the simulation results based on the proposed K-means-based rider sunflower optimization algorithm. Comparative analysis is done in Section 6. Finally, Section 7 concludes the research work.

1.2. Literature Review. In this literature review, five various existing methods are reviewed with its advantages and disadvantages. In 2018, Yelipe et al. [16] proposed an imputation-based class-based clustering (IM-CBC) with the help of class-based clustering classifier (CBCC) in the form of hybrid clustering to find the similarity between the two different medical records. This paper used fuzzy-similarity functions and Euclidean distance of K-means algorithm to find the similarity of the clusters. Then, we used support vector machine (SVM) and k-nearest neighbour (k-NN) for performing the classification. Finally proposed algorithm produced the result with high accuracy and performance. However, this method did not consider fuzzy measures for better classifying and predicting of medical data. Then, in that same 2018, Das et al. [4] proposed a modified Bee colony optimization (MBCO) approach for clustering the data with probability-based selection method. This optimization-based clustering shows faster convergence than other methods. This modified Bee colony optimization hybrid is with K-means algorithm to improve the performance and to achieve a global optimal solution and classification accuracy with the help of chaotic theory. But, this optimization method is not used with multiobjective optimization functions for initializing the clustering centroid to process high-frequency data streams. Then, in 2019, Al-Shammari et al. [17] proposed density-based clustering in the sense of dynamic framework for classifying the medical data with the help of piece-wise aggregate approximation and the density-based spatial clustering algorithm to enhance the better performance and maintenance of the dynamic cluster. But, this proposed algorithm is not considered about high frequency of incoming data streams for updating and maintaining the data clusters. In 2019, Chauhan et al. [18] proposed two-step clustering technology for patient's disorder analysis using different data variables for optimal clusters with different shapes and sizes. The main objective of this paper is diagnosis of liver disease at the earlier stage from the hidden knowledge with huge database. In 2020, Baliarsingh et al. [19] proposed a medical data classification by memetic algorithm-based SVM (M-SVM), which is the combination of social engineering optimizer and emperor penguin optimization. This algorithm can classify the medical data in very accurate manner. But, this method is not applied for large-scale data sets.

2. Challenges

The following challenges are faced by existing methods: the first one is finding proper cluster centroid and producing clustering results based on optimum cluster centroids which are not guaranteed. Second, data preprocessing, such as missing data, data redundancy, and data inconsistency, is the next main challenge in data clustering. Then, in medical data clustering, it is complex research, in real-time applications due to massive volume, unlimited incoming data, a huge amount of heterogeneous data, and high frequency of data [20]. To avoid this problem, here we proposed the most relevant features of the medical data clustering, so it can reduce the complexity of data analysis with less time and memory.

3. Proposed K-Means Clustering-Based RSFO for Medical Data Clustering

This section boons the proposed clustering algorithm K-means clustering based on RSFO, for medical data clustering. Rider sunflower optimization (RSFO) technique is designed by combining ROA and SFO. This technique hybridizes the advantages of both optimization such as ROA and SFO to define optimum centroid with faster convergence. Figure 1 illustrates the view of the proposed hybrid optimization-based clustering.

3.1. Data Preprocessing. The good quality of input data can produce good quality of output. The medical input data may have missing data or inconsistent data or noisy data. This type of uncleaned input data may affect the quality of output. To produce good quality of input data, data must be cleaned by preprocessing techniques. Data preprocessing must be done with every input data to clean noisy data, missing data, and inconsistent data. So, data preprocessing is very important part for processing the input medical data to smoothen from the huge amount of data and eliminate inconsistent and noisy data to produce better clustering results.

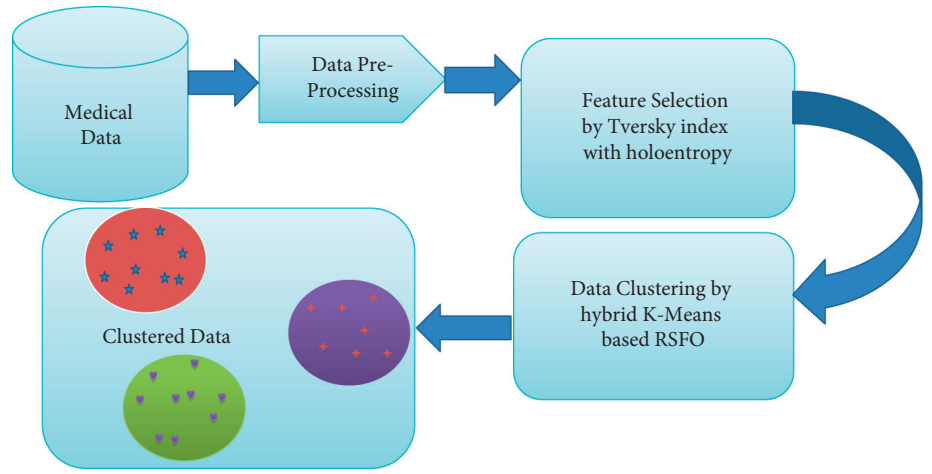

Figure 1: Pictorial representation of the proposed hybrid optimization-based clustering.

3.2. Feature Selection. The significance feature selection or feature extraction is the next important step to identify high relevant features to produce a better clustering solution. Here, all irrelevant features can be avoided or eliminated by using feature selection method. At the same time, complexity of the data is also reduced by reducing the number of features. There are different feature selection methods available to identify and remove irrelevant features, which depends on the data types used such as Laplacian score feature selection, and reference [21]explains about spectral feature selection for supervised and unsupervised learning [22], unsupervised feature selection with multisubspace randomization [23], etc. In this paper, feature selection is done by using the Tversky index with holoentropy.

3.2.1. Tversky Index Using Holoentropy. The Tversky index compares two data sets and finds similarities between the two data sets and is also used to extract the most relevant terms from the database, by using a feature evaluation function. At the same time, it can reduce the computational complexity to improve accuracy [24]. Tversky index uses Dice's coefficient and Tanimoto coefficient, which has a value between 0 and 1. The holoentropy is applied in Tversky index parameters α and β to find the relationship between features and attributes [25]. Holoentropy is calculated by using the product of entropy and weight function, which is given in the following:

$$\mathrm{HE}_{(\gamma)} = \omega \times \mathrm{Ent}(\gamma). \tag{1}$$

Here, $HE_{(\gamma)}$ refers to holoentropy, weight function is ω, and $Ent_{(\gamma)}$ is called as entropy measure. Here, the entropy $Ent_{(\gamma)}$ is calculated by

$$\mathrm{Ent}(\gamma) = -\sum_{i=1}^{u_\gamma} pi \log pi. \tag{2}$$

Here, $u_{(\gamma)}$ is the number of unique values in data and γ is used to apply on the Tversky parameters on α and β.

The procedural steps for RSFO are given as follows:

Step 1: initialize the initial parameters and randomly select the initial cluster centroids $C = \{Cen_1, Cen_2, \ldots, Cen_j\}$

Step 2: evaluate the fitness function by equation

$$F_{fits} = \frac{o_f}{Z}, \tag{3}$$

where "$\boldsymbol{Z}$" is the number of clusters and $\boldsymbol{O_f}$ is the objective function, which is calculated by the

$$0_f = \sum_{j=1}^{z}\left[\frac{\sum_{i=1}^{p} Dis\left(x_i^{(j)}, G_j\right)}{\left|U_{ij}\right|}\right]. \tag{4}$$

Here, "$\boldsymbol{P}$ "denotes the total number of data points and $|U_{ij}|$ is the total amount of data elements belonging to the cluster. Here, two data points distance measure is calculated by the

$$Dis\left(X_i, G_j\right) = \sqrt{\sum_{i=1}^{D}\left(X_i^{(j)} - G_j\right)^2}. \tag{5}$$

Cluster centroid will be recalculated by the

$$G_j = \frac{1}{N_j}\sum_{\forall X_j \epsilon C_j} X_i. \tag{6}$$

Find performance fitness value based on Euclidean distance measure between cluster centroid and data points of each intercluster and intracluster [26]. Then, cluster centroid will be found by rider sunflower optimization (RSFO).

$$F_{\mathrm{fits}} = \frac{\sum_{j=1}^{Z}\left[\sum_{i=1}^{P}\mathrm{Dis}\left(X_i^{(j)}, G_j\right)/\left|U_{ij}\right|\right]}{K}. \tag{7}$$

Update the optimal clustering centroid based on rider sunflower optimization (RSFO), which is given in following section.

3.2.2. Proposed RSFO Algorithm. For better clustering results, we are in need to produce optimum clustering centroids. So, to find a better clustering centroid, we used the RSFO algorithm, which is the combination of sunflower optimization (SFO) and rider optimization algorithm (ROA). This proposed model gets the rewards of both ROA and SFO to produce better clustering solutions with faster convergence. ROA [27] works by the behavior of different types of riders to the terminus. Here are four different types of riders such as bypass rider, attacker, follower, and overtaker, respectively. The sunflower optimization (SFO) works by the rotation of the sun. Sunflower always mimics the rotation, which is nature-inspired optimization [28]. This model can determine the good locations for better performance. At the same time, it uses high computation complexity due to high computation steps. To obtain a global optimal solution with better computation steps and fast performance, we used the hybrid ROA method with the SFO.

The procedural steps for hybrid K-means-based rider sunflower optimization is given as follows:

Step 1. Initialize the initial parameters

Step 2. Evaluate fitness function by Lagrangian optimization principle, which is given in the equations:

$$M = \sum_{p=1}^{K}\sum_{q=1}^{g}\left(w_{\mathrm{qp}}\left\|d_p - l_p\right\|^2 + \rho\ \ln\ w_{\mathrm{qp}} + \rho\ \ln\ \left\|d_q - l_p\right\|^2\right) = \sum_{q=1}^{g}\eta_q\left(-1\sum_{p=1}^{k}w_{\mathrm{qp}}\right). \tag{8}$$

Here, d_q refers to data object, user defined constant is denoted by ρ, lp refers to the center of the cluster, l_p is the fuzzy membership function, K indicates the total cluster centers, and g is the total data. Then, the above equation will derivate as follows:

$$\left(\frac{\partial M}{\partial w_{\mathrm{qp}}}\right) = \left\|d_q - l_p\right\|^2 + \frac{\rho}{w_{\mathrm{qp}}} - \eta_q = 0, \qquad w_{\mathrm{qp}} = \frac{\rho}{\left(\eta_q - \left\|d_q - l_p\right\|^2\right)}. \tag{9}$$

Since $\sum_{p=1}^{K} W_{qp} = 1$, which is given as

$$\sum_{p=1}^{K}\left[\frac{1}{\eta_q} - \left\|d_q - l_p\right\|^2\right] = \frac{1}{\rho}, \tag{10}$$

we should ensure $w_{qp} \geq 0 \geq 0$.

Step 3. Update the Position for the rider groups

For position updation, we used bypass rider to maximize the achievement rate. Bypass riders always track and follow a common usual path without other riders' information. The equation for position updation based on bypass rider is shown as

$$B_{t+1}(r,p) = \vartheta\left[B_t(t,p) * m(p) + B_t(\mu,p) * [1 - m(p)]\right]. \tag{11}$$

Here, the parameters ϑ, t, m, and μ indicate the random numbers from 0 to 1. Then, k indicates the number of iterations, which is defined by user. Assuming $\mu = r$, the equation is rewritten as

$$B_{t+1}(r,p) = \vartheta\left[B_t(t,p) * m(p) + B_t(r,p) * [1 - m(p)]\right]. \tag{12}$$

The sunflower optimization (SFO) updates the position or solution space by the rotation of sun. Sunflower always mimics the rotation of sun. Thus, the position updation of SFO is given by

$$B(r,p) = B_t(r,p) + y_r \times g_r. \tag{13}$$

Here, $B_t(r, p)$ denotes the current position at the time t, $B_{t+1}(r, p)$ is the updated position at the time t+1, $B_{t+1}(r, p)$ denotes the step of sunflower, and g_r refers to the direction of the sunflower.

$$B_t(r,p) = \frac{B_{t+1}(r,p)}{y_r} \times g_r. \tag{14}$$

For position updation, substitute the (15), which is the position updation of sunflower optimization in (13), that is the position updation of rider optimization.

$$B_{t+1}(r,p) = \vartheta B_t(t,p) * m(p) + \left(\frac{B_{t+1}(r,p)}{y_r}\right) \times g_r * [1 - m(p)], \tag{15}$$

$$B_{t+1}(r,p) = \vartheta B_t(t,p) * m(p) + B_{t+1}(r,p)[1 - m(p)] - y_r \times g_r * [1 - m(p)]. \tag{16}$$

Then, rearranging (15) and (16), we get

$$B_{t+1}(r,p) = \vartheta B_t(t,p) * m(p) + B_{t+1}(r,p) - \frac{B_{t+1}(r,p)m(p)}{y_r g_r} + y_r g_r m(p),$$

$$\frac{B_{t+1}(r,p)}{B_{t+1}(r,p)\vartheta} + \vartheta B_{r+1}(r,p)m(p) = \vartheta\left[\frac{B_t(t,p) * m(p)}{y_r t_r + y_r t_r m(p)}\right], \tag{17}$$

$$B_{t+1}(r,p)[1 - \vartheta + \vartheta m(p)] = \vartheta\left[\frac{B_t(t,p) * m(p)}{y_r t_r + y_r t_r m(p)}\right].$$

Then, the final equation will be written as

$$Bt + 1(r,p) = \frac{1}{[1 - \vartheta[1 - m(p)]]}\left[\frac{\vartheta[Bt(t,p) * m(p)]}{y_r t_r[1 - m(p)]}\right]. \tag{18}$$

Step 4. Defining the best solution

Here, the maximal fitness value is considered as a best solution and update rider optimization parameters for the best solution.

Step 5. Termination condition

The above steps 2 to 4 are repeated until the defined number of iterations are reached.

4. Experimental Results and Discussion

The experimental results provide for the proposed K-means-based rider sunflower optimization which is implemented using the programming language of Python 3.8.6 version in the Windows 10 operating system, Intel i5 core processor. Heart disease-based medical data set is used for conducting the experimentation from online source [29] to predict the heart disease risk factors with 300 instances and 7 main attributes such as age, random blood sugar, gender, body mass index, cholesterol, random blood sugar, smoking, systolic BP [30–33] with 500 iterations. These data were collected from Mr. Jims Johnson, Staff Nurse, whose goal is to provide the proper decision and suggestion to that particular heart patient. This decision can provide various suggestions to the patient such as healthy food diet, type of exercise, walking distance, medicine to take, walking distance, and regular medical checkup. For experimental results, only less quantity of data were taken. In real time, this algorithm can handle large quantity of data sets.

4.1. Performance Metric. The performance metrics for K-means-based rider sunflower optimization is to find data quality, similarity, and correct decision ratio from true positives, true negatives, false positives, and false negatives [34–37].

4.1.1. Accuracy

$$\text{Accuracy} = \frac{\text{AccT}^p + \text{AccT}^n}{\text{AccT}^p + \text{AccT}^n + \text{AccF}^p + \text{AccF}^n}, \tag{19}$$

where the parameters AccT^{p}, AccT^{n}, AccF^{p}, and AccF^{n} are indicating the total quantity of true positives, true negatives, false positives, and false negatives.

4.1.2. Jaccard Coefficient

$$\text{Jack}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}, \tag{20}$$

which is also called Jaccard similarity coefficient, which finds the similarity of two data sets that fall from the range 0% to 100%. When the percentage value increases, the similarity of two data sets also increases. Here, X and Y are two different data sets.

4.1.3. Random Coefficient

$$\text{Rand}_c = \frac{\text{RcT}^p + \text{RcT}^n}{\text{Rct}^p + \text{RcF}^p + \text{RcF}^n + \text{RcT}^n}, \tag{21}$$

where the parameters RcT^{p}, RcT^{n}, RcF^{p}, and RcF^{n} are indicating the total quantity of true positives, true negatives, false positives, and false negatives.

5. Simulation Results

Figure 2 shows 2-dimensional simulation results for the projected K-means-based rider sunflower optimization with different features of medical data in the sense of three different stages of the clusters.

In this figure, 2-D simulation results, i.e., red color indicates the high-risk factor of the heart disease; blue color indicates average risk factor; and green color denotes the less-risk factor of heart disease based on age vs cholesterol (in Figure 2(a)), age vs body mass index (in Figure 2(b)), and age vs random blood pressure (in Figure 2(c)). The abovementioned three diagrams differentiated based on the input features.

6. Comparative Analysis

The comparative analysis for the proposed K-means-based rider sunflower optimization using the performance metric is given below based on input size.

6.1. Comparative Result Analysis by Input Size. Figures 3(a)–3(c) show the qualified comparative study analysis by using input size which is varying from 50 to 300. The accuracy, Jaccard coefficient, and random coefficient proposed K-means-based RSFO input size 50 which are 67.037%, 60.748, and 61.907, as well as for the input size 300 are 90.026%, 89.3426%, and 92.767%, respectively.

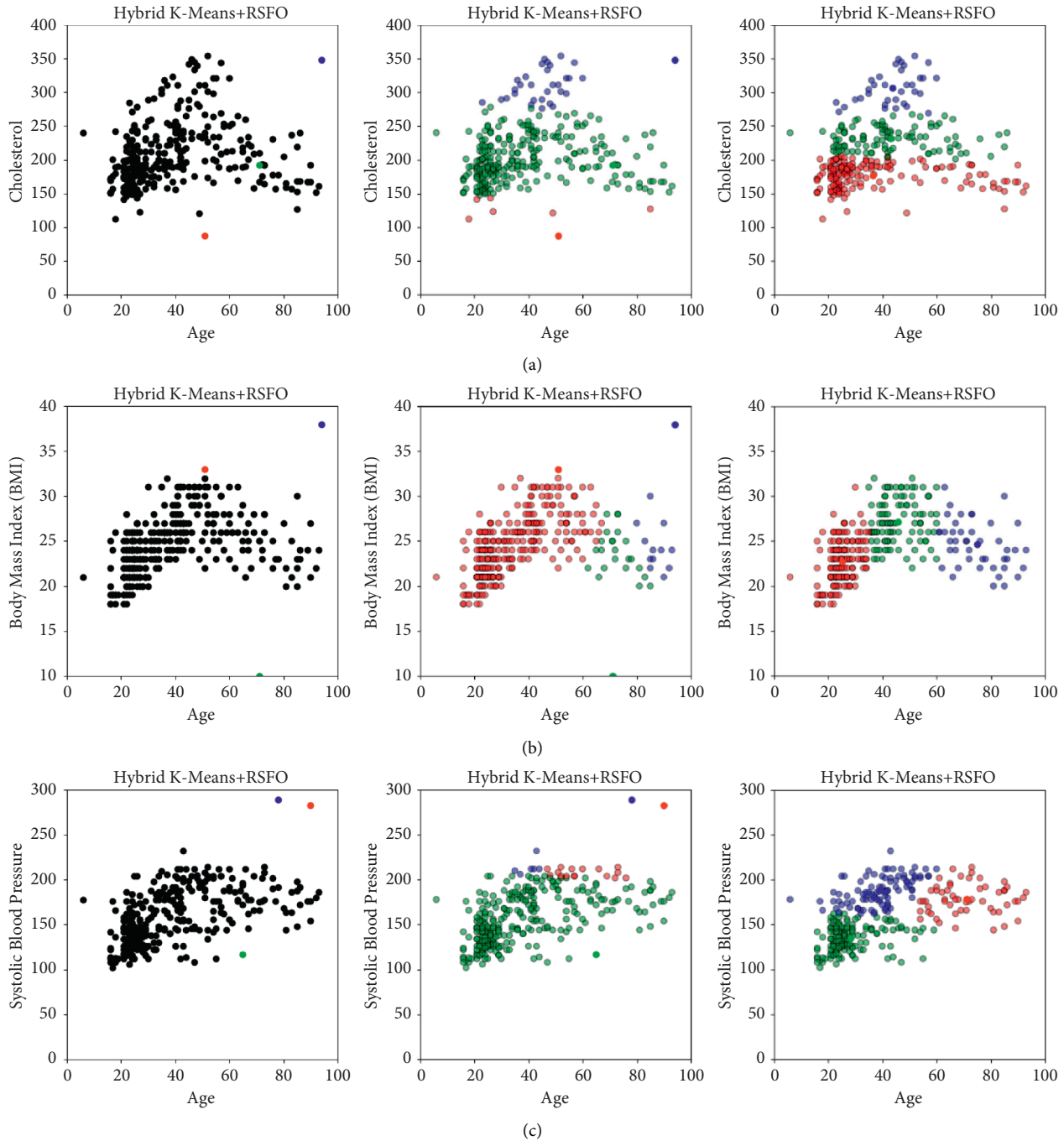

Figure 2: (a) Two-dimensional clustering simulation for age vs cholesterol by using hybrid K-means-based RSFO algorithm. (b) The 2-D clustering simulation of age vs body mass index by using K-means-based RSFO algorithm. (c) The 2-D clustering simulation for age vs random blood pressure using hybrid K-means-based RSFO algorithm.

6.2. Comparative Analysis by Table. The comparative analysis table is given in Table 1 with the performance measures of best accuracy, Jaccard coefficient, and random coefficient for the proposed K-means-based RSFO which are 90.026%, 89.3426%, and 92.767%, respectively. In this table, the proposed hybrid K-means and rider sunflower optimization technique are compared with existing K-means algorithm, K-harmonic means algorithm, and fuzzy C-means algorithm and show how much more accurate efficient solution is produced. Here, the proposed hybrid KM + RSFO technique produced more accurate efficient solution than existing methods.

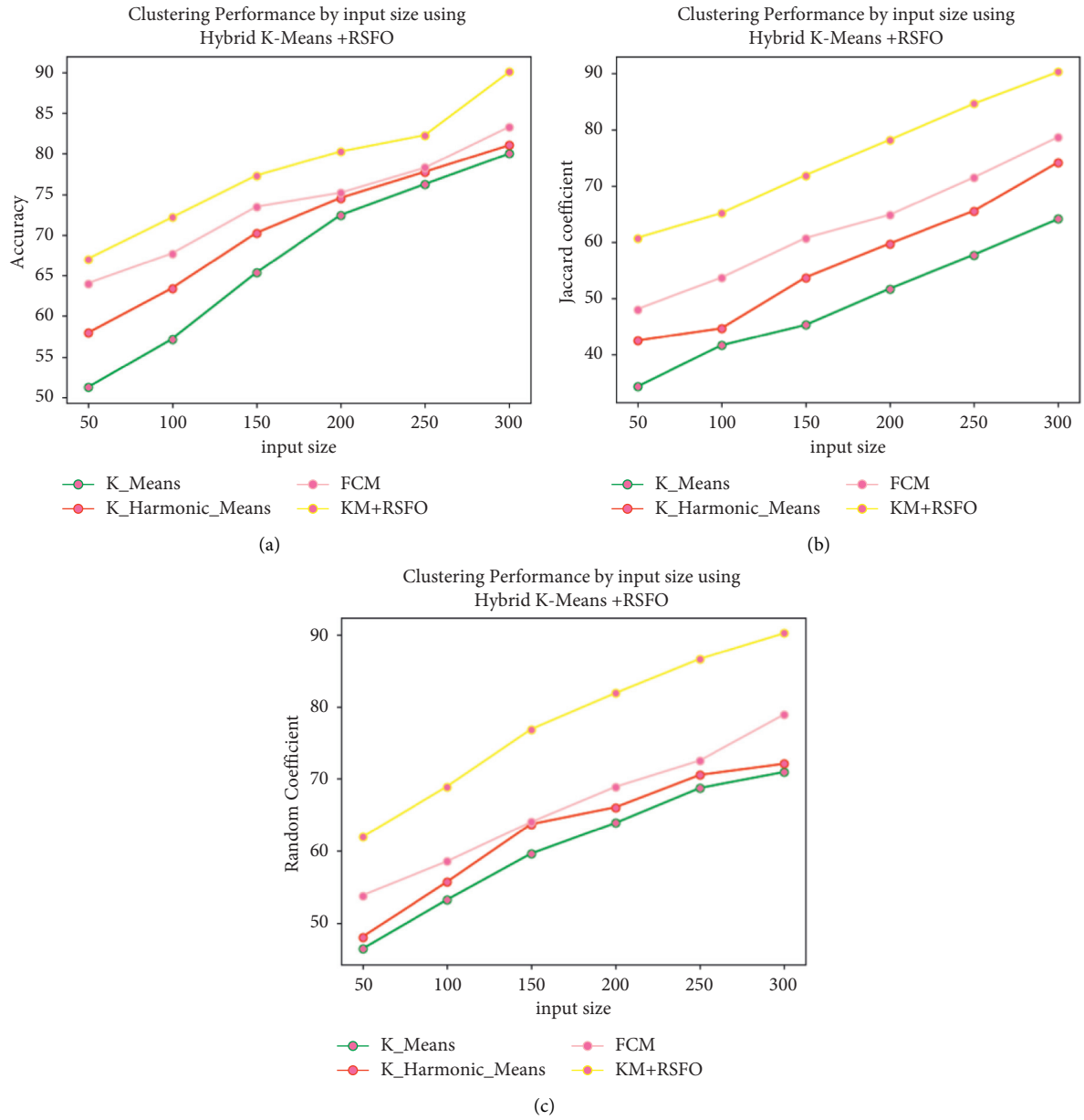

Figure 3: The comparative study analysis by input size using (a) accuracy, (b) Jaccard coefficient, and (c) random coefficient.

Table 1: Comparative result analysis table.

Input	Comparative metrics	K-means	KHM	FCM	KM + RSFO
Input size	Accuracy (%)	79.984	81.021	83.2842	**90.026**
	Jaccard coefficient (%)	64.109	74.253	78.735	**89.3426**
	Random coefficient (%)	70.953	72.095	78.936	**92.767**

Bold values show the result of the proposed KM+RSFO algorithm.

7. Conclusion

Thus, the proposed hybrid K-means-based rider sunflower optimization clustering algorithm for medical data analyses the risk factor of heart disease. The optimal centroid-based clustering solution is produced by using hybrid K-means-based rider sunflower optimization for heart disease-based medical data. The achievement of the proposed K-means-based RSFO algorithm is produced with best accuracy of 90.0236%, Jaccard coefficient of 89.3426%, and random coefficient of 92.767%. This hybrid clustering algorithm can get the advantages of both rider optimization and sunflower optimization thehniques. Deleted Here, RSFO is used to produce optimum clustering centroid and K-means

algorithm is used to produce fast clustering solution. So, this hybrid clustering algorithm can protect from premature convergence. As a future enhancement, this algorithm can be extended with multiobjective functions for more effective and better clustering centroid.

Data Availability

The medical data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. Russell and N. Peter, *Artificial Intelligence—A Modern Approach*, Pearson Education, London, UK, Third edition, 2009.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Elsevier, Amsterdam, Netherlands, 2012.

[3] S. Ayram and T. Kannan, "Introduction to partitioning based clustering methods with a robust example," Reports of the Department of Mathematical Information Technology Series. Software and Computational Engineering, University of Jyväskylä, Jyväskylä, Finland, 2006.

[4] P. Das, D. K. Das, and S. Dey, "A modified bee Colony optimization (MBCO) and its hybridization with k-means for an application to data clustering," *Applied Soft Computing*, vol. 70, pp. 590–603, 2018.

[5] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "PPHOPCM: privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing," *IEEE Trans Big Data*, vol. 5, no. 6, pp. 78–84, 2018.

[6] F. Bu, C. Hu, Q. Zhang, C. Bai, L. T. Yang, and T. Baker, "A cloud-edge-aided incremental high-order possibilistic c-means algorithm for medical data clustering," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 148–155, 2021.

[7] A. Jaya Mabel Rani and A. Pravin, "Optimization enabled block hole entropic fuzzy clustering approach for medical data," *The Computer Journal*, 2021.

[8] D. Binu and B. S. Kariyappa, "Rider NN: a new rider optimization algorithm-based neural network for fault diagnosis in analog circuits," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, pp. 2–26, 2018.

[9] G. F. Gomes, S. S. da Cunha, and A. C. Ancelotti, "A sunflower optimization (SFO) algorithm applied to damage identification on laminated composite plates," *Engineering with Computers*, vol. 35, no. 2, pp. 619–626, 2019.

[10] A. J. Mabel Rani and A. Pravin, "Multi-objective hybrid fuzzified PSO and fuzzy C-means algorithm for clustering CDR data," in *Proceedings of the 2019 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0094–0098, Chennai, India, April 2019.

[11] T. Velmurugan, "Performance Based Analysis between K-Means and Fuzzy C-Means Clustering Algorithm for Connection-Oriented Telecommunication Data," *Applied Soft Computing*, vol. 19, pp. 134–146, 2014.

[12] https://towardsdatascience.com/fuzzy-c-means-clustering-is-it-better--k-means-clustering-448a0aba1ee7.

[13] V. Kelner, F. Capitanescu, O. Leonard, and L. Wehenkel, "A hybrid optimization technique coupling an evolutionary and a local search algorithm," *Journal of Computational and Applied Mathematics*, vol. 215, no. 2, pp. 448–456, 2008.

[14] T. Cura, "A particle swarm optimization approach to clustering," *Expert Systems with Applications*, pp. 1582–1588, 2012.

[15] D. Bhutada, V. V. S. S. S. Balaram, and V. V. Bulusu, "Holoentropy based dynamic semantic latent dirichilet allocation for topic extraction," *International Journal of Applied Engineering Research*, vol. 11, pp. 1304–1313, 2016.

[16] U. Yelipe, S. Porika, and M. Golla, "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records," *Computers & Electrical Engineering*, vol. 66, pp. 487–504, 2018.

[17] A. Al-Shammari, R. Zhou, M. Naseriparsaa, and C. Liu, "An effective density-based clustering and dynamic maintenance framework for evolving medical data streams," *International Journal of Medical Informatics*, vol. 126, pp. 176–186, 2019.

[18] R. Chauhan, N. Kumar, and R. Rekapally, "Predictive data analytics technique for optimization of medical databases, Advances in Intelligent Systems and Computing," in *Soft Computing: Theories and Applications*, pp. 433–441, Springer, Berlin, Germany, 2019.

[19] S. K. Baliarsingh, W. Ding, S. Vipsita, and S. Bakshi, "A memetic algorithm using emperor penguin and social engineering optimization for medical data classification," *Applied Soft Computing*, vol. 85, 2020.

[20] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, no. 4, pp. 327–352, 1977.

[21] X. He, C. Deng, and P. Niyogi, "Laplacian score for feature selection," in *Proceedings of the NIPS'05 18th International Conference on Neural Information Processing Systems*, pp. 507–514, Columbia, Canada, December 2005.

[22] Z. Zheng in *Proceedings of the ICML '07 of the 24th international conference on Machine learning*, pp. 1151–1157, Corvalis, OR, USA, June 2007.

[23] H. Huang, X. Cai, and C.-D. Wang, "Unsupervised feature selection with multi-subspace randomization and collaboration," *Knowledge-Based Systems*, vol. 182, 2019.

[24] V. Likitha, S. Naik, and R. Manjunath, "Development of predictive model to improve accuracy of medical data processing using machine learning techniques," *International Journal of Scientific Research and Review*, vol. 7, no. 7, pp. 233–240, 2018.

[25] K. Hammouda and F. Karray, *A Comparative Study of Data Clustering Techniques*, 2000, http://www.pami.uwaterloo.ca/%20pub%20/hammou%20da/%20sde625-paper.pdf, p. 1, University of Waterloo, Ontario, Canada.

[26] https://en.wikipedia.org/wiki/Tversky_index.

[27] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications," *Expert Systems with Applications*, vol. 67, pp. 12–18, 2017.

[28] G. Wang, Y. Yuan, and W. Guo, "An improved rider optimization algorithm for solving engineering optimization problems," *IEEE Access*, vol. 7, pp. 80570–80576, 2019.

[29] M. H. Qais, H. M. Hasanien, and S. Alghuwainem, "Identification of electrical parameters for three-diode photovoltaic model using analytical and sunflower optimization algorithm," *Applied Energy*, vol. 250, pp. 109–117, 2019.

[30] http://rstudio-pubs-static.s3.amazonaws.com/24341_184a58191486470cab97acdbb%20fe78%20e%20d5.html.

[31] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2465571.

[32] S. D. Bhambere, "Oral health status, knowledge and caries occurrence in visually impaired students," *International Journal of Health Sciences & Research*, vol. 7, pp. 118–121, 2017.

[33] P. Andries and Engelbrecht, *Computational Intelligence an Introduction*, Wiley, Hoboken, NJ, USA, Second Edition, 2007.

[34] V. Kumar, "Implementation of Data Mining Techniques for Information Retrieval," Thesis, University of Cagliari, Cagliari, Italy, 2018.

[35] https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.

[36] https://towardsdatascience.com/fuzzy-c-means-clustering-is-it-better-k-means-clustering-448a0aba1ee7.

[37] S. Bhambere, "The long wait for health in India—a study of waiting time for patients in a tertiary care hospital in western India," *International Journal of Biomedical and Advance Research*, vol. 7, pp. 108–111, 2017.